Open Source Social Media as Sensors for Enabling Government Identification, Prediction and Response Applications

by

Swati Agarwal

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science
at

Indraprastha Institute of Information Technology, Delhi

Committee in charge:

Dr. Ashish Sureka, Advisor
Dr. Vikram Goyal, Co-Advisor
Prof. Dr. Jessica Rubart, Examiner
Dr P Radha Krishnal, Examiner
Dr. P. Krishna Reddy, Examiner

Winter February, 2017

Certificate

This is to certify that the thesis titled **"Open Source Social Media as Sensors for Enabling Government Identification, Prediction and Response Applications"** being submitted by **Swati Agarwal** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

Dr. Ashish Sureka (Advisor)
Principal Scientist, ABB Corporate Research Center, Bangalore, India

February, 2017
IIIT-Delhi, India

Prof. Vikram Goyal (Co-Advisor)
Associate Professor, Department of Computer Science, IIIT-Delhi, India

Biography

Swati Agarwal is a senior Ph.D. candidate at IIIT-Delhi working in Information Management and Data Analytics group. Her research interests are in the area of social computing and social informatics, machine learning, text mining and analytics, natural language processing and information visualization. Prior to enrolling in the Ph.D. program at IIIT-Delhi, she graduated from IIIT-Delhi with Masters in Technology in 2013 with the specialization in Data Engineering stream. During her Ph.D. education program, she went to IIIT-Hyderabad as a visiting researcher. She served as a program co-chair at South Asian Society of Criminology and Victimology. She also served as a technical program committee member at International Conference on Social Media Technologies, Communication, and Informatics 2017, Internal Conference of Intelligence and Security Informatics 2017, Workshop on Abusive Language Online, co-located with ACL 2017, International Conference on Computing Communication and Automation, and International Conference on Data Engineering and Applications. She is also chairing and coordinating a special track on Applications of Social Media in Security Informatics, along with SOTICS 2017. She also served as a member of the technical committee at various journals (JNCA 2016, Information Systems 2016, IFIP I3E 2016, ISFI 2016, ECIS 2016, ISFI 2015, Egyptian Informatics). She was awarded travel grants from ACM-India, IDRBT- RBI Govt. of India, Anita Borg Institute and Grace Hopper for attending conferences, Ph.D. symposium, and technical meet-ups. During her Ph.D., she presented an invited tutorial at Big Data Analytics conference 2015 held in Hyderabad, India. Her recent publications include International Journal of Web Engineering and Technology (IJWET 2017) from Inderscience Publishers, The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2017), International Conference on Advanced Data Mining and Applications (ADMA 2016) and European Intelligence and Security Informatics Conference (EISIC 2016).

Keywords: social media analytics, open source social media intelligence, text analytics modeling, user-generated data, complaints and grievances, hate and extremism promotion, religious beliefs and conflicts, civil unrest and protest, secret message communication

Abstract

Online Social media platforms such as Tumblr, Twitter (micro-blogging website) and YouTube (video sharing website) contains information which is publicly available or open-source. Open-source social media intelligence (OSSMInt) is a field comprising of techniques and applications to analyze and mine open-source social media data for extracting actionable information and useful insights. The focus of the work presented in this dissertation is on novel applications and techniques of OSSMInt in the government sector. We propose and develop several novel usage scenarios and applications around OSSMInt for government and broadly divide them into three categories: identification, prediction, and response applications. In particular, we present solutions, tools and techniques for analyzing data from micro-blogging website to analyze citizen complaints and grievances in the public sector [response]. The research presented in this dissertation also describes our work on analyzing data from Twitter micro-blogging website to early forecast a civil unrest and protest [prediction]. Furthermore, we build various applications around identification and detection that are useful for the government and security analysts. We demonstrate the application of OSSMInt for identifying religious conflicts within society by mining public opinions on Tumblr website and fill the gaps of offline surveys. The study presented in this dissertation propose solutions for enabling law enforcement agencies to detect, prevent and combat online radicalization and extremism (content, users, and communities) by mining data from Tumblr, Twitter and YouTube [identification]. We also propose to use the deep natural language processing analysis based techniques for automatic identification of racist and radicalized posts based on the intent of the author. Furthermore, we also propose and build an application for detecting secret message exchanged in an adversarial communication and capture the obfuscated terms in messages.

It is technically challenging to analyze social media content due to the free-form nature of user-generated data that raises several issues such as incorrect grammar, spelling mistakes, multilingual scripts, term obfuscation and usage of abbreviation and short-forms. In this dissertation, we present several techniques for data processing, text classification, and word obfuscation detection and information extraction for overcoming the noisy data problem. We also propose computational linguistic-based methods to address the challenges of ambiguity in the textual content. The central component of our proposed solution approach is the application of information retrieval and machine learning based techniques and algorithms. Our study consists of experimenting with a diverse range of machine learning algorithms such as unsupervised, semi-supervised and supervised learning (k-NN, SVM, Naive Bayes, Random Forest and Decision Tree) based algorithms. We also employ several ensemble learning based technique to improve the accuracy and performance of the baseline statistical models. We make the processed dataset used in our experiments publicly available for other researchers to replicate our experiments and benchmark against our proposed techniques. Data visualization is one of the major components of data analysis and interpretation. The study employs several basic and advanced data visualization techniques to present information in an intuitive manner to the end user.

I dedicate this thesis to my brother Abhishek. Thank you for always being there for me and giving me your unconditional support during each phase of my Ph.D. journey. Thank you, Bhai!

Acknowledgments

I take this opportunity to extend my sincere gratitude and appreciation to all those caring and helpful people around me who made this PhD thesis possible.

First and foremost, I would like to extend my sincere gratitude to my research guide and mentor Dr. Ashish Sureka for introducing me to this exciting field of research and accepting me as his PhD student. I would like to thank him for his dedicated help, advice, and inspiration for me to keep going on. His enthusiasm, integral view on research and his mission of providing high-quality work, has made a profound impression on me. It would have been difficult for me to complete this dissertation without his encouragement and immense support. I especially want to thank him for not only playing the role of a PhD guide but also teaching me various aspects of research, PhD life-lessons, and providing me several opportunities and platforms for helping me in growing as a researcher. I am grateful to him for facilitating me a flexible environment for research, doing open discussions as a collaborator and providing critical feedback as a reviewer at the same time. I am fortunate to have you as my advisor.

My special words of thanks should also go to my research collaborator Nitish. The biggest appreciation is when your collaborator wants to work with you on 'one more' project. I would like to thank Nitish for believing in me as a collaborator and working together on multiple research projects. He has been a great team and working with him helped me growing as a mentor as well. I would also like to thank Dr. Denzil for giving me his valuable time in discussing projects and giving me directions on evaluating my research ideas. I would like to acknowledge Nisha for being my first collaborator in my PhD. I would also like to thank Siddharth Dawar for taking his time on discussing with me my research ideas and solutions.

I would like to express my sincere thanks to ACM-India, IDRBT Govt. of India, Grace Hopper and Anita Borg Institute for providing me travel grants and funding to attend and present my research work in the conferences, symposiums, and technical meet-ups. I would like to thank Prof. Kamalakar and Dr. Lini for giving me an opportunity to work as a visiting researcher at IIIT-Hyderabad. I would also like to thank IIIT-Delhi, and it's staff for providing me an environment at the university for conducting my research and helping me in their own way.

I am thankful to Prof. Vikram Goyal, Dr. Donghoon Chang and Dr. Rahul Purandare for being my PhD monitoring committee members and giving me their valuable and timely reviews during yearly evaluations which helped me to shape-up my work at right time. I would also like to thank all my anonymous reviewers for reviewing my work and providing me their value-added feedback helping me in improving my work.

This entire thesis and majority of my papers are written in *ShareLaTeX*. I would like to thank Henry Oswald, Dr. James Allen, and Michael Cribbin for sharing the collaborative editing tool with LaTeXcommunity. This is my pleasure to acknowledge the joy of using *ShareLaTeX* within my research community.

My special regards to my teachers because of whose teaching at different stages of education has made it possible for me to see this day. Because of their kindness, I was able to reach a stage where I could write this thesis.

As always it is impossible to mention everybody who had an impact on this work however there are those whose spiritual support is even more important. I owe my deepest gratitude towards my family for their endless support and understanding of my goals and aspirations. I feel a deep sense of gratitude for my parents and elders who formed part of my vision and taught me good things that certainly matter in life. Their unfailing love and support have always been my strength. Their patience and sacrifice will remain my inspiration throughout my life. I would like to thank my sister Jigyasa and brother-in-law Ashish for always being there for me and helping me, guiding me and motivating me in every possible manner. I cannot express my gratitude towards you in words. I am grateful to my brother Abhishek for being my biggest support during this journey. I want to thank my younger and captain sister Surbhi for being a good listener and teaching me various aspects of life despite being the younger one. I thank each of you for having faith in me and cheering up the most during my success and encouraging me during any downfalls. Without your love and support, I would not have been able to complete much of what I have done and become who I am.

Contents

List of Figures

List of Tables

Thesis Publications

Google Scholars Citations: 97, **h-Index:** 7, **i10-Index:** 4, **Link:** `https://goo.gl/ZfdEew`

Journals

1. **Swati Agarwal**, Nitish Mittal and Ashish Sureka. (2017). A General Overview and Bibliometric Analysis of Seven ACM Hypertext and Web Conferences. International Journal of Web Engineering and Technology (IJWET), Inderscience Publisher. Volume 12. Issue 2.*,†

Peer Reviewed Conferences Publications

1. **Swati Agarwal**, Nitish Mittal and Ashish Sureka. (2017). Potholes and Bad Road Conditions-Mining Twitter to Extract Information on Killer Roads. The 18th International Conference on Web Information Systems Engineering. Springer. **[Under Review]**‡
2. **Swati Agarwal** and Ashish Sureka. (2017, May). Investigating the Dynamics of Religious Conflicts by Mining Public Opinions on Social Media. In Proceedings of the 21st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). Springer.‡
3. **Swati Agarwal** and Ashish Sureka. (2017, January). A Collision of Beliefs: Investigating Linguistic Features for Religious Conflicts Identification on Tumblr. In Proceedings of the 13th International Conference on Distributed Computing and Internet Technology (ICDCIT). Springer.‡
4. Nitish Mittal, **Swati Agarwal** and Ashish Sureka. (2016, December). Got a Complaint? Keep Calm and Tweet it!. In Proceedings of the 12th International Conference on Advanced Data Mining and Applications (ADMA). Springer.‡
5. **Swati Agarwal** and Ashish Sureka. (2016, August). But I did not Mean It!- Intent Classification of Racist Posts on Tumblr. In Proceedings of the 6th European Intelligence & Security Informatics Conference (EISIC). IEEE.‡
6. **Swati Agarwal** and Ashish Sureka. (2016, March). Investigating the Potential of Aggregated Tweets as Surrogate Data for Forecasting Civil Protests. In Proceedings of the IKDD 3rd International Conference on Data Science (CoDS). ACM.‡
7. **Swati Agarwal** and Ashish Sureka. (2015, March). Topic-Specific YouTube Crawling to Detect Online Radicalization. In Proceedings of the 10th International workshop on Databases in Networked Information Systems (DNIS). Springer.‡
8. **Swati Agarwal** and Ashish Sureka. (2015, January). Using Common-Sense Knowledge-Base for Detecting Word Obfuscation in Adversarial Communication. In Proceedings of the Workshop on Future Information Security (FIS) Co-located with Communication Systems and Networks (COMSNETS). IEEE.‡
9. **Swati Agarwal** and Ashish Sureka. (2015, February). Using kNN and SVM based One-Class Classifier for Detecting Online Radicalization on Twitter. In Proceedings of the 11th International Conference on Distributed Computing and Internet Technology (ICDCIT). Springer.‡

---

*The study presented in this article is a part of my Ph.D. work but not included in the thesis chapters.

†We discussed this work in brief in the appendix of this dissertation.

‡The article is included in the thesis chapters.

Peer Reviewed Conference Posters

1. **Swati Agarwal** and Ashish Sureka. (2016, August). Role of Authors Personality Insights for Racist Posts Detection. In Proceedings of the 6th European Intelligence & Security Informatics Conference (EISIC). IEEE.‡
2. **Swati Agarwal** and Ashish Sureka. (2015, May). A Topical Crawler for Uncovering Hidden Communities of Extremist Micro-Bloggers on Tumblr. In Proceedings of the workshop on Making Sense of Microposts (Micropost). co-located with International Conference on World Wide Web (WWW). ACM.‡
3. **Swati Agarwal** and Ashish Sureka. (2014, September). Learning to Classify Hate and Extremism Promoting Tweets. In Proceedings of the Joint Intelligence and Security Informatics Conference (EISIC + ISI). IEEE.‡
4. **Swati Agarwal** and Ashish Sureka. (2014, September). A Focused Crawler for Mining Hate and Extremism Promoting Users, Videos and Communities on YouTube. In Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT). ACM.‡

Comprehensive Examination Report

1. **Swati Agarwal** and Ashish Sureka. (2015). Applying Social Media Intelligence for Predicting and Identifying On-line Radicalization and Civil Unrest Oriented Threats. arXiv preprint arXiv: 1511.06858§†

ACM Newsletters

1. **Swati Agarwal**, Nitish Mittal and Ashish Sureka. (2017) Minority Ethnic Groups in Computer Science Research- What is the Bibliographic Data Telling Us?. In: ACM SIGCAS Newsletter, Volume 47, Issue 2.∗
2. **Swati Agarwal**, Nitish Mittal and Ashish Sureka. (2017) How healthy are ACM SIGWEB sponsored conferences?. In: ACM SIGWEB Newsletter, Issue Spring.∗
3. **Swati Agarwal**, Nitish Mittal and Ashish Sureka. (2016) A Scientometric Analysis of 9 ACM SIGWEB Cooperating Conferences. In: ACM SIGWEB Newsletter, Issue Autumn.∗
4. **Swati Agarwal**, Nitish Mittal and Ashish Sureka. (2016). A Glance at Seven ACM SIGWEB Series of Conferences. In: ACM SIGWEB Newsletter, Issue Summer.∗
5. **Swati Agarwal**, Nitish Mittal Rohan Katyal, Ashish Sureka and Denzil Correa. (2016) Women In CSR: What is the Bibliography Data Telling Us? In: ACM SIGCAS Newsletter Computers and Society, Volume 46, Issue 1.∗

Research Datasets Publications

1. **Swati Agarwal**, Nitish Mittal and Ashish Sureka. (2017, January). Syntactic Enhancement of Killer Road Complaint Tweets Posted on Twitter. Mendeley Data, v1 `http://dx.doi.org/10.17632/dm6s252524.1`‡
2. **Swati Agarwal** and Ashish Sureka. (2016, August). Religious Beliefs on Social Media: Large Dataset of Tumblr Posts and Bloggers Consisting of Religion Based Tags. Mendeley Data, v1 `http://dx.doi.org/10.17632/8hp39rknns.1`‡

§The article is an arxiv report written for the partial fulfillment of Ph.D. comprehensive examination.

3. **Swati Agarwal**, Nitish Mittal and Ashish Sureka. (2016, July). Enhanced Dataset of Citizen Centric Complaints and Grievances on Twitter. Mendeley Data, v1 `http://dx.doi.org/10.17632/w2cp7h53s5.1`‡

4. **Swati Agarwal** and Ashish Sureka. (2016, April). Semantically Analyzed Metadata of Tumblr Posts and Bloggers. Mendeley Data, v2 `http://dx.doi.org/10.17632/hd3b6v659v.2`‡

Chapter 1

Introduction

Online social media (OSM) websites are highly participative and collaborative platforms that enable users to create connections with other people, communicate them by discussing several topics, conveying their thoughts and share information via different mediums [1]. Unlike mainstream media, OSM platforms have low publication barriers which facilitate general users to participate in communication and post content [2]. Visiting a social networking website is the 4^{th} most popular activity on the Internet [3]. As per January 2017 statistics, there are over 3.7 billion active Internet users while about 73% of these users are active on different OSM websites[1]. Based on the type of content allowed on the websites, OSM platforms take different forms. For example personal blogs, micro-blogging services, discussion forums, video sharing portals, image hosting & sharing websites, social networking websites, social bookmarking social gaming, and question-answering websites. Facebook [4] is an online social media and social networking website while Twitter [5], Sina Weibo [6] and Tumblr [7] are the social networking as well as micro-blogging websites. Similarly, YouTube [8], Vimeo [9] and DailyMotion [10] are different video sharing websites. Imgur [11] is an image hosting and sharing website while Flickr [12] and Instagram [13] are both images as well as videos hosting websites. StackOverflow [14] is a question-answering based online community while Reddit [15] is an online social news and media aggregation website. Due to various features like anonymity, low barriers to publication, social networking, and wide reachability among global users make social media websites popular among their users. According to January 2017 statistics, there are approximate 217 active OSM websites [16] while few OSMs are more popular than others. As per January 20, 2017 statistics, Facebook (founded in 2004) is the most popular social networking website on the Internet [17] witnessing 1.8 billion active users, YouTube (founded in 2005) has 1.3 billion users [18], Twitter (founded in 2006) has 313 million monthly active users [19] posting more than 500 million tweets per day and Tumblr (founded in 2007) has 334 million active blogs [20] creating 65 million posts [21] every day. In this Section, we discuss a brief background of the most popular social media platforms that are also studied in the work presented in this dissertation.

YouTube

YouTube is the most popular video sharing and hosting website that allows users to upload content on the website in the form of videos (including small clips and longer videos). YouTube has very low publication barriers, and users can upload content with a valid YouTube account. Figure 1.1 shows the example of a video "social media analytics tutorial" uploaded by the user (referred as uploader on YouTube) "Thomas Joslyn"[2]. Uploaders on YouTube can provide tags, title, and description along with their video to make

[1] `http://wearesocial.com/blog/2017/01/digital-in-2017-global-overview`

[2] `https://www.youtube.com/channel/UCnWf9n47xYsN-NaF_GTyMQg`

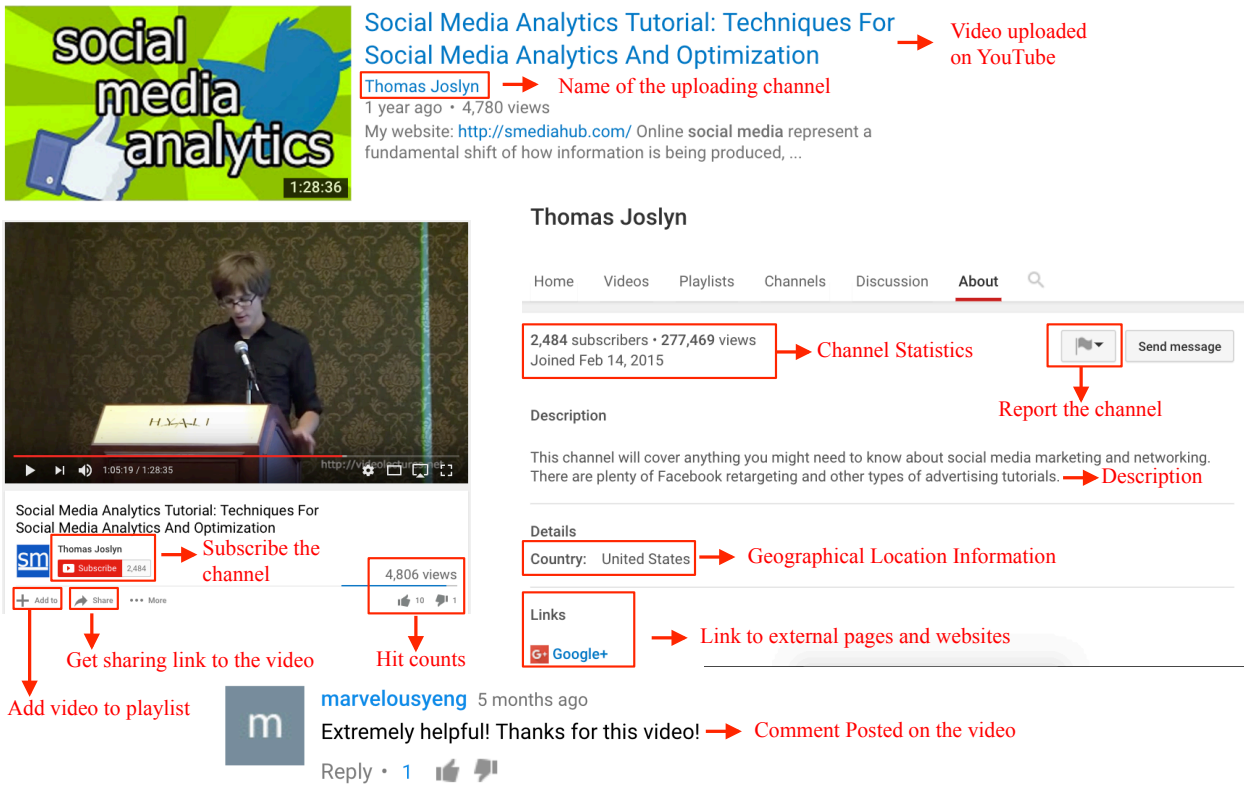


Figure 1.1: Illustrating Snapshot of a Video and Channel Metadata on YouTube Website. Further Demonstrating the Common Activities Performed by YouTube Users on Published Videos (comment on the video, reply to an already published comment, add to the playlist, like/dislike the video) and Existing Channels (subscribe the channel, comment on users' profile, report the channel).

their content easily searchable. Figure 1.1 demonstrate some of the commonly practiced activities that can be performed on the website; e.g. liking or disliking a video. YouTube allows users to add videos to their personalized playlists and share videos on other platforms. Users on YouTube can post comments on already published videos, share their views and connect with other users sharing the similar interests [22]. Users can further connect with each other by subscribing their channels. In order to make their channels easily searchable and find other users with similar interests, YouTube allows them to add a description and profile information about themselves and the channel. Some of this information is publicly available to other users (links to other profiles- Google+, Twitter) while some information can be made private by the channel (location, friends, subscriptions, playlists and activities).

Twitter

Twitter is the most popular online micro-blogging and social networking website. Similar to other online social media websites, Twitter has low publication barriers and allows users to post content in the form of tweets. Unlike YouTube, Twitter primarily allows users to create the textual post and attach multimedia content (images, URLs, and videos) as external entities [23]. Figure 1.2 shows a snapshot of a Twitter profile and various activities that can be performed on the website. The textual posts on Twitter called as tweets

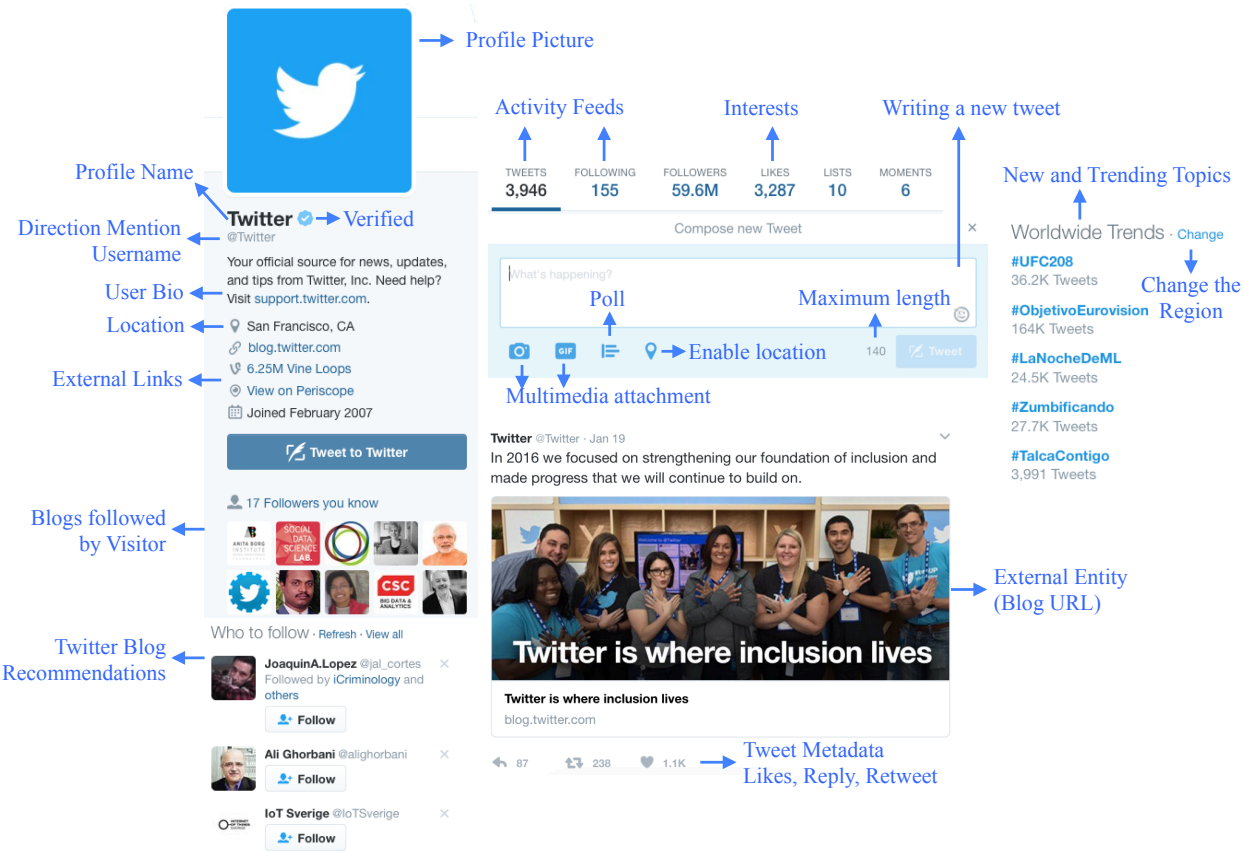


Figure 1.2: Snapshot of a Twitter Profile and Various Activities Performed by the User on the Website. Demonstrating the Various Contextual Metadata and Fields Available with the Tweets (like, re-tweet, reply), User Profile (follow a blog, tweet to a user, similar profiles suggestions) and Twitter Website (search a tag, publish a tweet, trending hashtags).

have a limit of 140 characters and hence called as microposts. While, in YouTube, the tags are provided as additional information, on Twitter, tags are the integrated part of the tweets and called as hashtags (# symbol followed by the keyword). Hashtags in Twitter are the most important and commonly used entity for various purposes; e.g. making the tweets easily searchable by other users sharing the same interest, expressing the opinions, the highlight of the post or concluding phrase for the tweet. Twitter allows users to interact with each other by following (not a mutual relation) their existing blogs on the website. Bloggers on Twitter can like the tweets posted by other users and reply them and share them on their blog with or without adding comments (called as re-tweet). Each blogger on the website is assigned a unique username that can be used to search or send messages to the profile (@ symbol followed by the username). Further, users can add a short bio (160 characters), location and links to their external profiles to make their profiles recognizable to other bloggers sharing same interests. The information can be made both publicly available as well as private.

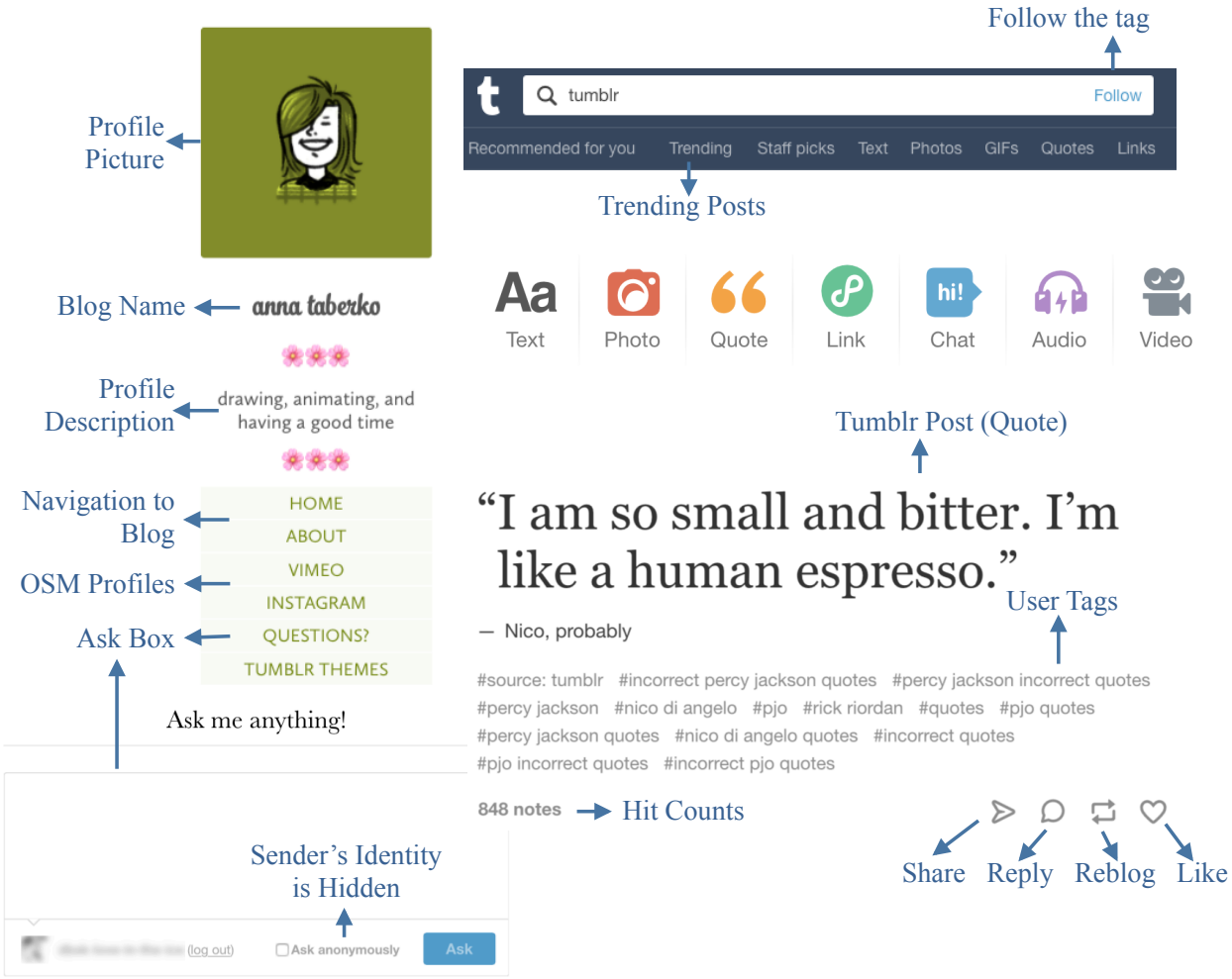


Figure 1.3: Shows the Fields Available on a Blogger Profile and a Published Post on Tumblr Website. Further Demonstrates the Various Activities Performed on the Blogger's Profile (sending message, follow the blogger), Post (like, re-blog, share, add a comment) and Tumblr Website (search a tag, follow/track a tag, trending posts, recommended posts).

Tumblr

Tumblr is the second most popular micro-blogging service commonly used by people of young age group (15-35). Tumblr allows bloggers to post 8 different types of content including images, videos, text, quote, external links, chat posts, audio, and answers (response to the received questions from other bloggers) [24]. Figure 1.3 shows an example of a Tumblr profile (referred as a blog), Tumblr post and search activities that can be performed on the website. Tumblr not only allows users to search for a post using the keywords attached to the post but it allows them to follow multiple tags and receive future updates on the posts associated with those tags. Similar to Twitter, bloggers on Tumblr can like and share a post and can reblog it on their profile with or without embedding comments. To receive updates on their dashboard, users can follow (not a mutual relation) other bloggers sharing the same interests. Further, as the Figure 1.3

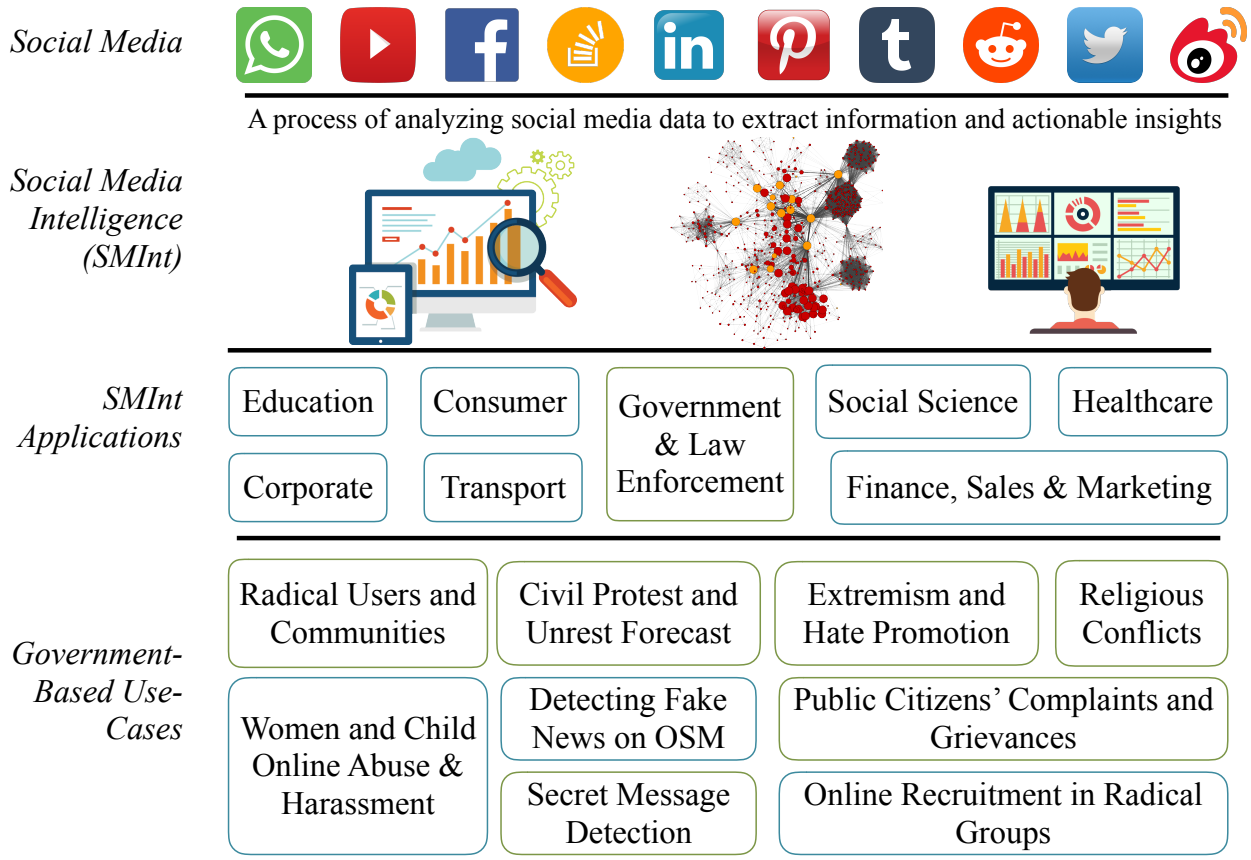


Figure 1.4: High-Level Block Diagram Demonstrating Several Applications of Open Source Social Media Intelligence (OSSMInt). Further, Showing the Applications of OSSMInt for Government and Law Enforcement Agencies.

shows users are allowed to send messages and submit questions while some of the bloggers allow anonymous questions (referred as submissions and ask) as well. Unlike Twitter and YouTube, Tumblr allows bloggers to customize their blogs and create their own themes and add as many pages as they want. Bloggers add these pages and links to make their blogs easy to navigate and provide quick access to the content posted on similar topics and tags. For example, edits, tutorials, reblogs, my_posts, and answers. Similar to YouTube, the list of followers and followings can be made publicly available as well as private.

1.1 Open Source Social Media Intelligence

One of the key features of social media is that the data posted on these platforms is public and accessible as open source data. Though, users are allowed to create private accounts and hide their posts but the majority of the content on OSM websites is public. Facebook, Twitter, YouTube, Tumblr and Sina Weibo are some of the open source OSMs that allow users to extract public data using REST APIs [25] [26] [27] [28] [29] unlike corporate organization websites and personal blogs that are closed-source platforms. Social Media Intelligence (SMInt), also known as social computing and social informatics is a domain comprising of applications built upon intelligence collected and inferred from social media data using various data

mining, machine learning, information retrieval, and text analytics techniques [30]. Open Source Social Media Intelligence (OSSMInt) is a sub-field within SMInt that focuses on extracting useful information and actionable insights from publicly available and overt sources of data on social media platforms [31]. Research shows that the Internet and social media websites have facilitated users to publish anything that they see or feel. Such publications (user-generated data) can be considered as the readings of "human-sensors" about the physical world that we sense [32]. There are several applications that can be built by applying OSSMInt techniques on this human-sensor data. For example, applications in education [33], transport [34], corporate sector [35], social science [36], government [37] and healthcare [38] domains. The aim of this thesis is to build applications of OSSMInt that are useful for the government and law enforcement agencies. Developing applications to enable such usage scenarios requires solving several technically challenging problems which needs research and technology advancement. Current tools and techniques are not enough to fully address the applications defined by us and the gaps between what is available and what is required forms the motivation of our work. Figure 1.4 shows a high-level presentation of some of such applications. Figure 1.4 shows a few applications proposed and implemented by us as well as the few applications developed by other researchers. For example, religious conflict detection and public complaints and grievances mining from micro-blogging websites are some of the novel contributions of our worl. Based on the types of social media intelligence required in building these applications, we divide them into three broad categories: 1) Identification, 2) Prediction and 3) Response applications.

1. **Identification Applications** are the detection-based applications that extract social media content (contextual and linguistic) for identifying relevant information. Such applications apply text analytics, classification, and machine learning based techniques on data for identifying unusual and hidden information from the data. For example, finding emerging topics [39] and high-quality content [40] on social media. Other identification applications include spam detection [41], sentiment analysis [42], and profile cloning detection [43]. Identification applications for government include applying OSSMInt for detecting the sensitive or harmful content which is useful as well as important for the government. For example, discussion on the topics of race and religion, online abuse and harassment, sharing secret messages, hate promotion, users, and communities spreading extremism among other users on social media.

2. **Prediction Applications** uses data mining and text analytics based approaches for an early forecast of events. For example, using text mining and predictive machine learning approaches for analyzing financial news articles for stock market prediction [44], mining online users' communication for predicting natural hazards and disasters [45], and mining online polls and user-generated data for predicting election results [46]. Prediction applications for government include the applications that are useful for security analyst, law enforcement agencies, and local civil force. For example, mining online social media for an early forecast of mobilization of civil protests and riots in a city. These applications include extracting and mining users' communications and online news media for predicting the location (where the protest is going to happen), time (when the protest is going to happen) and density (how many people are going to participate the riot) of the event.

3. **Response Applications** include usage of identification applications for detecting useful content and further responding to the authors of those posts. For example, gathering information from online review sites and blogs for mining consumers' opinion on various products and services [47]. Response applications for government include the identification of complaints and grievances reported online by the public citizens for the government authorities and facilitate public agencies to response those complaints, provide better services and improve their connections with public citizens.

1.1.1 Technical Challenges

The task of mining open source social media data and building the identification, prediction, and response (IPR) based applications is technically challenging. Challenges like missing information, unstructured data, and multilingual content contaminate the accuracy of natural language processing techniques. Further, due to the free-form nature of social media text, the user-generated content (UGC) is highly likely to have spelling and grammar mistakes. Lack of ground-truth data, the presence of imbalance data and ambiguity in the intent of authors make building these applications challenging. We discuss each of these challenges below:

1. **Massive Size, High Velocity and Rich User Interaction:** The volume and variety of data in terms of modality such as free-form text, images, audio and video generated every day on social media are huge. The massive size of the data poses hard computational challenges such as data processing and storage for researchers or application developers interested in analyzing the data [30]. In addition to the massive data size, the rich user interaction (such as friends, followers, subscribers, like and reblog relationship) possible in social media increases data complexity, dimensions, and variability that needs to be addressed. Further, in addition to the overall quantity, the speed (for example few megabytes per second or millions of rows per hour) at which data is generated poses data analytics challenges. Real-time processing and storing of such high-velocity of data is computationally challenging from the perspective of data ingestion [48].
2. **Imbalance Data:** Due to the massive flow of data being uploaded on social media portals, the relative percentage of positive class records is very small (0.000001%). Therefore, to classify the relevant content (such as extremist content, public complaint, religion based posts) the experimental dataset becomes highly imbalanced and causes a large number of false positives.
3. **Intent Mining:** Social media provides a platform for users to interact with each other by sharing and consuming information. Exchange of this information reflects a variety of intentions. For example, information sharing, query, promotion, sarcasm, humor, and negative sentiments. However, only a very small percentage of data contains explicit intent or a specific meaning [49]. In the domain of online radicalization detection, religious conflict identification it is highly important to understand the psychology and the intent of extremist authors. Mining relevant intent from ambiguous, unconstrained and natural language microposts is technically challenging.
4. **Multilingualism:** The web and social media are inherently multilingual. Content is posted in several different languages and processing them through automated algorithms requires linguistic resources for each language. Mining multilingual information is important and even essential for several government-based application due to the global diversity, user demographics, and wide reachability of the social media platforms and its users.
5. **Noisy Content:** Due to the low publication barriers, a huge amount of content on social media is of low quality (such as spam) and in general of low relevance or importance (such as posting a message on what one had for breakfast or lunch) [50]. While the relevance of content is contextual and based on the topic of discussion, certain content is of low relevance and importance irrespective of the context. Moreover, there are several issues such as grammatical mistakes, spelling errors, usage of non-standard acronyms and abbreviations, emoticons and incorrect capitalization of terms due to the informal nature of the content.
6. **Spam and Fake Accounts:** Spam, irrelevant and unsolicited messages as well as the fake accounts are common in social media platforms [51]. Such content not only decreases the value of the website and the user experience but also poses technical challenges for social computing researchers in terms of data cleaning and pre-processing before building accurate data mining models.
7. **Data Annotation and Ground Truth:** Creating data annotation and ground-truth is a basis for several machine learning tasks such as predictive modeling and classification. Creating and annotating

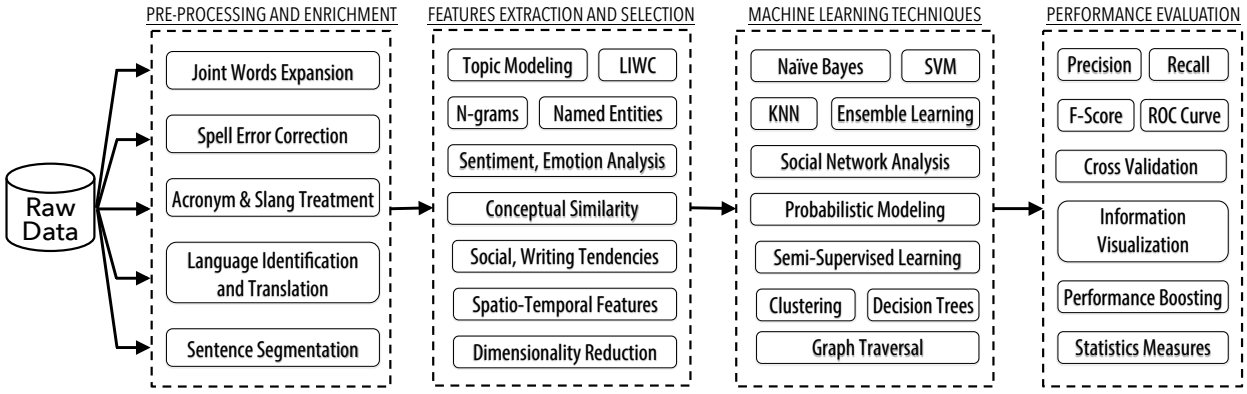


Figure 1.5: A General Framework Illustrating Various Data Mining and Information Retrieval Techniques used in my Work, Primarily Divided into 4 Categories- 1) Data Pre-Processing, Enhancement and Enrichment Techniques, 2) Features Extraction and Selection Techniques, 3) Machine Learning Techniques for Classification and Network Analysis and 4) Performance Evaluation Metrics

high-quality ground-truth data at a large scale is effort intensive if done manually and a non-trivial technical challenge if done (semi)-automatically [52].

8. **Manipulation, Fabrication and Adversarial Behavior:** Deception, manipulation, misinformation, adversarial behaviors, and credibility is a major issue in social media. Research shows that a lot of content posted on social media is factually incorrect and is rumor [53]. Fake information, rumors, and manipulated content are not only a social media abuse but challenges for researchers and developers in building computational frameworks for Intelligence and Security Informatics (ISI) based applications. Further, it is impossible to verify a rumor, fake news, and secret message communications due to the lack of facts and ground truth.

9. **API Limitations:** As discussed above, open source OSM platforms provide Application Programming Interfaces (APIs) for developers to query website and extract open source data. Some of the private information on user and data can be accessed using open authorization (OAuth) [54] on the website. However, extracting each and all public information using API is another limitation. Some of the open source social media RESTful (Representational state transfer) APIs like Twitter, Tumblr and YouTube allow developers to extract all available information on the website while APIs like Facebook and Instagram allow to fetch only limited public information. The API limitations and restrictions further increase the disproportion of imbalance data and quality of extracted content.

1.2 Thesis Contributions

In this thesis, we identify several novel applications on how social analysts in government or law enforcement agencies can leverage the large volumes of social media data and extract useful insights from them. Research shows that there are several existing machine learning and data mining techniques for analyzing social media data and gaining insights from it. However, there are significant research gaps between the available technologies and required methodologies for building the government-based IPR applications. In order to fill these gaps, it does not only require the algorithmic contributions but also requires the scientific discovery and technology advancement. In addition to identifying these problems, this thesis also proposes

novel methods and technologies to build such applications. Figure 1.5 shows a block diagram framework demonstrating several data mining and machine learning applications used in our thesis to enrich raw experimental datasets, extracting features and performing classification and prediction tasks for building desired applications. The central component of our proposed solution approach is the application of information retrieval and machine learning based techniques and algorithms. This thesis consists of experimenting with a wide range of machine learning algorithms such as Support Vector Machines, Naive Bayes, Random Forest and Decisions Tree. We also employ several techniques such as ensemble classifiers to boost the accuracy of the baseline statistical models.

The key and specific contributions of our work presented in this thesis are as follows:

1.2.1 OSSMInt Applications in Government Domain

1.2.1.1 Mining Public Citizens' Complaints and Grievances [55] [56]

Given a public post on Twitter, we identify if the tweet is a complaint and grievance report or not. Studies in previous literature are limited to conducting empirical analysis on online and offline data and examine the use of Twitter for client feedback and sentiments about the services provided by government agencies. Whereas, some studies are focused towards mining social media communications for building predictive models for situation awareness. We, however, analyze the tweets reported directly to the official Twitter handle of public service agencies and government accounts. We propose a text analysis based ensemble classifier for identifying the complaints and grievances reports on Twitter. Further, due to a variety of departments under one public service, the complaint reports have a diverse range of topics. Therefore, to identify the low-level details about the complaint, we propose to use contextual and linguistic features that indicate the relation between a complaint and concerned department. Since, the public agencies' account on Twitter are public and anyone can directly mention them in their tweets it is not necessary that every report is a complaint report. We divide such reports into 3 categories: appreciation (tweets published to praise and appreciate the work done by a particular public service or addressing the previous complaints), information sharing (sharing news articles and external information for awareness) and promotional (advertisement of policies and promotional events by other official Twitter handle of similar public agency) tweets. We identify features that are strong indicators of differentiating a complaint report from such non-complaint tweets.

We observe that due to diverse nature of public services and type of complaints, several specific categories of complaint reports require different structure for reports. We extend the similar idea of complaints and grievances identification for identifying such reports. We conduct a case-study on the complaints reports posted on Twitter about the poor conditions of roads (referred as killer roads) such as road irregularities, dysfunctional facilities, and potholes causing accidents, risk, discomfort and poor experience to public citizens. In order to identify such complaints, we investigate the efficacy of spatial (geographical location metadata) and linguistic features to discover insights from less informative complaints on killer roads. Furthermore, due to the free-form nature of social media and user-generated text, the tweet report are full of noise and poses several challenges for text analyst and NLP researchers. We address the challenges of free-form text in tweets and capture the dependencies between noisy text and semantics. In our thesis, we make an algorithmic contribution by proposing a micropost-enrichment algorithm that is generalized and enhances the raw tweets syntactically as well as semantically.

1.2.1.2 Religious Conflicts Identification [57] [58]

The powerful emergence of religious faith and beliefs within political and social groups, now leading to discrimination and violence against other communities has become an important problem for the government and law enforcement agencies. In this thesis, we propose another application of OSSMInt that online

social media can be used for detecting religious conflict within society. Current state-of-the-art reveals that over the past 3 decades, social science researchers have been conducting offline surveys for identifying religious disputes. While, the immense amount of data available on social media in the form of comments, communications, discussions has been largely ignored in existing works and raising three major limitations: 1) subjectivity of opinions in the surveys, 2) generalized claim of conflicts and 3) identity of people participating in the surveys. We address the challenges and limitations of offline surveys and propose an approach to leverage the power social media for identifying religious conflicts within society. We conduct a survey among a random group of people who do not use social media and blogs and people who actively participate in social media activities. We show them various posts created around the topic of religion and race and based on the survey, we define 11 categories of opinions that can identify the contrast of conflict in religious posts. As stated above, the social media data can be used as the readings of human sensors about the physical world that we sense or feel. Therefore, we identify the various linguistic features from social media posts that reflect the emotions and personal opinions of people while publishing the post. We decode these linguistic features using machine learning and natural language processing techniques. We further propose a text-analysis based model for classifying social media posts into these dimensions. We conduct our experiments on Tumblr micro-blogging website and get our data annotated by the people who have a 2 to 3 years of experience of using the website. In order to address the challenge of creating ground-truth for a large-scale data, we propose a multiclass semi-supervised classifier that actively learns from small-scale training data and classify unlabeled posts. We investigate the efficacy of our approach across various dimensionality reduction techniques *e.g. principal component analysis, attribute selection correlation* and evaluate our performance in form of classifier accuracy.

1.2.1.3 Identification of Extremist Content, Users and Communities [48] [59] [50]

We observe that the objective of publishing such religious and race-targeted posts on social media is not only for expressing their opinions but some of these posts are also created with a common agenda of promoting hate towards a targeted community or an individual. Such group of like-minded people takes the leverage of social media and freedom-of-speech to post extremist content and promote their ideology. Due to the immense popularity of social media, the presence of such content on these websites is a major concern for the law enforcement agencies and security analysts. Current state-of-the-art reveals that text classification (automatic and semi-supervised learning), clustering (unsupervised learning), exploratory data analysis and keyword based flagging approaches are commonly used for identifying extremist content on social media. While, link analysis technique is used for crawling through navigation links for identifying similar users and locating their hidden communities. We investigate the application of topical crawler-based approach for uncovering the extremist bloggers and communities on social media platforms. The specific contribution of this work is to demonstrate the effectiveness of contextual metadata such as the body or description, tags, and caption (or title) of a post in identifying radicalized content. We conduct a series of experiments on three different social networking platforms- Twitter, Tumblr, and YouTube. We conduct our experiments on Twitter since Twitter is the most popular micro-blogging website. Videos are the easiest and most effective way to share and express opinions and thoughts. Therefore, we conduct our experiments on YouTube which is the most popular video hosting and sharing website. Tumblr is another most popular micro-blogging service that facilitates many such features which are not there in other social networking websites. For example, tracking of tags and keywords, sending anonymous messages, creating descriptive posts, and post a wide variety of multimedia content. We propose several text-analysis-based models (n-gram model, one-class SVM, and KNN classifiers) for classifying extremist posts and measure their performance using standard metrics of information retrieval. We further demonstrate the effectiveness of various relationships among users *e.g. like, reblog, follower, subscription* and graph traversal algorithms for identifying the hidden communities. We use social network analysis for capturing the insights about major influencers and

core users in the community.

We further extend the idea of radicalized and racist posts identification to include not only the content of the post but also the intention or objective of the author. The prior studies in literature mostly use quantitative text analysis and keyword-spotting based approaches for detecting persuasion behavior and differentiating radical groups from non-radical groups. However, the dependency on the user behavior and authors personality traits has not yet been considered. We propose a hypothesis that the authors' personality traits are the strong indicators of the intent of the author and can be used to discriminate the posts having malicious intent from naive posts. We contribute to the current state-of-the-art of intent classification of racist posts by 1) using natural language processing techniques for identifying the focused and targeted topic of posts while the topic related key terms are missing from the content, 2) performing sentiment enrichment on the posts and emphasizing on target-specific emotions *e.g. anger, disgust, sadness, joy, fear*, social tendencies (personality traits of the Big Five personality theory) *e.g. openness, conscientiousness, extraversion, agreeableness* and *emotional range* and author's language & writing cues *e.g. analytical, confident* and *tentative* style of writing and 3) identifying the semantic role of each term present in the content and identifying the hidden phrases playing major role in the post.

1.2.1.4 Term Obfuscation Detection in Adversarial Communication [60]

Due to the constant monitoring and surveillance by Intelligence agencies on social media, intercepting mail, mobile phone and satellite communications terrorist and criminals use textual or word obfuscation to prevent their messages from getting intercepted by the law enforcement agencies. Textual or word substitution consists of replacing a red-flagged term (which is likely to be present in the watch-list) with an "ordinary" or an "innocuous" term. For example, the word attack being replaced by the phrase birthday function and bomb being replaced by the term milk. Research shows that terrorist use low-tech word substitution than encryption as encrypting messages itself attracts attention. Prior studies in the literature use probabilistic or distributional models, Sentence Oddity Measures (SMO) and Pointwise Mutual Information (PMI) based approaches for identifying out-of-context terms in a given sentence. However, the major limitation of the existing approaches is that they are able to predict the suspicious sentence only if the first noun of the sentence is substituted and therefore, the substituted term is already known to the analysts. We instead consider a sentence as a bag of words and investigate the application of ConceptNet, a lexical resource and commonsense knowledge base for identifying any term that has been substituted in the sentence. We frame the given problem as a textual reasoning and context inference task and utilize ConceptNet knowledge base to compute the conceptual or semantic similarity between any two given terms and define a Mean Average Conceptual Similarity (MACS) metric to identify out-of-context terms. We make our proposed approach generalizable by conducting experiments on three different open source datasets of different writing structure.

1.2.1.5 Social Media as Sensors for Predicting Civil Unrest Events [61]

In addition to the previous usage scenarios, we propose another application of OSSMInt for government where real-time identification of mobilization and planning of events can be used for building a forecasting model for civil protests. In countries like the USA, Australia and India where protests are legal, early detection or prediction of such events is valuable for government, tourism and law enforcement agencies. Existing studies use keyword-based flagging, probabilistic models, clustering, named entity recognizers, logistic regression, and dynamic query expansion as the conventional techniques for predicting upcoming protest events. However, the existing studies are conducted on the data posted during the event or after the event has happened increasing the possibility of bias in the dataset. Further, due to the high velocity and massive size of data uploaded on social media data, mining each and every post for building predictive model impedes the performance of the model. We address the challenge of noisy content present in real time stream data by performing a content-based characterization and semantic enrichment on raw tweets. To address the

limitations of prior literature, we use trend analysis based approach (captured along the sliding window) and collect data posted 7 days prior to the event. We build a one-class text classification and analytics model for classifying crowd-buzz & commentary and mobilization & planning microposts related to a protest or civil disobedience. We extract linguistic features from the posts classified as a topic related (crowdbuzz and mobilization posts) and discard all irrelevant posts. We investigate the efficiency of spatiotemporal features and named entities for predicting a civil protest event. The key contribution of this work includes predicting the location (where the protest is going to happen), time (when the protest is going to happen). In order to evaluate the performance of our approach, we conduct our experiments on two civil protests happened in Australia and USA.

1.2.2 Publishing Research Datasets

Prior literature shows that the large volume of scholarly articles and research requires the technology advancement for managing and analyzing the content of the literature in an efficient manner. However, a major limitation of the academic research is the unavailability of public research datasets for comparing proposed methods with the existing techniques [66]. Research shows that sharing and reusing base research datasets increases the efficiency and quality of research. Sharing of research datasets within the community helps to capture errors, training new researchers and discourage duplicate and fraud data collection [67]. Due to the larger scientific community benefits, datasets has now become a significant part of the scholarly records, and publication of datasets has become a good practice among the researchers, either formally (linking with their journal articles, citing the datasets in their articles) or informally (sharing them on their personal homepages) [68].

In social media computing community, conferences like ICWSM[3] have made efforts in the direction of sharing datasets and encouraging researchers to make their results available for comparison and benchmarking. Earlier, the datasets are shared on the personal homepages of researchers which is hard to access. Therefore, several websites like KDnuggets[4], CrowdFlower[5], data challenges such as KDD-Cup[6] encourage researchers to publish their data more and more frequently. In social computing community, there are several observatories under WebScience Trust[7] that publish social media research datasets. Some of these observatories are Observatory on Social Media[8] by Truthy group at Indiana University, COSMOS[9] by Social Data Science Lab at Cardiff University, Stanford Network Analysis Project (SNAP)[10] and Southampton WebObservatory at Southampton University [11]. However, these datasets are not linked to the related journal or conference articles and are also not citable. Whereas, various data hubs and data hosting platforms like The Dataverse Project[12], Machine Learning Dataset Repositories[13], Mendeley Data[14] and OpenML[15] facilitates researchers to make their dataset recognizable within the community and citable in research articles by generating a unique DOI for each share.

[3] http://www.icwsm.org/data/
[4] http://www.kdnuggets.com/datasets/index.html
[5] https://www.crowdflower.com/data-for-everyone/
[6] http://www.kdd.org/kdd-cup
[7] http://www.webscience.org/web-observatory/
[8] https://truthy.indiana.edu
[9] http://socialdatalab.net
[10] http://snap.stanford.edu
[11] https://webobservatory.soton.ac.uk
[12] http://dataverse.org
[13] http://mldata.org
[14] https://data.mendeley.com/datasets
[15] https://www.openml.org

As discussed in Section 1.1.1, collecting dataset and creating ground truth for security informatics domain applications is technically challenging. One of the major contributions presented in this thesis is that in order to conduct our experiments all datasets that we created, we make them publicly available in an organized format. Instead of uploading the raw and unstructured data, we pre-processed the collected data and organized it in a tabular form and shared it online on data hosting website. We published our datasets on Mendeley Data and generated the DOI links for them. Table 1.1 shows the list of research dataset collected for the purpose of experiments and applications presented in this dissertation. To make our results available for benchmarking, we also share the extracted features and enriched metadata along with the raw dataset. Agarwal et al. [64] consist of enhanced dataset (C3GT) of tweets posted on 4 Indian public agencies' accounts on Twitter (@RailMinIndia, @dtptraffic, @DelhiPolice and @IncomeTaxIndia) collected for extracting information on public citizens' complaints and grievances. In extension to C3GT dataset [64], Agarwal et al. [62] contains syntactically enhanced dataset of complaints reports posted on the official Twitter handle of Union minister and ministry of road, transport, and highways, the government of India (@MORTHIndia and @nitin_gadkari). The published datasets were collected to conduct experiments presented in Chapter 2. Agarwal et al. [65] consist of the largest dataset (RBSM) of Tumblr posts and bloggers collected by using web-crawling on the website. The published dataset is collected for the purpose of identifying religious conflicts on Tumblr website, and therefore each post contains at least one tag which is frequently used in religion based posts. The study conducted on this dataset is discussed in the Chapter 3. Agarwal et al. [63] consist of the semantically analyzed metadata of Tumblr posts and bloggers collected using bootstrapping method on Tumblr website. The published dataset is the first Tumblr dataset shared publicly and collected for the purpose of identifying intent based racist and radicalized posts (presented in Chapter 4).

1.3 Organization of Thesis

The rest of this thesis is organized as follows. We discuss each contribution in the subsequent chapters. Chapter 2 propose text analytics based solution approaches to enrich social media text and build response based applications. We then discuss the performance of proposed approaches to two case studies of response based applications by mining public complaint and grievances reports on Twitter. In Chapter 3, we propose an approach to overcome the challenges of offline surveys and leverage the power of social media for identifying religious conflicts within society. In Chapter 4, we present an in-depth and comprehensive study on online extremism detection. Chapter 4 proposes several machine learning, text analytics, and social network analysis for identifying radicalized content, users, and communities. We then discuss the implications of Big Five Factor Model in social informatics for intent-based identification of radicalization and racist content. In Chapter 5, we present an approach to detect secret messages communicated between terrorists and decode obfuscated term in the message. We propose to use a commonsense knowledge base for computing the conceptual and semantic similarity between the words and capturing the out-of-context term in the message. In Chapter 6, we present an ensemble identification and prediction based modeling based approach for improving the accuracy in forecasting civil protests and unrest. In Chapter 7, we give a brief summary of our thesis concluding with the general and specific takeaways and discuss the future directions of the work. In the end, we provide an overview of two of our studies as the Appendix (Chapter A) which are conducted during the Ph.D. education program but are not a part of the dissertation chapters.

Table 1.1: Shows a Summary of Research Datasets Collected for the Purpose of Experiments and Case-Studies Presented in this Dissertation. The Listed Datasets are also Available on Mendeley Data Published Online for the Research Community.

Dataset	**KRCT- Killer Road Complaint Tweets [62], Platform: Twitter**
Method	Real time data extraction using Twitter Search API
Metadata	Tweets (post text, enriched hashtags, multimedia attachment, @username mentions), users (locations, verified account)
Size	$81,304$ raw tweets, $3,302$ unique enriched tweets, $2,604$ unique users
Description	consists of tweets collected in a time span of 4 weeks (from July 18, 2016 till September 13, 2016) posted to two public agency accounts of Govt. of India: @MORTHIndia and @nitin_gadkari. The published dataset is enriched syntactically using hashtag expansion, spell error correction, sentence segmentation, @username expansion, slang conversion.
Dataset	**SETP- Semantically Analyzed Tumblr Posts [63], Platform: Tumblr**
Method	Bootstrapping using Tumblr Search API
Metadata	Posts (title, description, type, tags, uploader, notes- like and reblog), blogger (name, title, description, number of posts)
Size	$3,228$ raw Tumblr posts, $2,456$ unique enriched posts, $2,224$ unique bloggers, and $10,217$ unique tags
Description	contains the Tumblr metadata of labeled posts and bloggers collected via bootstrapping method for the purpose of racist and radicalized intent identification. The dataset also contains various features (semantic tagging, emotions, language cues and author personalities) extracted after semantically analyzing the textual post.
Dataset	**C3GT- Citizen Centric Complaints Tweets [64], Platform: Twitter**
Method	Real time data extraction using Twitter Search API
Metadata	Tweets (post text, enriched hashtags, multimedia attachment, @username mentions), users (locations, verified account)
Size	$3,700$ annotated and enriched tweets
Description	contains the public complaint tweets posted on 4 public service accounts of Indian Government (@RailMinIndia, @IncomeTaxIndia, @DelhiPolice and @dtpTraffic) collected in a time span of 4 weeks (11 April 2016 to 8 May 2016). In this dataset, we also share a sample of tweets enriched using hashtag expansion, spell error correction and slang expansion.
Dataset	**RBSM- Religious Beliefs on Social Media [65], Platform: Tumblr**
Method	Timestamp and keyword based web-crawling using Tumblr Search API
Metadata	Posts (title, description, type, tags, notes, like, reblog), Notes (note type, timestamp, blogger id), blogger (name, title, description, number of posts)
Size	107,586 raw posts, 89,803 unique and process posts
Description	contains 8 different types of Tumblr posts consisting of tags commonly used in religion based posts. The shared dataset consists of linguistic and contextual metadata of Tumblr posts, bloggers, tags and Notes available on Tumblr website since 2007.

Chapter 2

Public Services in India- Mining Twitter to Extract Information on Public Complaints and Grievances

2.1 Introduction

Recently, there has been an increasing trend and adoption of social media by government organizations for not just disseminating information but also collecting information such as complaints and grievances from citizens (a phenomenon referred to as *citizensourcing*) [69] [70] [71]. In particular, social media platforms like Facebook [4] and Twitter [5] are gaining popularity as social-media based grievance management system or platforms on which citizens can lodge complaints. Twitter is one of the most widely used micro-blogging websites on Internet. Due to the wide reachability and connectivity among its users, Twitter is being used by the National Government to reach out the public. For example, in India, ministry of the railway (@railminindia), state police (@delhipolice and @mumbaipolice), ministry of road and transport (@morthindia), traffic police (@dtptraffic) and income tax department (@incometaxindia) have some of the most active Twitter accounts. Similarly, department of children & youth affairs press office account (@DCYAPress), department of health (@roinnslainte) are two of the official Twitter handles of Ireland government. Official channel for Prime Minister Theresa May's office (@Number10gov) and national government (@GOVUK) are two most popular Twitter handle of Government of United Kingdom. However, unlike other countries, one of the primary objectives of Indian Government's Twitter accounts is to not only reach out to the public but to also address their complaints and grievances [1,2]. It is seen that sometimes users also attach pictures and videos, *e.g, the location of the incident or showing the severity of issue reported* in their reports providing evidence to their complaints. Based on our analysis of several Indian Government Twitter accounts, we find that an active Government Twitter handle receives an approximate of 5 tweets per minute. Based on our analysis of several Indian governments' Twitter handle data, we found that 50% of the tweets posted in an hour are complaints and grievances reported from various regions of India. The government bodies on Twitter forward these complaints and redirect authors to the concerned department for resolving their complaints efficiently. Based on our inspection on Twitter website, we observe several other complaints on Twitter that are posted

[1] http://www.thehindu.com/business/Industry/government-to-introduce-twitter-seva-for-startups/article8483730.ece

[2] https://blog.twitter.com/2016/modi-s-government-is-transformingindia-through-twitter-in

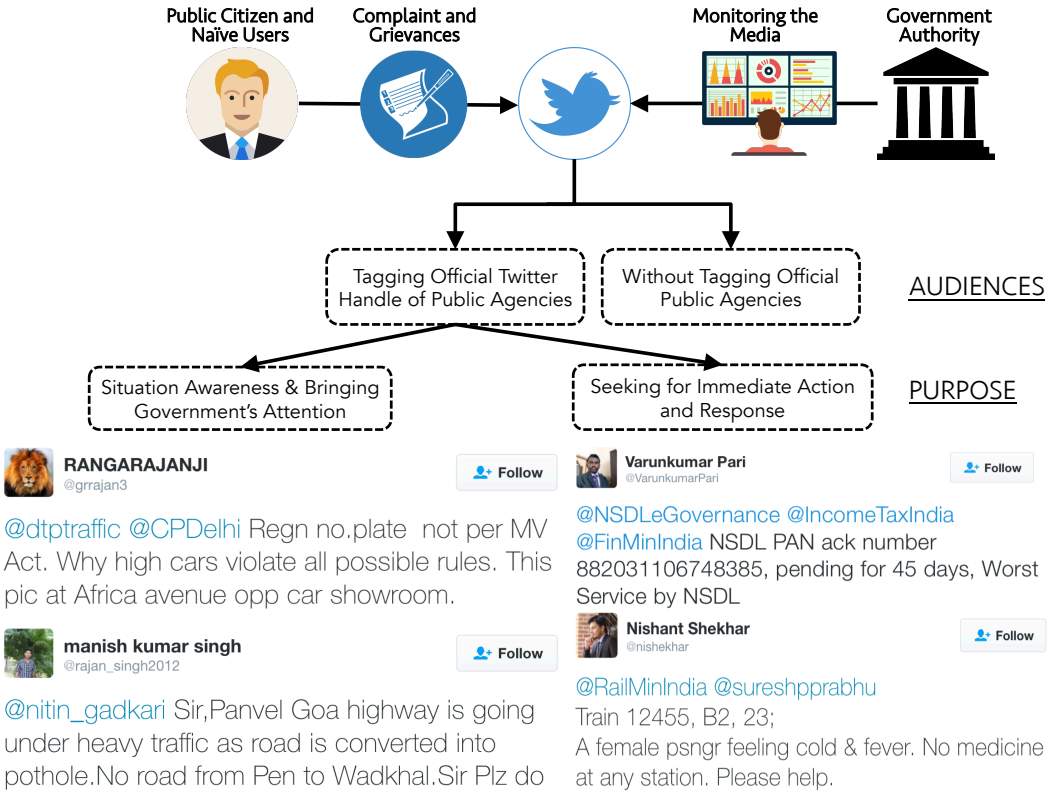


Figure 2.1: High-Level Block Diagram of Usage of Twitter Micro-blogging Website by Public Citizens for Posting Different Types of Complaints- 1) Bringing Attentions of Government to the Issues and 2) Seeking Immediate Response on Complaints

without mentioning the official handle of Government authorities. Further, the complaints reported directly to the public agencies have different purposes and audiences. Based on the content and targeted audiences of these reports, we find following two types of complaints and grievances tweets:

1. complaints which are reported to spread awareness among other citizens and to bring government's attention to the issues reported in the complaint.

2. complaints which seek for immediate action and response from the concerned authorities.

Figure 2.1 shows the high-level block diagram representation of usage of Twitter by public citizens for posting different types of complaints reported for different purposes. Figure 2.1 reveals that in the complaint ("@dtptraffic @CPDelhi Regn no.plate not per MV Act. Why high cars violate all possible rules. This pic at Africa avenue opp car showroom.") reported by Rangarajanji (@grrajan3), the author mentioned the incident violating the traffic rules while it is not affecting the person directly, but the purpose of the complaint is to bring the attention of concerned department and resolve the issue accordingly. Whereas, in the tweet ("@RailMinIndia @sureshpprabhu Train 12455, B2, 23; A female psngr feeling cold & fever. No medicine at any station. Please help.") posted by Nishant Shekhar (@nishekhar), the author requests for an emergency medical assistance and hence seek for an immediate action. However, despite the objectives of the complaints and presence of direct mention to their official Twitter handle, it is important for the government and state authorities to mine each and every complaint posted on Twitter and address on time. Nevertheless, the manual analysis and gaining insights from tweets is infeasible due to the high velocity and volume of the tweets posted on the website.

Further, to manually identify the type of the reports and resolve them, many complaints remain unaddressed.

The research work presented in this chapter is motivated by the need to develop a solution to automatically resolve the challenges of manual inspection. Twitter allows users to post a maximum of 140 characters and therefore involves usage of slang and abbreviations. Due to the presence of free-form text tweets do not have a defined structure or language format and hence are high likely to have grammar and spelling errors. As discussed in Chapter 1, the presence of multilingual texts and scripts in tweets, it is challenging to identify the linguistic features for building NLP (Natural Language Processing) based applications. Text classification or categorization and information extraction from tweets is thus a technically challenging problem. Further, filtering these complaint reports from non-complaint tweets is technically challenging due to the wide range of complaints. Our research aim is to build a text analysis based model to address the NLP challenges in microposts (very short text such as tweets). In particular, the research aim of the work presented in this chapter is to investigate text classification based techniques for automatically identifying complaints tweets and assigning them to predefined labels based on the topic of the content. Our research aim is also to investigate information extraction and visualization to extract useful information and insights from the complaint reports. Furthermore, our aim is to create an annotated dataset and make it publicly available to the research community.

2.2 Related Work

In this Section, we discuss the closely related work of prior research in context to work presented in this chapter. We conduct a literature survey in the area of mining Twitter data for identifying complaints and reports on social media websites. Based on our survey, we find that while there has been a lot of work in the area of mining Twitter for identifying product and services based consumer insights and opinions; the field of automatic identification of citizen complaints is a relatively unexplored area. Further, based on the types of complaints reported on the website and proposed techniques, we divide our literature survey into three lines of research. We discuss the closely related work to work presented in this chapter in following subsections:

2.2.1 Usage of Twitter Micro-blogging Website for Reporting Complaints and Grievances

Heverin et al. [72] examine the use of Twitter by city police departments in large U.S. cities (cities with populations greater than 300,000) that have active Twitter accounts. Their analysis reveals that city police departments use Twitter to converse directly with the public and news media to disseminate crime and incident related information. Anderson et al. [73] present a study on Twitter adoption across American municipal police departments serving populations over $100,000$. Their analysis reveals that there is a strong adoption of Twitter in regional government. However, there is a weak support from the organization for addressing the complaints in comparison to the need and demand from the public. Meijer et al. [74] present an empirical analysis of Twitter usage by the Dutch police and conduct an analysis of 982 Twitter handle. Their study reveals that mostly Twitter is being used for external communication by the police officers, but the mutual interest of other police officers is substantial. Edwards et al. [75] present a study on webcare, i.e. the act of engaging in online communication with citizens. They investigate 4 cases of webcare of Dutch public organizations by addressing the client feedback and related sentiments. Vanessa et al. [69] present a study for analyzing the behavioral similarities and differences of 3-1-1 phone service (formal) and Twitter (informal) channels for reporting issues in the community. They present a supervised learning method to

automatically classify the complaints and label them to show the comparison of types of reports posted on two channels.

2.2.2 Mining Public Complaints and Communication on Twitter for Building Prediction Models for Situation Awareness

Research shows that Twitter is not only being used to identify public complaints and address them, but also for mining citizens' complaints and building predictive models for spreading awareness. Kumar et al. [76] present an application of Twitter to atomically determine road hazards by building language models based on Twitter users' online communication. Their proposed system aims at identifying potential road safety hazards that pose driving risks. Gu et al. [77] propose a methodology to mine tweet texts to extract incident information on both highways and arterial roads and conduct a case-study of two regions: the Pittsburgh and Philadelphia Metropolitan Areas. Fu et al. [78] describe an approach to extract and analyze real-time traffic related Twitter data for incident management purpose. Their experimental results reveal that mining social media data is a promising approach for early incident detection and can be used as a supplemental source for incident data collection. Eleonora [79] present a real-time monitoring system for traffic event detection from Twitter stream analysis and conduct a case-study for the Italian road network. They demonstrate detection of traffic events almost in real time, often before online traffic news websites. Mai et al. [80] demonstrate the use of data from public social interactions on Twitter as a potential complement to traffic incident data. Schulz et al. [81] present a solution for a real-time identification of small-scale incidents using microblogs, thereby allowing to increase the situational awareness by harvesting additional information about incidents. Napong et al. [82] present a study on social-based traffic information extraction and classification consisting of mining Twitter for traffic congestion, incidents, and weather. Panagiotopoulos et al. [83] present a study on examining the use of Twitter for making public awareness during emergencies such as heavy snow and public riots.

2.2.3 Steps taken by the Indian Government

In order to address the complaints and grievances of citizens, the Indian government also initiated several policies and organizations. The aim of these organizations is to mine online complaints specific to the concerned department and address them in a timely and efficient manner.

1. **TwitterSeva:** Ministry of External Affairs, India launched a program to address public complaints on one shared portal. The portal was introduced to unite 200 social media handle, and any tweet mentioning these accounts is directly forwarded to TwitterSeva[1] platform. The proposed policy was introduced to address the complaints of local Indian embassy and regional passport office. Further, in extension to the TwitterSeva, several other services #DOTSeva, #BSNLSeva, #MTNLSeva, and #PostalSeva were introduced to address the complaints of telecommunication and stakeholder departments. Another extension of TwitetrSeva, #MociSeva was introduced for the consumers seeking assistance from the Ministry of Commerce & Industry.
2. **Cybercell:** In order to address the online complaints against women harassment, an online cybercell was introduced by the Union minister for women and child development (@Manekagandhibjp), Govt. of India[3]. The objective of this cell is to address the complaints and grievances of online trolls and right-wing abuse against women in India.

[3] http://www.huffingtonpost.in/2016/07/08/maneka-gandhi-sets-up-cyber-cell-to-check-online-abuses-on-women/

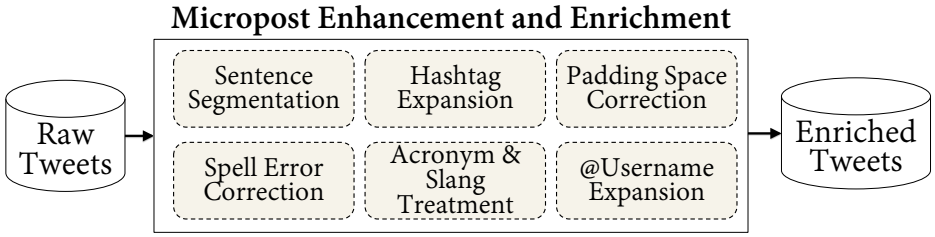


Figure 2.2: High Level Framework Design of Micropost Enrichment Algorithm- The proposed framework addresses the challenge of noisy data in the tweets by expanding joint hashtags, normalizing text, expanding @username mentions, correcting spelling errors and correcting slang and abbreviations in a raw tweet.

2.3 Research Contributions

Previous studies [69] [72] [73] [74] [75] focus on conducting empirical analysis on online and offline data and examine the use of Twitter for client feedback and sentiments about the services provided by police departments. Similarly, some studies [76] [77] [78] [79] [81] [82] perform Twitter stream analysis and traffic news websites analysis for building early prediction based model for determining road safety hazards and real-time traffic congestion. In contrast to the existing studies, the specific and key contributions of the study presented in this chapter are the followings:

1. We address the challenge of free-form text in tweets and capture the dependencies between noisy text and semantics of the complaints. We propose a multistep micropost enrichment algorithm that identifies the natural language text-based errors in the tweets and corrects them to increase the efficacy of features identification.

2. We propose a text analysis based ensemble classifier for identifying the complaints and grievances reports posted by public citizens on official Twitter handles of public agencies and the Indian government. The proposed model uses various contextual and linguistic features that indicate the relation between a complaint and the concerned department. We also identify features that are strong indicators of differentiating a complaint report from non-complaint tweets.

3. We further explore the similar idea and investigate the efficacy of spatial (geographical location metadata) and linguistic features to discover insights from less informative complaints regarding issues on bad road conditions. For example, potholes, dysfunctional or lack of streetlights and under-construction roads causing life risks to public citizens (referred as the killer roads in this chapter). We build a content-disambiguation and text-analysis based model to enrich our linguistic features for identifying exact problem reported in the tweets.

4. We publish the first ever databases of citizens' complaints reported on the official Twitter handle of public agencies accounts of Indian Government. Agarwal et al. [64] consists of enhanced dataset of citizen-centric complaints on Twitter collected for extracting information on public citizens' complaints and grievances. Agarwal et al. [62] consist of syntactically enriched tweets reported to the ministry of road, transport, and highways to complaint about risky, hazardous and poor road conditions. We make our datasets [64, 62] publicly available for the research community so that our results can be used for benchmarking, comparison and further extension.

Table 2.1: Concrete Examples of Original and Enriched Complaint Tweets Before and After Performing the Text Pre-processing and Enhancement of the Content

	Before	After
SSN	@AamAadmiParty Umm nice.What about Vadra, Khurshid , Sheila Dixit **...** LoL. Where is the 370 pages report **???**@nitin_gadkari	@AamAadmiParty nice. What about Vadra, Khurshid, Sheila Dixit. LoL. Where is the 370 pages report? @nitin_gadkari @ArvindKejriwal
	@nitin_gadkari @PMOIndia Jewar Toll Plaza is charging toll tax & not accepting 500/1000 Rs Notes also**...**why so**???**	@nitin_gadkari @PMOIndia Jewar Toll Plaza is charging toll tax & not accepting 500/1000 Rs Notes also. why so?
HTE	@IncomeTaxIndia unearths Rs 52.5cr **#blackmoney** from Amritsar Rice miller via @timesofindia	@IncomeTaxIndia unearths Rs 52.5cr black money from Amritsar Rice miller via @timesofindia.
	@ArvindKejriwal**#OddEvenDobara** being broken in broad daylight 2 men inside @abpnewstv @dtpTraffic	@ArvindKejriwal Odd Even Dobara being broken in broad daylight 2 men inside @abpnewstv @dtpTraffic.
SEC	material **snt** by railways n 21/3. Current status:railways saying we haven't **gt** material.	Material sent by railways n 21/3. Current status:railways saying we haven't got material.
	my **frnd geting** call **frm** 8757969668 claiming to b from Naptol asking to dposit 12500 so that they will deliever Safari car.	My friend getting call from 8757969668 claiming to b from naaptol asking to deposit 12500 so that they will deliver safari car.
AST	Sir **y** is it so **dat** the vendor in S-9 of shramjeevi exp. Is taking more charge on a bottle of amul kool?	Sir why is it so that the vendor in S-9 of shramjeevi express. Is taking more charge on a bottle of amul kool?
	Pls tel n c JI **dat** failure of laws on cops led to **dis** in US learn from it.	Please **tel** and see JI that failure of laws on cops led to this in US learn from it.
UME	**@MORTHIndia @nitin_gadkari @PMOIndia @narendramodi** pls take appropriate action. I ve raised so many complains. No response whatsoever	MORTHINDIA Nitin Gadkari PMO India Narendra Modi pls take appropriate action. I ve raised so many complains. No response whatsoever
	@mansukhmandviya @MORTHIndia @CMOGuj @vijayrupanibjp whn will u cleared VASAD BAGODARA HIGHWAY PROJECT ITS 2 MUCH NEEDED DAILY ACCIDNT HAPEN	Mansukh Mandaviya MORTH CMO Gujarat Vijay Rupani whn will u cleared VASAD BAGODARA HIGHWAY PROJECT ITS 2 MUCH NEEDEDDAILY ACCIDNT HAPEN

2.4 Micropost Enrichment Algorithm

As discussed in Section 2.1, the automatic identification of complaint reports is a technically challenging problem due to the free-form nature of social media text. The user-generated data on social media websites contains various natural language issues such as incorrect grammar, spelling mistakes, term obfuscation and usage of abbreviation and short-forms. We propose a micropost-enrichment algorithm consisting of several techniques for data pre-processing and text enrichment for overcoming the problem of noisy data in the text. The proposed algorithm is a multistep iterative process that takes a raw tweet as an input and provides a syntactic and semantically enriched tweet. Figure 2.2 shows the high-level block diagram of the proposed algorithm primarily consisting of 6 phases: Sentence Segmentation (SSN), Hashtag Expansion (HTE), Padding Space Correction (PSC), Spelling Error Correction (SEC), Acronyms & Slang Treatment (AST), and @Username Mentioned Expansion (UME). Algorithm 1 shows the step by step pseudo code of the proposed algorithm. Table 2.1 shows concrete examples of the tweets before and after executing the each phase of proposed micropost-enrichment algorithm individually. We discuss each of these phases in the following subsections:

Algorithm 1: Micropost Enrichment Algorithm

Data: Tweet $t \in T$, Set of hashtags H present in t, Set of @usernames U mentioned in t, Bing Search API Key K, Lexicon of domain specific slang DL, Standard slangs SS
Result: Enriched tweet $t' \in T'$
Function `SentenceSegmentation(`t`)`
1 t' = t.replace(consecutive dots, dot)
2 t' = t'.replace(consecutive ?, ?)
3 t' = t'.replace(consecutive !, !)
4 append white space before and after #
Function `HashtagExpansion(`t'`, `H`)`
5 **for** ***all*** $h \in H$ **do**
6 if h contains underscore then split h from underscore
7 if h contains numeric value then split h from numeric value
8 **if** *h contains uppercase letter* **then**
9 **if** *consecutive uppercase letters* **then**
split from max index of uppercase and split from index-1 location
else
split from uppercase letter
10 use porter stemming algorithm
Function `SpellingErrorCorrection(`t'`, `K`)`
11 create set wg of consecutive 3-word grams of t'
12 **for** ***all*** $wg \in WG$ **do**
13 wg'=FetchText(Include_Results_For(BingSearch(K, wg)))
14 t'=t'.replace(wg, wg')
Function `SlangConversion(`t'`, `SL`, `DL`)`
15 if DL present in t' then replace DL with the mapped keyword
16 if SS present in t' then replace SS with the mapped keyword
Function `UsernameExpansion(`t'`, `U`)`
17 **for** ***all*** $u \in U$ **do**
18 expanded_user=Profile_Name(TwitterAPI(u))
19 replace u from t' as expanded_user
20 T'=`SentenceSegmentation(`T`)`
21 T'=`HashtagExpansion(`T'`, `H`)`
22 T'=`SpellingErrorCorrection(`T'`, `K`)`
23 T'=`SlangConversion(`T'`, `SL`, `DL`)`
24 T'=`UsernameExpansion(`T'`, `U`)`

2.4.1 Sentence Segmentation

In the first phase of the micropost-enrichment model, we enrich the syntactic structure of a tweet. Steps 1 to 4 in Algorithm 1 shows the steps for sentence segmentation in a tweet. We remove all the URLs and filler terms (such as umm, hmm, errr) from the tweets. We also replace special characters appearing consecutively with one character. For example, "?????" and "!!!!" are replaced with "?" and "!" respectively. Due to the 140 character limit of tweets, users avoid using spaces after special characters and hashtags (#) that lead to the poor hashtag identification. For example, Twitter does not identify the hashtags written consecutively or without spaces before "#". For example, in a tweet consisting of the following pattern of hashtags "#DwarakaExpressWay#DelhiRoads#Potholes", Twitter does not recognize any of the hashtags. Figure 2.3 shows concrete examples of tweets without consisting of whitespaces before hashtags

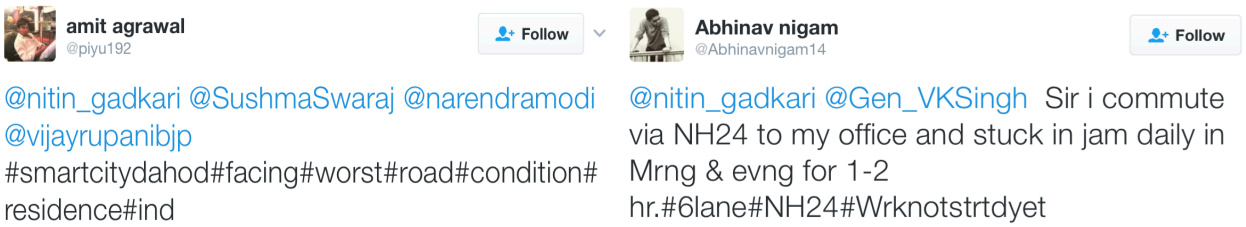


Figure 2.3: Snapshot of Concrete Examples of Tweets Consisting of No Whitespace Before Hashtags Making them Unrecognizable as Distinct Hashtags

and therefore making them as one long word. Whereas in a tweet consisting of a whitespace before "#" such as "#DwarakaExpressWay" "#DelhiRoads" "#Potholes", three distinct hashtags are extracted using Twitter API. Therefore, in order to enrich the hashtags list, we perform a cleaning on the retrieved tweets and add a padding whitespace before every # and later trim all extra whitespaces occurring consecutively.

2.4.2 Hashtag Expansion

Hashtags in a tweet are the key descriptive phrases written by simply adding a hash symbol # before the phrase. These hashtags are used to tag the tweets and make them easily searchable for other users. We observe that with the latest trend on Twitter, users create more descriptive hashtags by combining two or more strings (character sequences). For example, #GoodWorkRailwayPolice, #borivalitrainchaos, #WeRequestModiGovt and #ScamQueenOnRoad. Since hashtags are the user generated phrases, there is no one standard approach or a defined structure to create a joint hashtag. Therefore, expanding such hashtags is a technically challenging problem. We propose a four-step approach to split the strings in a joint hashtag and semantically enhance the tweets in our experimental dataset. Steps 5 to 10 of Algorithm 1 shows the steps used for hashtag expansion in a tweet.

1. **Common Separator:** We expand the hashtag by simply splitting it around common separators ('_' and '-') used on Twitter. For example, #strong_action and #no-strong-action-by-police are converted into 'strong action' and 'no strong action by police' respectively.

2. **Uppercase Letters:** Unlike common hashtag separators, due to the presence of acronyms and abbreviations, splitting a hashtag from uppercase letters can increase the noise in the expansion. Therefore, we split a hashtag by keeping consecutive uppercase together until the last upper case in a string (if an acronym is followed by a lowercase letter). For example, #MarchForDemocracy is converted into 'March For Democracy', #CharchaOnRWH is converted into 'Charcha On RWH' and #FANTomorrow is converted into 'FAN Tomorrow'. We, however, do not expand a hashtag that contains only uppercase letters. For example, #CCTV and #FDDI.

3. **Alphanumeric String:** We split an alphanumeric phrase in a set of all numeric and character strings. If the expanded strings contain any uppercase letter, we expand it further by using Step 2. For example, #TheriJoins100crClub is converted into 'Theri Joins 100 cr Club'.

4. **Porter Stemming:** We use Porter stemming algorithm [84] to identify the longest substring (a small character sequence within a large sequence) in a hashtag and split the string at that location. For example, #seriousissue and #havesomesenseofchecking are converted into 'serious issue' and 'have some sense of checking' respectively. If a hashtag is created by joining only Hindi language words, then we do not perform any expansion on the hashtag. For example, #swacchbharat and #swaranshatabdi.

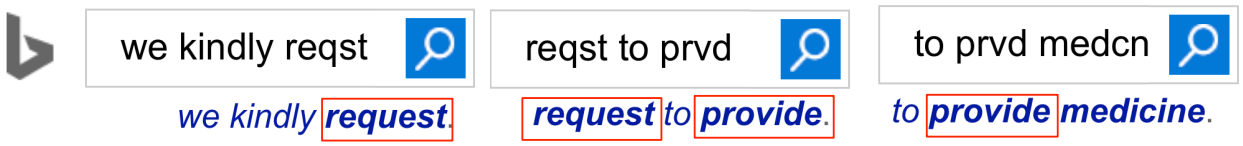


Figure 2.4: Snapshot of a Worked-out Example (splitting a sentence into group of 3-grams) of Spell Correction Framework using Bing Search Engine

2.4.3 Padding Space Correction

Similar to the addition of space padding before hashtag (#), we add a whitespace before all special characters (comma, period, question mark, colon, semicolon, underscore and exclamatory marks). We further trim the consecutive extra whitespaces, fixing the presence and absence of spaces before and after the special characters. We, however, do not add padding whitespaces before the special characters appearing in @username mentions. For example, @nitin_gadkari, @poonam_mahajan, and @PIB_India.

2.4.4 Spelling Error Correction

Due to the presence of free-form and user generated text, a tweet is high likely to have spelling and grammar mistakes. In this phase, we address the challenge of spell errors in a tweet by correcting them using n-gram model. Steps 11 to 14 of Algorithm 1 shows the steps used for hashtag expansion in a tweet. We split our tweet from each special character (_, -, ?, !) and create a set of n-consecutive word grams (3 in our approach). For example, for a tweet consisting of words 'a b c d' has a set of two 3-grams 'a b c' and 'b c d'. We query each n-gram on a Search Engine using GET method while the language of spell checking is set as English. We use Bing Search Engine API [85] for spell correction since for our experimental data, Bing Search Engine [86] predicts the spellings more accurately in comparison to Google, Yahoo, and DuckDuckGo search Engine even though Bing Search allows to make less than 7 queries per seconds which are lesser than other search engines. We extract the translated n-gram resulted as "Including results for" in Bing Search and store the results for each n-gram separately and compute their extent of similarity with queried n-grams. For the first and last words in a tweet, we replace them with the terms corrected by Bing Search. Whereas, for the terms appearing in multiple n-grams, we replace them by the term corrected in the majority of n-grams. For example, for a given sentence, 'pleas answr my query asap', we create a set of three 3-grams [n1: 'pleas answr my', n2: 'answr my query', n3: 'my query asap']. Here, 'answr' is replaced by 'answer' only if both n1 and n2 are corrected as 'please answer my' and 'answer my query'. Figure 2.4 shows an another worked-out example of proposed spell correction framework using Bing Search Engine.

2.4.5 Acronyms and Slang Treatment

In this phase of proposed micropost enrichment algorithm, we expand the Internet Slangs and normalized text of a tweet written in 'sms' language in three stages: domain specific slang, standard slang, and user slang. We conduct a manual inspection on Twitter and identify several acronyms and slang commonly used in complaints and their respective definition. For example, rly (railway), NH (National Highway), Toll, RTO (Regional Transport Office), RC (Vehicle Registration Certificate), KM (KilloMeters), STN (station), Acc (accident) and RD (Road). If a term does not exist in domain specific slang keywords then we replace it with user slang. User slang are the slang that are commonly used on social media but are not present or recognized as standard slang used for those words. For example, PL is used for please whereas, it is defined as "parent looking" in a standard slang dictionary. We identify the 1, 2 and 3 character-grams present in

Table 2.2: Concrete Examples of Tweets Recorded Before and After Complete Execution of Micropost-Enrichment Algorithm

Before	After
Its 11:30 AM, no one at service window no 131@rtoahmedabad. plnty of ppl waiting frm past hour. @anandibenpatel @vijayrupanibjp @nitin_gadkari.	Its 11: 30 AM, no one at service window no 131@rtoahmedabad. Plenty of people waiting from past hour. Anandiben Patel Vijay Rupani Nitin Gadkari.
I'm seeing most of d ministers r so active in wrk n social media too like @narendramodi @manoharparrikar @nitin_gadkari 1/2.	I'm seeing most of the ministers are so active in work and social media too like Narendra Modi Manohar Parrikar Nitin Gadkari 1/2.
Sir d directional boards 2 d delhi airport r worst plz update dem immediately daily thousands of people r facing dis problem @nitin_gadkari.	Sir the directional boards 2 the delhi airport are worst please update them immediately daily thousands of people are facing this problem Nitin Gadkari.

the dataset and create an exhaustive list of such slang words. In the third stage of slang and acronyms treatment, we replace the remaining slang with their standard definitions by sending them through a POST request on NoSlang- largest portal for Internet slang dictionary & translation [87]. We, however, do not replace any numeric character unless it is an alphanumeric string. For example, semantically, 4 can mean either word "for" or "four" while "r8" is used for "right" and "b4" is used for "before".

2.4.6 Username Expansion

Expansion of direct mentions replaces the Twitter @screenname with the user profile name making them easily recognizable by the named entity recognizers. For example, @Dev_Fadnavis, @Ra_THORe and @dtptraffic are expanded to Devendra Fadnavis (Chief Minister of Maharashtra), Rajyavardhan Rathore (Minister Of State for Information & Broadcasting) and Delhi Traffic Police respectively. Steps 17 to 19 of Algorithm 1 shows the steps used for hashtag expansion in a tweet.

In order to minimize the noise in output tweet, we apply domain specific slang conversion after every other phase of data enrichment. We observe that applying spell correction phase directly after hashtag expansion can rather create the noise in the data. For example, a hashtag #CMODelhi is expanded as "CMO Delhi" corrected as "COM Delhi" while CMO is an abbreviation used for "Chief Minister of". Similarly, we apply domain specific slang conversion after @username expansion as due to the limited number of characters in the username, the @screenname contains several abbreviations. For example, the username of "@PMOIndia" is "PMO India" that is expanded as "Prime Minister of India" using the abbreviation and slang treatment. Table 2.2 shows examples of user complaints tweets recorded before and after executing all phases of micropost enrichment algorithm.

2.5 Experimental Setup

As discussed in Section 2.1, the different complaints reported to official public agencies have different purposes- 1) complaints that seek for immediate action and acknowledgment from the concerned authorities and 2) complaints that are reported for general awareness and posted for bringing the attention of the government to the reported issue. In this Section, we address these different types of complaints and present two case studies on mining Twitter micro-blogging website for extracting actionable information and

@dtptraffic No one in Uttam Nagar follows traffic rule, no traffic police personnel is available. No enforcement, no fear. @AlokVermaCPDP

@IncomeTaxIndia My mom suppose to receive her #PANCard by 20June. Not received & No response from #FirstFlight. @bookcomplaint #FFCLisCRAP

@DelhiPolice 10000₹ wrongly deducted frm my ac in Lakshmi Nagar.can u help me in refunding the amount.

@RailMinIndia @sureshpprabhu rain water is coming inside on seat through window sealing. Bedsheet n blanket both got wet. PNR NO.8248914739

Figure 2.5: Concrete Examples of Citizen Complaints Reported to Government's Official Twitter Handlers- Tweets address various public issues such as traffic violation, inconvenience in train coach, wrongly detection of money from bank account.

insights. We conduct two case studies on Twitter for identifying complaint reports posted on an official Twitter handle of public agencies. It is required to perform two case-studies because both requires focused attention, variation in the pre-processing and enrichment pipeline. In both case studies, the identification of some complaints is specific to the structure of tweets while some complaints are general. Since we cannot have one-size-fits-all, we need to analyze these two types of complaints separately as two different case-studies.

2.5.1 Mining Twitter to identify complaints and grievances posted to public service agencies (Case-Study 1)

Public complaints such as scam, fraud, corruption, illegal transactions and money deduction from banks, no-refund on ticket cancellation, delay of train and buses, poor facilities and amenities in trains and on stations, traffic violations, bribes, crimes happening on road and streets. General issues that are faced by public citizens and seek for immediate actions by the concerned authorities. It is seen that people post complaints and report incidents causing discomfort and poor experience of government facilities to the citizens online. Identification of such complaints and grievances on Twitter is important for the public agencies. According to Railway Statistics, in the first week of January 2017, 55 trains were delayed while 6 trains were canceled due to heavy fog n North India[4] causing discomfort to passengers. While, heavy rainfall in Delhi-NCR (National Capital Region- includes New Delhi and urban areas surrounding it in neighboring states of Haryana, Uttar Pradesh, and Rajasthan) caused a massive traffic jam on roads for hours. These complaints were reported on Twitter and requesting for an immediate response from the concerned authorities. Based on the types and frequency of complaints reported on Twitter, we identify several public agencies account on Twitter that receives complaints on a daily basis and seek for immediate action. Figure 2.5 shows examples of such complaints posted on various public services accounts of the Indian government on Twitter. However, due to the number of complaints reported on these accounts the overall response rate for these complaints is very low. We examine several government accounts on Twitter and find that official public service agencies like Ministry of Information and Broadcasting, Ministry of Railway, State Police, Income Tax India, Telecommunications, Finance, Ministry of Urban Development, Agriculture & Farmers Welfare and Consumer Affairs receiving frequent complaints from public citizens. While only some of the official handles like Railway Ministry of India, Income Tax India, Delhi Traffic Police, and Delhi State Police are active accounts on the website. Further, there are several accounts of officials and organizations belonging to the same department on Twitter. For example, @CPDelhi is an official account of Commissioner of Police, Delhi while @DelhiPolice is an official twitter handle of Delhi state Police. Based

[4] http://www.ndtv.com/india-news/55-trains-delayed-six-cancelled-due-to-fog-in-north-india-1644432

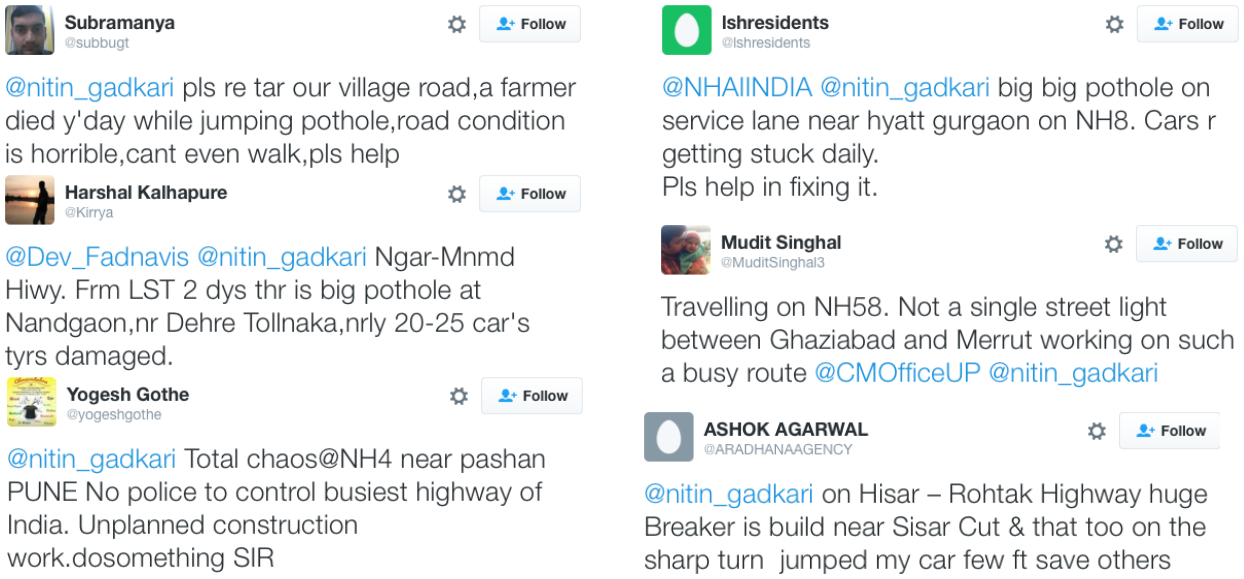


Figure 2.6: Concrete Examples of Complaints on Killer Roads and Citizens' Discomfort Reported to Official Twitter handle of Ministry of Road, Transport and Highways, Government of India- Addressing Various Road and Transport Related Issues such as Pothole, Dysfunctional Streetlights, Unplanned Construction on Highway and Breakers near Intersection and Sharp Turns.

on the activity feeds on such similarly grouped accounts, we identify four Indian government related Twitter accounts for the purpose of our data collection and experimentation: @RailMinIndia (Railway Ministry of India), @dtpTraffic (Delhi Traffic Police), @DelhiPolice (Delhi Police) and @IncomeTaxIndia (Income Tax Department, Government of India). We conduct experiments on these four official government accounts to test the generalizability of our approach. We select only these 4 accounts as they are active and spans diverse topics (not focusing on similar kind of complaints).

2.5.2 Mining Twitter to extract information on bad roads complaints (Case-Study 2)

Bad road conditions such as road irregularities, roughness, potholes, bumps, patchy surface and poorly designed speed breakers make the road risky and hazardous for drivers resulting in accidents[5]. Similarly, several crashes happen due to blind or improper curves, temporary diversions, traffic on under repair and under-construction roads. Dysfunctional street lights or dim lights, encroachments on roads by shops, hawkers and broken or hard to see road signboards are also major causes of accidents[6]. Bad road conditions based (the focus of this case study) causes are different than accidents happening due to indiscipline or careless driving (driver's fault such as speeding or drunken driving) or poor weather conditions (rain, storm, and

[5]http://www.thehindu.com/news/cities/bangalore/traffic-police-have-a-list-of-121-spots-that-are-accidentprone-in-bengaluru/article8701749.ece

[6]http://indianexpress.com/article/cities/pune/pune-roads-battles-with-accidents-and-crime-as-street-lights-dysfunction/

snow). As per the statistics of Ministry of Road Transport and Highways (MORTH) [88], Government of India, on an average 400 road deaths take place every day in India[7]. According to MORTH statistics, in 2015, India had 501, 423 road accidents causing 146, 133 deaths while 10, 727 people were killed in the crashes happened because of potholes. In the first quarter of 2016, over 300 deaths were reported in Uttarakhand state where accidents happened due to the sharp turns and absence of crash barriers. According to the reported accidents and newspaper statistics, Delhi recorded the highest number of road accident deaths (17 deaths per hour) in India in 2015[8]. While, in Maharashtra, Meghalaya, Madhya Pradesh and Uttar Pradesh state, the maximum number of deaths happened due to the potholes and under-repaired roads and highways[9]. While our case-study is on Indian tweets and we have provided several statistics on India, bad road conditions and tweets on road irregularities is seen in other countries also[10,11]. As discussed in Section 2.1, due to the increasing trend of adoption of social media by Indian Government and public agencies to reach out to the people, public citizens use Twitter to post their complaints and report the incidents to the concerned authorities [89]. The complaints on killer roads contain the information about road irregularities and other issues causing high risks and discomfort to the citizens. Figure 2.6 shows concrete examples of bad road conditions complaints reported to the Indian Government's official Twitter handle. In order to collect our dataset of complaint reports, we identify two official Twitter handle of concerned departments of Indian Government. @nitin_gadkari[12] is the official Twitter account of Mr. Nitin Gadkari- current Union Minister of Road Transport & Highways and Shipping in India. @MORTHIndia[13] is the official Twitter account of Ministry of Road Transport and Highways, Government of India. We observe that public citizens also report their complaints to @narendramodi and @PMOIndia which are the official Twitter accounts of current Prime Minister of India, Mr. Narendra Modi. However, in order to remove the noise from our data and create the dataset under the scope and focus of our study (killer road complaints), we select only @nitin_gadkari and @MORTHIndia accounts for our experimental dataset collection.

2.5.3 Experimental Dataset Collection

We conduct our experiments on an open source dataset collected in real-time from Twitter. We use the official Twitter REST API [26] for downloading the tweets posted to the selected accounts in real time. We search the tweets mentioning the screen name (along with @ symbol) of these accounts. Twitter REST API allows us to extract only past 7 days of data. Therefore, for each account selected in case study 1, we collect data (Test Data 1) for 4 weeks (from 11 April 2016 to 8 May 2016). Whereas, for @nitin_gadkari and @MORTHIndia, we collect data (Test Data 2) for 8 weeks (18 July 2016 to 13 September 2016). Figure 2.7a displays the statistics of the experimental dataset (Test Data 1) consisting of tweets (original tweets obtained, filtered and sampled) posted over a four-weeks duration (11 April 2016 to 8 May 2016). Figure 2.7a reveals the total number of tweets collected for the four accounts were: @RailMinIndia- 36182, @dtpTraffic- 1524, @DelhiPolice- 1720 and @IncomeTaxIndia- 383. Table 2.3 shows the statistics of Test Data 2 collected using Twitter API. Table 2.3 shows that we were able to collect a total of 81, 304 tweets consisting of 17, 511 original tweets, 11, 092 replied tweets and 52, 701 re-tweets. We conduct our study on English language tweets

[7] http://www.ndtv.com/india-news/every-day-400-people-die-in-road-accidents-in-india-shows-government-data-1403899

[8] http://auto.ndtv.com/news/road-traffic-deaths-highest-in-delhi-1417473

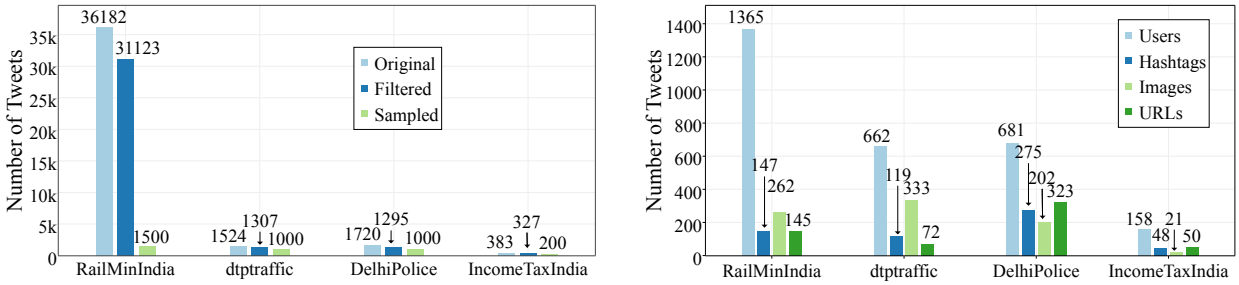
[9] http://timesofindia.indiatimes.com/india/400-road-deaths-per-day-in-India-up-5-to-1-46-lakh-in-2015/articleshow/51919213.cms

[10] https://www.washingtonpost.com/news/wonk/wp/2015/06/25/why-driving-on-americas-roads-can-be-more-expensive-than-you-think/

[11] http://metro.co.uk/2017/01/26/is-this-the-most-potholed-roundabout-in-the-country-6406995/

[12] https://twitter.com/nitin_gadkari

[13] https://twitter.com/MORTHIndia



(a) Number of Tweets in Experimental Dataset

(b) Contextual Metadata Statistics

Figure 2.7: Experimental Dataset Statistics- Illustrating the statistics of number of tweets collected, filtered and sampled for each account. Further, showing the variation in number of sampled tweets consisting of various entities (URL, hashtag, image, @usermention)

Table 2.3: Statistics of Experimental Dataset- Illustrates the number of Original and Unique tweets Collected After Data Extraction, Filtering and Sampling. Showing the Number of Sampled Tweets Consisting of Contextual Metadata (Image, Video, URL, Hashtag, @username mention). TS denotes the Timestamp (dd/mm) of the Tweet.

Collection	**Total**	**Original**	**Re-tweet**	**Replied**	**English**	**Unique**	**Sampled**	**Users**
	81,304	17,511	52,701	11,092	13,368	13,208	3,302	2,604
Tweets	**Image**	**Video**	**URL**	**Hashtag**	**@mention**	**Lat-Long**	**Min TS**	**Max TS**
	598	0	854	821	1673	8	20/07	Max 13/09

only. Therefore, we use "detected language" [26] feature of Twitter API and filter the tweets identified as non-English or undefined ('und'). Further, the aim of our study is to identify linguistic features of complaint reports and build a text classifier. Therefore, we filter the posts which have no text and contains the only image, @username mentions, video, and URLs. Figure 2.7a and Table 2.3 shows the number of tweets remaining in after applying the filter operation on Test Data 1 and Test Data 2 respectively. To remove bias from the data, we perform a random sampling and select a subset of the dataset from the filtered tweets. The number of tweets after filtering and random sampling for the four accounts of Test Data 1 (@RailMinIndia, @dtpTraffic, @DelhiPolice and @IncomeTaxIndia) are 1500, 1000, 1000 and 200 respectively. While, as the Table 2.3 reveals that after filtering and random sampling (25% of unique tweets) of records, there are $3,302$ unique and English-language tweets posted by $2,604$ unique users.

2.5.4 Experimental Dataset Charcaterization

2.5.4.1 Multimedia Content

In addition to downloading the textual content of the tweet, we use the API to extract contextual metadata such as the type of tweet (re-tweet, reply or original), hashtag, URL, image, video and user mentions in the tweet. We also extract the general details of the blogger such as the @username and geo-location (if available). Figure 2.7b displays the number of sampled tweets containing distinct users, distinct hashtags, image and URL present in Test Data 1. Figure 2.7b reveals that the number of tweets with media attachment

Table 2.4: Frequently Occurring Hashtags and Topics- Illustrating the examples of hashtags and related topics that are most discussed in the experimental dataset. The count shows the frequency of each tag in dataset.

Hashtag	count	Topic	Hashtag	count	Topic
@DelhiPolice			@dtpTraffic		
#OddEven	14	Vehicle Rule	#OddEven	53	Vehicle Rule
#IPSAKnowledgeSeries	13	IPS Discussion	#OddEvenDobara	44	Vehicle Rule
#sexualharassment	6	Harassment	#kotlamubarakpur	3	Car on Fire
@IncomeTaxIndia			@RailMinIndia		
#SovereignUnnathi	2	Construction Project	#RailDrishti	15	Initiative
#Theri	2	Raid	#Latur	10	Relief Operation
#Aadhaar	1	Unique ID	#RailwayZoneForVizag	8	Metro

vary for each account. For example, 50% of sampled tweets in @dtpTraffic dataset contains external images. Whereas, in @IncomeTaxIndia and @RailMinIndia datasets it varies from 15% to 20%. We observe a variety of topics being discussed in the tweets indicated by several distinct hashtags.

Similarly, Table 2.3 shows the number of tweets (after random sampling) consisting of various distinct contextual metadata present in Test Data 2. Table 2.3 reveals that only 18% (598 out of 3,302) of the tweets in our experimental dataset collected for killer road complaints contain external images and 25% (821 out of 3,302) of the tweets contain distinct hashtags while only 8 out of 3,302 tweets contain geo-location information of the users posting these tweets.

2.5.4.2 @Username Mentions

In addition to the multi-media content, we extract all the direct mentions in our datasets (Test Data 1 and Test Data 2) and compute their frequency. For Test Data 1, we observe that several tweets contain direct mentions to related official government Twitter handles. For example, out of 1000 tweets in our dataset for @DelhiPolice, we observe 131 direct mentions to @CPDelhi which is the official Twitter handle of the Commissioner of Police of Delhi. There are 97 mentions to @ArvindKejriwal (Chief Minister of Delhi), 44 mentions to @HMOIndia (Home Minister of India), 34 mentions to @PMOIndia (Office of the Prime Minister of India), and 26 mentions to @narendramodi (Prime Minister of India). Similarly, we observe 720 direct mentions to @sureshprabhu (Minister of Railways, Government of India) in 1500 tweets belong to the @RailMinIndia dataset. The top direct mention (total count of 26) in the dataset on @IncomeTaxIndia is @arunjaitley who is the Finance Minister and Minister of Corporate Affairs in Government of India. The Twitter handle of Chief Minister of Delhi (@ArvindKejriwal) is the top direct mention (total count 139) in our dataset for @dtpTraffic.

Similarly, based on the statistics of Test Data 2, we find that other than the official Twitter handle of @nitin_gadkari and @MORTHIndia, citizens mention several official accounts of Indian government- specific to region and problems mentioned in their complaints. For example, out of 3,302 sampled tweets, 487 tweets have direct mention to @narendramodi which is an official Twitter handle of Mr. Narendra Modi- current Prime Minister of India. While, 365 tweets have mentioned @PMOIndia- the official Twitter handle of office of the Prime Minister of India. There are 229 mentions to @Dev_Fadnavis (Mr. Devendra Fadnavis, Chief Minister of Maharashtra state) and 139 mentions to @mlkhattar (Mr. Manohar Lal, Chief Minister of Haryana state). Similarly, we observe 69 direct mentions to @sureshpprabhu (Mr. Suresh Prabhu- Minister for Railways, Government of India). While, other than @nitin_gadkari and @MORTHIndia, the official Twitter handle of Prime Minister of India (@narendramodi) has the maximum number of direct mentions in our dataset.

Table 2.5: Concrete Examples of Frequently Occurring 7 and 8 Character-gram Strings in the Experimental Dataset of Each Public Service Account

Account	Hashtags
@DelhiPolice	traffic, missing, abusing, arrested, detained, criminal, communal
@dtpTraffic	flyover, oddeven, parking, pillion, crossing, redlight, hospital
@IncomeTaxIndia	website, efiling, pending, invoice, property, interest, marriage, passport
@RailMinIndia	toilets, sleeper, medical, delayed, cleaning, security, drinking, stoppage

2.5.4.3 Topic Modeling

Due to the diverse range of topics being discussed and reported in Test Data 1, we perform Topic Modeling on our exprimental dataset. Table 2.4 displays frequently hashtags, their count, and topic in Test Data 1. Table 2.4 reveals that the controversial odd-even traffic rule imposed by the Delhi government aimed at controlling the air pollution levels is one of the highly discussed topics on @dtpTraffic and @DelhiPolice. We also observe tweets on sexual harassment and a car on fire issue having direction mentions of @dtpTraffic and @DelhiPolice. Tweets on a government initiative (called as Rail Drishti) and bringing metro to a populated city were some of the topics on the @RailMinIndia dataset. Topics on a unique identity card and an income tax raid were topics on @IncomeTaxIndia. We also compute frequently occurring character n-grams in our experimental dataset. Table 2.5 shows some of the frequently occurring 7-gram and 8-gram strings in the dataset of all 4 accounts. The frequently occurring character n-grams indicate the type of issues being discussed in the respective twitter accounts.

2.6 Proposed Solution Approach

2.6.1 Filtering Non-Complaint Tweets

As discussed in Section 2.1, all public service accounts have open Twitter handles, and anyone can mention them in their tweets. Therefore, not all tweets that are posted to these agencies are complaint reports. In order to classify complaints and grievances (C&G) tweets, we build a one-class classifier since the class of other tweets is unknown. Conventional binary and multiclass classification algorithms classify the input data into pre-defined labels. However, the challenge arises when the input does not belong to any class or is irrelevant to the classification task. In one-class classification algorithms, the training data contains the sample of the relevant class (referred to as the positive class or target class) [90]. While for the other class training data does not include any instances. The one class-classification algorithm is trained on the samples of only positive class instances and it classifies as many instances as possible from target class and minimizes outliers. Therefore, it requires identifying strong and discriminatory features to find the best separation of both target and unknown class instances.

In the complaints and grievances reports identification problem, we, however, observe that various features are strong indicators of tweet not to be a complaint report. Despite having only one target class, we build another independent one-class classifier that identifies the tweets that are certainly not complaints and grievance reports (non-C&G).

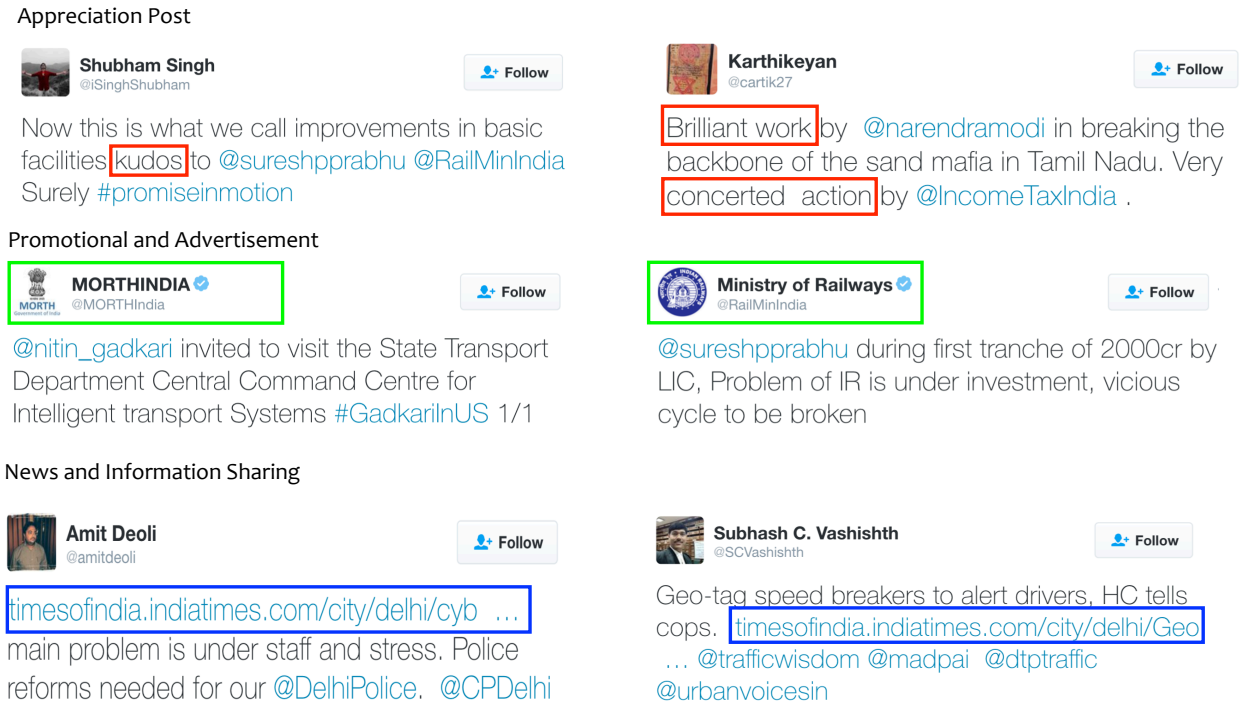


Figure 2.8: Concrete Examples of Non-Complaint Tweets Posted on the Official Twitter handle of Public Service Agencies- classified into 3 categories: appreciation, information sharing and promotional tweets.

2.6.1.1 Appreciation, Information Sharing and Promotion (AISP) Tweets Classifier

Based on our observation and analysis of non-complaint tweets, we divide them into four broad categories: appreciation, queries, information sharing and promotion tweets. Figure 2.8 shows examples of tweets classified into each of these categories. Since a query post can be ambiguous with a complaint tweet, we first classify other three categories of tweets and identify query tweets in later steps. We perform AISP tweet classification on the enhanced version of our experimental dataset (refer to Section 2.4).

1. **News Update or Information Sharing:** Tweets with the presence of an external URL (not image or video attached) are marked as non-complaint tweets. We identify such tweets by using the URL count feature of Twitter REST API. However, we do not classify tweets that has link to another tweet (referred as a quote).

2. **Promotion:** We observe that there are several tweets which are posted by other official accounts of same public agencies. For example, @sureshpprabhu (Minister of Railway Department of India) posting about a new policy and mentioning @RailMinIndia in his tweets. We classify such tweets by checking verified account value of the blogger using Twitter REST API. If the tweets are posted by the verified accounts, then we mark them as non-complaint posts.

3. **Appreciation Post:** We create an exhaustive lexicon L_K of appreciation keywords and convert them into their lemma form by using Stanford's CoreNLP API [91]. We create a bag of words L_W by converting each word of a tweet to their lemma form. If there exists an intersection between L_K and L_W, we compute the *joy* (emotion tone) feature of that tweet. If the confidence score of joy

Table 2.6: Grouped Triplets of Frequent N-grams Occurring in Public Agency Specific Complaint Reports.

Account	N-Grams	Grouped Triplets
@DelhiPolice	bribe, abuse, harassment, FIR, phone, action, report, complaint	<{bribe, abuse, harass}, {FIR, phone}, {action, report, complaint}>
@dtpTraffic	bribe, challan, violation, abuse, harassment, jam, commotion, congestion, accident	<{bribe, challan}, {violation, abuse, harassment}, {jam, commotion, congestion, accident}>
@RailMinIndia	train number, train name, coach, pnr number, bribe, corruption, report, complaint, action	<{train number, train name, coach, pnr number}, {bribe, corruption}, {report, complaint, action}>
@IncomeTaxIndia	pan number, ack, FIR, TIN, complaint, report, investigation, refund	<{pan number, ack, FIR, TIN}, {complaint, report, investigation}, {refund}>

feature of a tweet is above a certain threshold, then we classify it as an appreciation post. In order to compute the threshold value, we take a sample of 50 posts annotated as appreciation post and compute their joy confidence score using IBM Watson's Alchemy Tone Analyzer API [92]. The Tone Analyzer API uses natural language processing techniques and linguistic analysis to detect three types of tones from input text: emotions, social tendencies, and writing style. We analyze the content of a tweet and compute the emotions of the user. Alchemy API analyzes the text and generates scores on two levels (document level and sentence level) for five categories of emotions: anger, sadness, joy, fear and anxiety. We, however, compute the score for joy feature of the tones. Since the text length of tweets is very short, we select only sentence level measures of these tones. We compute the average of these confidence scores and record it as the threshold for testing tweets.

2.6.2 Features Extraction and Selection

2.6.2.1 Case Study 1

As discussed in Section 2.5.4, the complaints reported in Test Data 1 contain a diverse range of topics and do not contain a specific format or structure of report. Further, due to the free-form nature of social media and user-generated text these complaint reports use different terms for complaining similar incidents. Based on our observation on complaint reports posted on official Twitter handle of @railminindia [93], @dtptraffic [94], @delhipolice [95] and @incometaxindia [96], we identify five features that can be used to discriminate complaint reports. We discuss each of these features in the following subsections:

2.6.2.1.1 Frequent N-Grams

As discussed in Section 2.5.4, we find several character n-grams that occur very frequently in complaint posts and defines the topic of the complaint. We create a triplet of such n-grams and defines a new feature vector to avoid keyword-based-flagging method. We create these triplets based on the frequent n-grams that are common for each account (to keep our approach generalized) and n-grams that are unique to an account and not seen in other complaints. Table 2.6 shows the triplets created for each account. We identify n-grams that are similar to each other to merge them into one item of a triplet. For example, in a triplet <I1, I2, I3>, if we find a term I4 that is similar to I1 them we merge them together and expand our triplet as <I1/I4, I2, I3 >. Table 2.6 shows the grouped triplets for each account. We further avoid keyword-spotting approach and improve the efficiency of our feature vector. We use WordNet lexical resource [97] to identify the synonyms of the terms present in the tweets and compare them with the words present in our lexicon. For example, if the triplet in our feature vector has the word "investigation" and an unknown tweet contains

Table 2.7: Sample of Closed Domain (department specific) Keywords used in Citizens' Complaints and Grievances Specific to a Public Agency

Account	Key-Terms
@DelhiPolice	bribe, abuse, harassment, FIR, phone, action, report, complaint
@dtpTraffic	bribe, challan, violation, abuse, harassment, jam, commotion, congestion, accident
@RailMinIndia	train number, train name, coach, pnr number, bribe, corruption, report, complaint, action
@IncomeTaxIndia	pan number, ack, FIR, TIN, complaint, report, investigation, refund

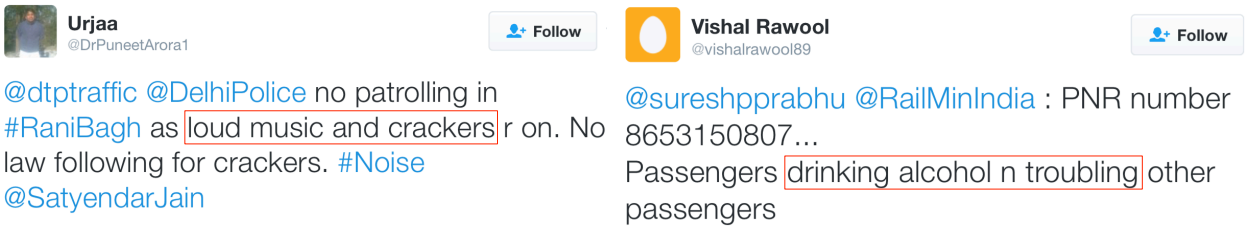


Figure 2.9: Examples of Complaint Reports posted by Citizens Facing Discomfort and Inconvenience due to the Reported Issue.

a term "enquiry", keyword spotting method will assign a value of 0. Whereas, "enquiry" is another term used for "investigation" and hence will be detected using a lexical database approach.

2.6.2.1.2 Closed Domain Key-Terms

Online government and public service agencies have several different departments and receives certain types of complaints that are specific to these departments. Based on our observation and manual inspection, we create a lexicon of the keywords that are specific to these complaints without going into low-level details of the type of complaints and to keep our features generalized. Table 2.7 shows a sample of such keywords for each account.

2.6.2.1.3 Events and Substances

We find that not all complaints posted to these government accounts are related to the problems faced by the reporters. But these reports are also of common concern and bringing the attention of public service agencies to various issues. For example, violation of traffic rules and alcohol consumption. In order to find such activities in the tweets, we create two more feature vectors "events" and "substances". We use IBM Watson Concept and Relationship Extraction API [98] and extract various events and substances being performed and reported in the tweet. The Concept and Relationship Extraction API [98] by IBM Watson performs linguistic analysis on input text and identifies the abstract concepts that are not necessarily to be the named entities. The Concept and Relationship Extraction API uses various lexical resources and knowledgegraphs

(dbpedia[14], yago[15], wordnet[16], and freebase[17]) as the back-end databases and enrich the concepts based on the semantic information (keywords, entities, sentiment, and verb normalization) of the terms used in the text. Figure 2.9 shows the concrete examples of tweets reporting about the specific events in their complaints.

2.6.2.1.4 Location

It is seen that while reporting complaints about public issues, people often mention their locations in their tweets. For example, an accident happened at location L_1 or a train did not arrive at time at platform P_1. Therefore, we use location parameter as another feature vector for classifying complaint tweets. We use IBM Watson Relationship Extraction (IRE) API to identify the named entities in a given tweet. However, we find that due to the presence of free form text and people names in locations, IRE is not able to identify the locations with 100% accuracy and predicts many places as person and organization. For example, M. B. Road, Gandhi Nagar, and AIIMS metro station. Therefore, we merge people, organization and location entities and further use Google geocoding API [99] to identify the terms that are locations. We apply IRE before performing geocoding as it groups the words that are likely to be one entity. For example, Preet Vihar, Nehru Place, and Hauz Khas Village.

2.6.2.1.5 Media Presence

Twitter allows users to attach multimedia files to their tweets such as video and pictures. Attaching a video and picture in the complaint tweets increases the credibility of their report as it provides evidence and helps concerned authorities to understand the severity of the problem. We identify the presence of media files (video or picture) in a tweet and create a Boolean feature vector in our model.

2.6.2.2 Case Study 2

It is seen that citizens post complaints about daily life issues specific to road and transport while only some of these complaints get addressed and resolved. It happens due to the large volume and high velocity of data posted to these accounts. We observe that all reported complaints do not contain necessary and sufficient information about the issue faced by the people and hence are left unaddressed or unresolved despite considered by the authorities. Another fact that we come across is that unlike other department complaints, killer road related complaints can have fewer chances to get addressed based on the amount of information available in the tweets. For example, in a railway department, the following two complaints "no blankets are available in S3 coach of Hemkund Express." and "Boarded Hemkund Express from Haridwar Junction. No blankets are available on the train. My PNR number is 12345" have similar possibility to get addressed. Whereas, in the complaints related to killer roads, a tweet "under-construction roads near Nehru Place metro station are causing accidents." is high likely to get addressed in comparison to the tweet "Nehru place roads are accident prone. Delhi roads are worst". We create a hypothesis that to successfully address a complaint about killer roads it is important to identify the three features:

1. To know the type of issue faced by the citizens so that the complaint can be forwarded to the concerned department

[14] http://wiki.dbpedia.org/about

[15] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/#c10444

[16] http://wordnet.princeton.edu/

[17] https://developers.google.com/freebase/

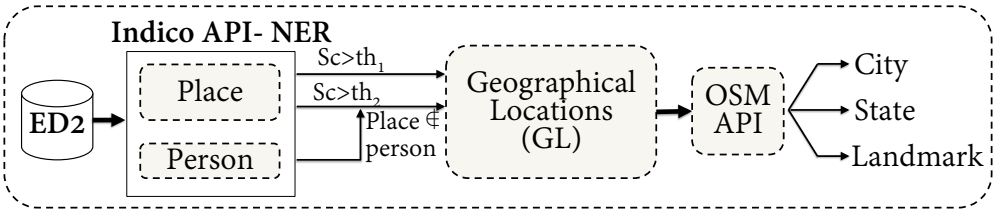


Figure 2.10: Proposed Framework for Identifying the Geographical Location (region, landmark and exact location) Reported in the Complaints)

2. It is required to know the region (or city) from which the complaint is reported
3. The pinpoint location (or landmark) of the problem [100]

We discuss each of these features and the proposed method for their identification in the following subsections:

2.6.2.2.1 Geographical Location Identification

Twitter microposts are the user-generated data and do not have a defined structure for writing. Therefore, despite correcting the spell errors and performing sentence segmentation on tweets, we find that existing named entity recognizers are not able to identify locations in tweets with 100% accuracy. Further, we find that in Indian locations scenario, the majority of places are identified as persons' names. For example, for the locations "Mahatma Gandhi Marg", "Vasant Vihar Colony", "PVR Saket"- "Gandhi", "Vasant" and "Saket" are identified as persons' names. Therefore, we use a combination of Named Entity Recognizer and GeoCoding API to extract the geographical entities mentioned in tweets. Figure 2.10 demonstrates the high-level framework for identifying a geographical location in tweets. We use Indico Text Analysis API [101] to extract the specific places and person referred in the tweets. Test Analysis model of Indico API allows a machine to discover a variety of knowledge from plain text input using Machine Learning and Deep Learning methods. Given a tweet $t_i \in ED2$ as plain text input, Indico API returns a list of JSON object of three key-value pairs. Each object represents a substring of tweet t_i- identified as a person or place, prediction confidence score and position of substring (starting and ending character) in the input tweet. The confidence score varies between 0 to 1 representing the confidence of text analysis model for identifying it to be a place or person.

Using Indico API, for a given tweet input t_i, we extract all the locations $L = \{L_1, L_2, ..., L_i \mid 0 \leq i \leq length(t_i)\}$ and persons $P = \{P_1, P_2, ..., P_j \mid 0 \leq j \leq length(t_i)\}$. Based on the score of verified locations, we divide the confidence score CS of locations into three categories- low ($CS < 0.2$), medium ($0.2 \leq CS < 0.5$) and high ($0.5 < CS$). For each location $L_i \in L$ with a confidence score $CS_{L_i} > 0.5$ and $0.2 \leq CS_{L_i} < 0.5$, we filter them as L_{High} and L_{Medium} respectively. While we discard the entities identified as a location with a confidence score below 0.2. Table 2.8 shows examples of tweets and the confidence score of entities to be identified as location. As discussed above, due to the ambiguity in location names, we observe that many places are identified as persons and vice versa. Therefore, we discard the places having a medium confidence score but also identified as person names. The extracted geographical location entities for a tweet t_i are $GL = L_{High} \cup (L_{Medium} - P)$. Further, if $GL_m \in GL$ is a substring of $GL_n \in GL$ then we discard GL_m and filter only GL_n entities identified as locations. For example, in the third example shown in Table 2.8, if "Vasant Vihar" is identified as a location then so are the "Vasant" and "Vihar". Therefore, we take the longest subsequence of the strings identified as locations and select "Vasant Vihar" as the location.

In order to identify the type of a location (region or landmark), we use OpenStreetMap (OSM) API [102]. OSM API takes geographical locations GL as an input and identifies their geocode (latitude and longitude). Further, it also identifies the type of place (such as road, building, school, highway etc) based

Table 2.8: Concrete Examples of Location and Person Names Identified using Indico Text Analysis Model- Illustrates the entities and their confidence score to be a place or person's name

Tweet	Place	Person
Accident at **Wadkhal**. Pls confirm if traveling towards **Alibaug**. Truck broke into 2 due to pothole.	<Wadkhal:0.73, Alibaug: 0.71>	-
Pathetic highway since ages between **Roorkee** to **haridwar**. Corrupt ministry. I bet if you drive on own	<haridwar:0.95, Roorkee: 0.60, Pathetic highway: 0.03>	<Roorkee: 0.06>
sir can u please resolve RTR flyover mess in **Vasant Vihar**, **delhi**. Daily thousands of liters of fuel is wasted in traffic there	<delhi:0.92, Vasant Vihar:0.53, RTR: 0.05, Vasant: 0.03, Vihar: 0.02>	<sir:0.37, Vihar: 0.03, Vasant Vihar:0.02>

Table 2.9: List of All Map Features and Amenities Identified by OpenStreetMap API for the Tweets Present in Our Experimental Dataset and Labeled as Landmark

Private, administrative, locality, canal, residential, hospital, hamlet, aerodrome, station, trunk, commercial, construction, neighborhood, common, bank, motorway_junction, suburb, kindergarten, river, bus_stop, bus_station, school, restaurant, unclassified, motorway, hotel, secondary, peak, primary, water, public_building, bicycle_parking, service, house, stadium, railway

on the using the tags associated with their basic form. Each tag represents the geographic feature of the place. We group these features into two categories: region and landmark. We define a region as the name of a village, town, city, district and tertiary. Whereas, we define a place to be a landmark if it tells the pinpoint location, nearby area or locality. For example, geographic features such as school, church, hospital or building are labeled as landmarks. Table 2.9 shows the name of all amenities labeled as landmarks.

2.6.2.2.2 Problem or Issue Identification:

As discussed above the public agencies accounts on Twitter are open and therefore, any one tag them in their complaints. However, during our inspection on Twitter, we observe that despite being a complaint report, not all tweets posted on @MORTHIndia [103] and @nitin_gadkari [104] are related to killer roads. Further, to identify the complaint tweet, it is important to detect the information about topic or cause of the issue. For example, in a tweet "*Big potholes on Delhi roads. It's risky to drive at nights*", identification of term "pothole" is important to know that the tweet belongs to a killer road category. Tweets are user generated data that contains free-form text and high likely to have different terminology for reporting similar complaints. For example, "streetlights on highway not working", "dim streetlights on highway", "highway is full of dark", "there is a complete blackout on highway" are reporting the same complaint using different terminology.

In our proposed method, we address the challenge of keyword based flagging methods for identifying the issues and problems reported in tweets. Figure 2.11 shows the high-level research framework for topic

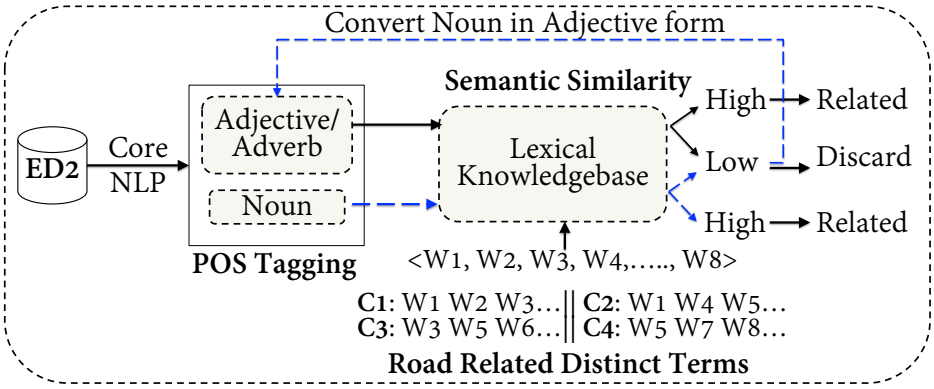


Figure 2.11: High-level Block Diagram of the Proposed Framework for Mining Tweets for Identifying Issues Related to Bad Road Conditions.

Table 2.10: Snapshot of the Key-terms Related to Various Issues Reported in Killer Road Complaints- used to identify the topic of problem by computing the conceptual similarity with terms occurring in unknown samples.

Category	Related Terms
Dysfunctional facilities	street light, traffic light, drainage
Risky and Hazardous	pothole, street light, traffic light, hawkers, breaker, sign board, construction, intersection, diversion, animal, turn, crash, barrier
Poor Conditions	road, highway, traffic, flyover, underpass, bypass, chowk, expressway
Indiscipline and Carelessness	barrier, repair, sign board, pothole, parking, police

identification. We apply Stanford CORE NLP parser [91] on the tweets identified as unknown after executing AISP classification (ED2). Stanford CORE NLP parser performs part-of-speech (POS) Tagging on each tweet and identify the nouns, adjective, and adverbs in our posts. We filter the terms already tagged as places or persons during geographical location identification. In order to address the limitations of keyword based flagging methods, we compute the semantic similarity between the terms presented in a tweet and the terms related to road, transport, and highway (RTH) complaints. Research shows that a lexical database can be used to compute semantic similarity between two terms [60] [105]. We use Conceptnet commonsense knowledgebase [106] as a lexical resource and identify the conceptual similarity between the phrases presented in the tweets and RTH complaints. ConceptNet is an ontology-based semantic network in which each node represents a concept, and each edge represents the relation between two concepts. ConceptNet is an open source knowledge base consisting of concepts extracted from common and informal sentences and daily basic understanding [107]. It not only identifies the shortest path between two concepts but also identifies the common-sense associations that users make among the concepts [108] [109]. Therefore, due to the free-form structure of tweets, we select ConceptNet over WordNet [97] [110] and further, ConceptNet is an extension over WordNet and DBPedia [111] lexical resources.

We conduct a manual inspection on the Twitter handle of @MORTHIndia and @nitin_gadkari and identify several issues reported by public citizens about dangerous road conditions. Based on our observation, we define four categories of killer road complaints: 1) Dysfunctional facilities, 2) Risky and Hazardous (accident prone), 3) Poor Conditions and 4) Indiscipline and Carelessness. For each category, we define certain key

Table 2.11: Examples of ConceptNet Distance between Terms Present in Complaint Tweets and Issues Related to Killer Roads.

	highway	traffic	flyover	animal	construction
congestion	0.1	0.34	0.03	0	0.01
street_light	0.85	0.50	0.28	0.12	0.02
intersection	0.73	0.4	0.31	0.03	0.04
hawker	0.26	0.53	0.35	0.12	0

terms and compute their conceptual semantic similarity (using ConceptNet) with the terms present in the tweet. Table 2.10 shows the list of all key terms defined in the categories mentioned above. Table 2.10 also reveals that there are several entities such as street light, signboard, and pothole that occur in more than one category. Therefore, we create a lexicon of all distinct words and compare them with the terms present in the tweet and identified as a Noun. The high confidence score of similarity computation shows that the tweet is high likely to be in the associated category. We use association method of ConceptNet to access the concepts and compute their semantic similarity. We use ConceptNet5.0 API [112] and find the association between two given concepts using [assoc/c/en/concept1?concept2/] request while we filter results computed for only English-language concepts using [filter=/c/en/].

Given a tweet $t = \{t_1, t_2, t_2, ..., t_n\}$, a set of killer road complaints related key-terms $K = \{C_1 \cup C_2 \cup .... \cup C_j\}$ where $1 \leq j \leq 4$ and a network of knowledgebase $N_{concept}$. We find a $t_i \in t$ where $POS(t_i) \in \{NN, NNP, NNS, NNPS\}$, $t_i \notin GL$ and $t_i \notin P$ (refer to Section 2.6.2.2.1). Then, $\forall K_p \in K \mid K = \{K_1, K_2, ...K_m\}$, we compute the confidence score with each t_i as $CS_{t_iK_p}$ =Conceptual_Similarity$(t_i, K_p, N_{concept})$. Therefore, for each $t_i \in t$ and $K_p \in K$, we get a two-dimensional vector $Score_{m,n}$ matrix.

$CS_{tK}=$

	t_1	t_2	.	.	t_i	.	.	t_n
K_1	0.4	0.5	.	.	0.23	.	.	0.50
K_2	0.3	0.02	.	.	0.7	.	.	0.4
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
K_p	0.24	0.25	.	.	0.3	.	.	0.7
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
K_m	0.7	0.8	.	.	0.6	.	.	0.9

If, $\exists K_p \in K : CS_{t_iK_p} \geq th$ then the topic of a complaint is said to be true for tweet t. Whereas $\forall K_p \in K : CS_{t_iK_p} \leq th$, we convert the adjective form of noun words if available. If $\forall t_i \in t$ and $\forall K_p \in K$, the confidence score between concepts $CS_{t_iK_p} \leq th$, the topic of a complaint is said to be false for t. Table 2.11 shows concrete examples of conceptual similarity score (confidence) score computed between various concepts- noun terms present in the tweets and topic-related keywords. For each t_i, we select all K_p such that $CS_{t_iK_p} \geq th$ and assign various categories (Dysfunctional Facilities, Risk and Hazardous, Poor Condition and Indiscipline & Carelessness) to the tweet. $\forall j$, C_j=TRUE for tweet t, if $K_p \in C_j \mid 1 \leq j \leq 4$. We, however, find several such examples, where the problem identification is challenging not only for the machine but also for human annotation. Such reports contain humor, ambiguity, and sarcasm and increase the number of false alarms in problem identification. For example, "Nitin Gadkari Come to Gurugram! Enjoy the fun of Venice for free! Great offer by Haryana Government!", "Nitin Gadkari NH17 is National Horror neglected by Govt. After Govt", and "But this highway is like a black spot on moon".

2.6.3 Classification

2.6.3.1 Case Study 1

In this phase of our proposed framework, we build an ensemble learning based Support Vector Machine (SVM) classifier. We split our data (tweets classified as unknown in AISP classifier) into 1:3 ratio and use 25% and 75% of the tweets as training and testing dataset. We train our model from feature vectors created in Phase 2 and perform one-class classification on each tweet in testing dataset. SVM classifier learns the features from the tweets verified and annotated as complaint tweets and identify whether a given tweet (in the testing dataset) is a complaint tweet or not. If a tweet is not identified as C&G, then it is classified as unknown. Research shows that the performance of an SVM classifier can be improved by modifying the kernels of the classifier [113]. Therefore, to investigate the efficacy of our proposed approach and identified features (linguistic and contextual), we run our classifiers for SVM by varying the kernel functions of SVM. We test our methodology for linear, polynomial and RBF (Radial Basis Function) Kernels. Further, we ensemble all three kernels and execute them in cascaded and parallel manner. For each SVM classifier (linear, polynomial, RBF kernels and ensemble classifier), we get a set of tweets classified as complaints and grievances tweets and another set of unknown tweets.

2.6.3.1.1 Query Identification

Another indicator of a non-C&G tweet is a query post where users post their general questions about the things that are related to these public services. In order to perform a verification of unknown tweets classified by C&G classifier, we identify the query posts (refer to Section 2.6.1.1) in unknown tweets. However, due to the presence of free-form text, it is hard to determine a query post just by spotting a 'Wh' word in the starting of a sentence or a question mark at the end of the phrase. We tag part-of-speeches in each tweet and defines 5 patterns that are strong indicators of a tweet to be a query. "{}" denotes the optional part-of-speech.

1. Modal/VBP + PRP + VB + NN(P/S) or PRP
2. Modal/VBP + NN (P/S) + VB
3. WHADVP/WHNP/WP + Modal/VBZ + DT + VB + NNP/PRP
4. If + NNP/PRP + VBZ/Modal + NNP/PRP
5. WHADVP/WHNP/WP + VBZ/Modal + RB + NNP/PRP + JJ

We further avoid the pattern 'WP + DT + NN (P/S)' as it is commonly used pattern in complaint tweets.

2.6.3.1.2 Complaint Type Identification

To identify the type of complaint, we perform topic modeling on all tweets identified as complaint and grievances reports. Since, the type complaints posted on Twitter vary for each account and have a large dimension, for example, in @RailMinIndia, cleanliness, delay of a train, refund-issue, waiting room, platform, berth allocation, poor service and assistance in coach and many more related complaints. Similarly, in @dtptraffic, various complaints on traffic rules violation, yellow line violation, bribery, illegal challans, riding motorbikes without helmets and similar complaints with different issues can be there. Therefore, we use natural language processing technique and remove the dependency with keyword spotting methods. We use Alchemy concepts and taxonomy API [114] and label these complaints into the most likely topic and sub-topic defined in the taxonomy hierarchy. For example, riding without helmet or driving without a number plate both are traffic violation related complaints. Similarly, unhygienic food serving or low-quality facilities to passengers are labeled as poor assistance in coaches.

2.6.3.2 Case Study 2

As discussed above, users can direct mention a public agency account in their tweets while reporting a complaint. However, due to no restriction to the direct mention, users tag these accounts in many non-complaint and AISP tweets. Based on our inspection of available information in the tweets posted on @MORTHIndia and @nitin_gadkari, we define three dimensions of tweets: Useful tweets, Nearly-Useful tweets, and Irrelevant tweets. We discuss each of these tweets in the following subsections:

1. **Irrelevant Tweets (IRT):** We define a tweet as irrelevant if the post is not about the poor conditions and irregularities of roads or highways causing life risks, discomfort, hazard or poor experience to the citizens. Table 2.12 shows examples of such tweets posted on official accounts of Government authorities of Road and Transport Department. Table 2.12 reveals that the tweets labeled as irrelevant are either off-topic or the authors discuss the problems related to road and transport; however the complaints are not about the poor conditions of roads or faulty and dysfunctional facilities.

2. **Useful Tweets (UT):** Useful tweets U_t are the posts which are a clear indicator of complaints and can be used to identify the low-level details of the issue faced by the citizens. Given a tweet t_i and a set of named entities $X =< T_c, T_l, T_p >$, we define t_i as a useful tweet- $t_i \in U_t$ if $t_i \in N$ (dataset) and $t_i = \{X, T_o \mid \forall x \in X : P(x)$ where $P(x) \neq \phi\}$. While, T_c denotes the name of city or region, T_l denotes exact geographical location or landmark, T_p denotes the problem or issue reported in the complaint and T_o denotes other words in the tweet. Table 2.12 shows examples of tweets consisting of detailed information about the city, landmark location, and the problem reported in the tweet. For each $t_i \in U_t$, we refer them as Experimental Dataset 3 i.e. $ED3$.

3. **Nearly-Useful Tweets (N-UT):** Nearly-Useful tweets are the tweets posted for complaining a report but containing incomplete or insufficient information about the issue. For example, missing the exact location of the problem, ambiguity in reporting the issue, defining the problem on an abstract level and lacking the details. Given a tweet t_i, we define it as a nearly-useful tweet $t_i \in NU_t$ if $t_i \in N$ and $t_i = \{X, T_o \mid \exists x \in X : P(x)$ where $P(x) = \phi\}$. Table 2.12 shows examples of tweets posted as complaint reports but providing insufficient information. For simplicity of tweets, we removed the @username mentions from the tweets and used the corrected and enriched form of tweets (after applying micropost-enrichment algorithm). As Table 2.12 reveals, in the first tweet of the nearly useful (N-UT) category, the author has mentioned the problem of potholes in the city of New Delhi while the exact location of potholes is missing from the complaint. Similarly, in the second example, the author has mentioned the landmark but did not specify the city or region name. Since multiple towns and areas can have the locations with the same name makes it difficult to understand the low-level details about the complaint. For each $t_i \in NU_t$, we refer them as Experimental Dataset 4 i.e. $ED4$.

2.6.3.2.1 Enrichment of Incomplete Reports

In this phase of proposed approach, we convert nearly useful tweets into useful tweets by extracting possible information about missing entities. As discussed above, a complete complaint report is comprised of three major components: city T_c (or landmark or region), pinpoint location T_l, and the problem T_p reported in the tweet T. However, based on the type of missing information, it is not always possible to convert an incomplete or nearly-useful tweet to a useful and complete report. For example, in the absence of issue expected to be reported in the complaint, it is hard to identify the problem component since the complaint is subjective and can be about any topic or it can be completely irrelevant as well. Therefore, we discard the tweets identified as nearly-useful tweets with no issue mentioned in the reports.

We apply the geographical location hierarchy model (bottom-up) and use graph backtracking method to identify the region or locality for a given pinpoint location in the tweet. Figure 2.12 shows the concrete

Table 2.12: Concrete Examples of Reports Identified as Irrelevant Tweets (IRT- not specific to a killer road complaint), Useful Tweets (UT- providing low-level details about the complaint) and Nearly-Useful Tweets (N-UT- reporting incomplete information) for Addressing and Resolving the Complaint

Type	Tweets
IRT	1. can we take some action on such wastage of public time n fuel as it's wastage of natural resources 2. Team works Sir Great decision in building the nation's future in this circumstance 3. Why state governments highways still charging tolls ?
UT	1. Sir your attention please. these are craters on Mum-Goa Highway near Chiplun. Being hopeful 2. please fix the big crater dug up by MNGL across the road, that connects the Mont Vert Tropez Society to Wakad road in Pune 411057 3. Huge pothole on NH160 in Igatpuri Talke village (80km frm Thane) Cars breaking down continuously
N-UT	1. people are struggling with pothole roads in national capital Delhi, do something about that 2. huge pothole on road at kapurbawdi, even during ganpati festival,where is authority 3. Street full of hawkrs no street light working whole city under chaos Ranchi has gone to worse please help

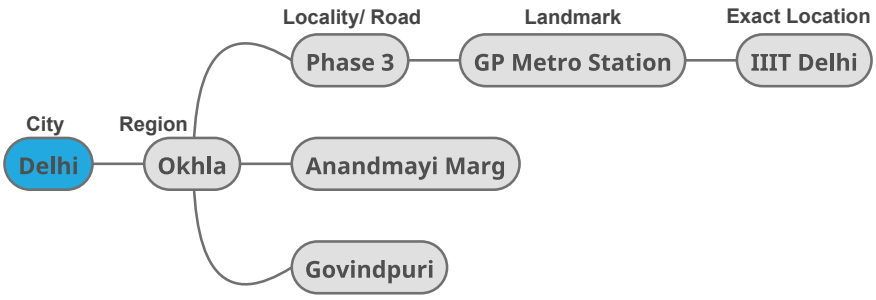


Figure 2.12: A Concrete Example of using Geographical Location Hierarchy Model to Enrich Location Component in the Tweets.

example of a location hierarchy. For a given pinpoint location T_l, we use OpenStreetMap API and extract the detailed information associated with a geographical location e.g. for IIIT-Delhi, we are able to extract GovindPuri Metro Station as landmark, Phase 3 as locality in Okhla region which is located in Delhi city. We use OpenStreetMap API [102] to enrich the region and city information T_c for each tweet consisting of only pinpoint locations T_l. Since it is possible for two different locations to have the same name, we further use Twitter user location to verify the identified city information. We extract the location of all distinct users available in our experimental datasets 4 (ED4- Nearly Useful Tweets). If the city or state information mentioned on their Twitter profile matches with the location identified using OpenStreetMap API then we show the 100% accuracy of the location component enrichment; otherwise, we proceed with the

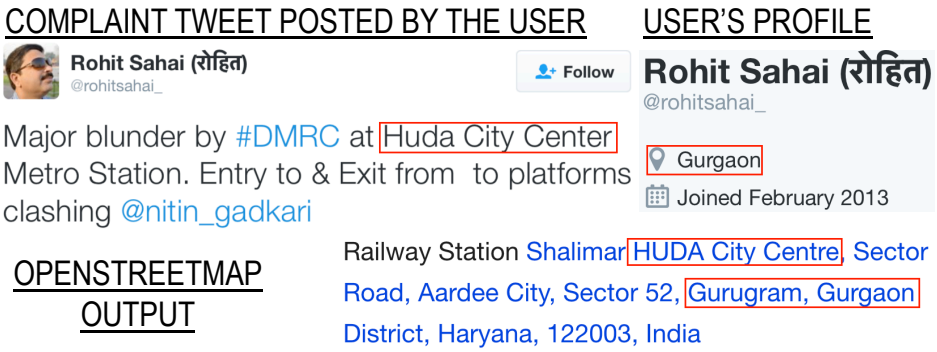


Figure 2.13: A Concrete Example of Location Identification Performed using OpenStreetMap and Verified using User's Profile Information.

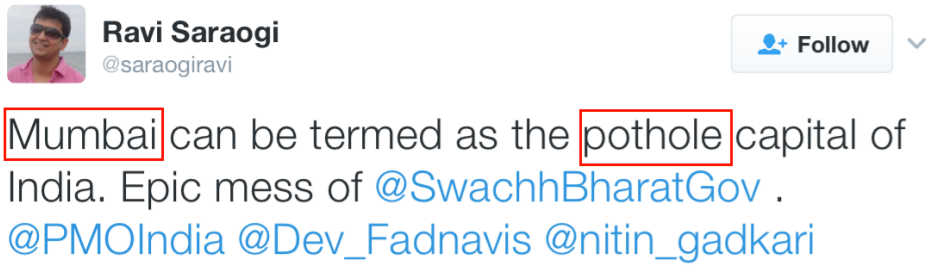


Figure 2.14: A concrete example of tweet consisting of only city name and problem while the exact location of reported issues is missing.

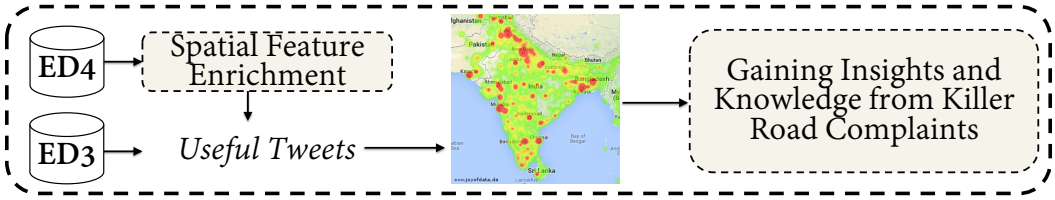


Figure 2.15: A Block Diagram Presentation of Extraction of Information and Insights from Public Complaints.

enriched results without claiming the 100% accuracy in enrichment. Figure 2.13 shows a concrete example of a location verification method. Figure 2.13 shows that "Huda City Center" is detected as a location entity. We furter use OpenStreetMap to extract the place or city name as Gurgaon which is also mentioned as the users' location in his Twitter profile.

As illustrated in Figure 2.12, for a given region or city, it is not possible to trace in the geographical location hierarchy model and identify the pinpoint location of the problem. For example, for the tweet shown in Figure 2.14, the reported problem aims at the generalized picture of Mumbai roads, and therefore it is infeasible to identify the exact location of the problem. While using OpenStreetMap, only state information can be extracted as Mumbai is a city and the capital of Maharashtra state which does not provide any information about the problem's location.

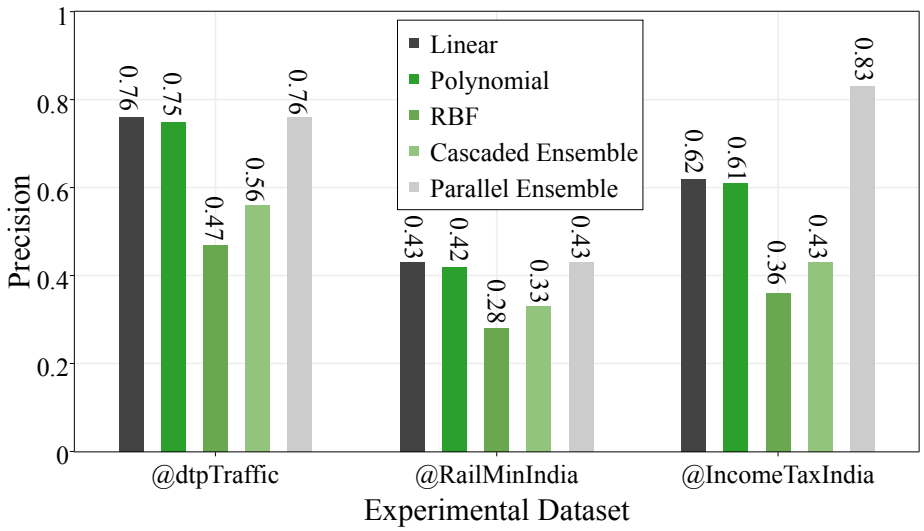


Figure 2.16: Confusion Matrix Results for C&G Tweets Classifiers (SVM with 3 different kernel functions and Ensemble Classifiers)- Column charts illustrate that linear kernel outperforms other kernels and by ensembling all kernels in cascaded or parallel boost the overall performance of each kernel.

2.7 Empirical Analysis and Evaluation Results

2.7.1 Case Study 1

In this Section, we present the accuracy results of each classifier and also discuss the influence of various kernel functions in SVM on the accuracy of proposed approach. We evaluate the accuracy of our classifier by comparing the observed results against actual labeled class. We conduct our experiments on the sampled tweets collected for @dtpTraffic (1000), @DelhiPolice (1000), @RailMinIndia (1500) and @IncomeTaxIndia (200) and report the accuracy of AISP classifier for each account. Proposed AISP classifier identifies 47 (A:12, IS:27, P:8), 132 (A:20, IS: 99, P: 13), 121 (A:41, IS:76, P:4) and 35 (A:4, IS:30, P:1) tweets as AISP for @dtpTraffic, @DelhiPolice, @RailMinIndia and @IncomeTaxIndia respectively. Based on our experimental results, we are able to correctly classify 124, 34, 32 and 103 tweets for @DelhiPolice, @dtpTraffic, @IncomeTaxIndia and @RailMinIndia respectively. While there is a misclassification of 8, 13, 3 and 18 tweets in similar order of accounts.

We execute our C&G classifier on unknown posts classified in previous phase (@dtpTraffic: 953, @DelhiPolice: 868, @RailMinIndia: 1379 and @IncomeTaxIndia: 165). We split these unknown tweets into training and testing dataset. Since we implement a one-class classification algorithm, we train our model only for target class (complaints and grievance reports) tweets while test dataset contains tweets that belong to complaint or others categories. For @dtpTraffic, we use 239 and 714 tweets as training and testing tweets respectively. Similarly, for @RailMinIndia, @DelhiPolice and @IncomeTaxIndia, we use 345, 217 and 42 tweets as training dataset respectively. While, 1034, 651 and 123 tweets are used as testing datasets for these accounts. Since the accuracy measures are biased towards the majority class in the dataset, we evaluate the performance of our classifiers using the standard information retrieval metric i.e. precision. We, however, find that due to a high imbalance of complaint reports in @DelhiPolice experimental dataset, our model does not find enough tweets for training the model (<50). Therefore, we discuss the complaint classification results for remaining three accounts.

Figure 2.16 shows that linear kernel in SVM outperforms RBF kernel with a reasonably high margin (varying from 20% to 30%). The reason that linear kernerlized SVM outperforms other kernels is due to the sparse nature of our feature vector model. Further, the vector model contains a high dimensional feature space consisting of frequent n-grams, closed-domain keywords, events & substances, location, and media presence. Since the features selected for classification are discriminatory and contains relevant information to the target class, despite having a high dimensional vector model, the linear kernelized SVM avoids overfitting of test objects and outperforms other kernel support vector machine classifications.

Figure 2.16 also shows that for @dtpTraffic (linear kernel), we are able to achieve the maximum precision rate i.e. 76% (184/(184+58)). Whereas, for @IncomeTaxIndia and @RailMinIndia, we were able to achieve a precision rate of 62% (31/(31+19)) and 43% (188/(188+252)) respectively. Our results reveal that using the linear kernel in one-class SVM classifier, we are able to achieve an overall accuracy up to 60%. Whereas, there is an overall misclassification of up to 10% in identifying complaint tweets as unknown. The column charts in Figure 2.16 shows that linear and polynomial kernels give similar results with a difference 1% to 2% in precision rate. Using SVM polynomial kernel, in @dtpTraffic experimental dataset, we were able to identify complaint tweets with a precision of 75% (170/(170+56)). While, for @RailMinIndia and @IncomeTaxIndia, we were able to identify complaints tweets with a precision rate of 42%(139/(139+189)) and 61% (22/(22+14)) respectively. To compute the efficacy of our approach for correct classification, we record an overall misclassification of 12% (complaint tweets wrongly classified as unknown) for all accounts for polynomial kernel SVM classifier.

2.7.1.1 Boosting of Baseline Approach

As discussed in the literature, the performance of SVM classifier can be boosted by modifying the kernels or combining more than one classifiers [113]. Therefore, to boost the efficiency of our proposed approach, we ensemble our SVM classifiers (3 different kernels) together and evaluate their performance while arranged in cascaded and parallel fashion. We compare the accuracy results of ensemble classifiers with three classifiers executed individually. Our results reveal that similar to individual classifiers, ensemble classifier also gives best results for @dtpTraffic (maximum precision). For a given testing dataset (dtptraffic: 714, RailMinIndia: 1034, IncomeTaxIndia: 123), parallel ensemble SVM classifier outperforms cascaded ensemble classifier. Using a combination of linear, polynomial and RBF kernels in a parallel manner, we were able to achieve a precision of 75%, 83% and 39% for @dtpTraffic, @IncomeTaxIndia and @RailMinIndia respectively. While there is an overall misclassification of 18% in identifying complaint tweets as unknown. Figure 2.16 reveals that by arranging these kernels in cascaded order, it decreases the performance of overall classification from 10% to 20%. For example, for @dtpTraffic and @IncomeTaxIndia datasets, we achieve a precision of 56% and 43% respectively which are approximate 20% lesser than the individual precision of linear kernel SVM classifier. In comparison to cascaded ensembling, in parallel ensemble classification, we are able to boost the accuracy for @IncomeTaxIndia dataset by 21% whereas, for @dtpTraffic, the performance is maintained with a precision of 76%.

In parallel ensemble learning model, the performance of each kernel SVM is independent to other kernels used for ensembling. Whereas, in cascaded or serial ensemble learning, the performance of one model is dependent on the performance of previous kernel models. Therefore, the ordering of various kernelized SVMs also impacts the overall performance or accuracy of ensemble classification. In our serial ensemble classification model, the linear SVM is followed by polynomial and RBF kernel SVMs. The lower performance of RBF kernel impacts the overall accuracy of classification despite the high performance of linear kernel model.

We label complaint tweets of each account using taxonomy and concept feature. Our results reveal that for @dtpTraffic account, maximum complaints are posted regarding traffic light violation, illegal tax, bribe payment and license related issues. Similarly, in @RailMinIndia experimental dataset, maximum complaints belong to theft, food assistance, cleanliness and train delayed issues. Unlike @RailMinIndia and @dtpTraffic,

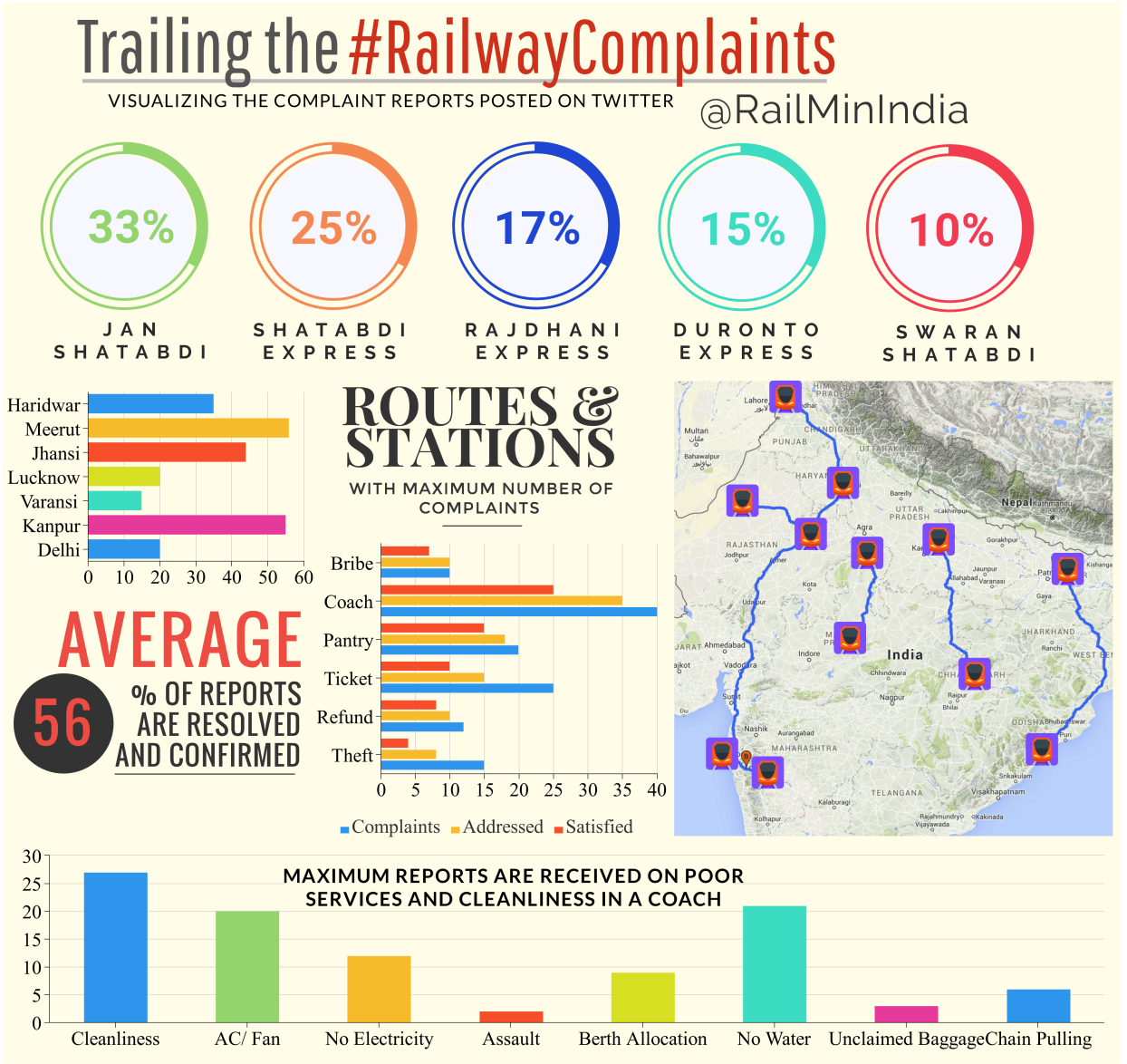


Figure 2.17: A Dashboard Representation of Complaints and Grievance Reports Received by @RailMinIndia. Illustrating the Percentage of Different Types of Complaints for Popular Trains, Routes and Stations in India.

@IncomeTaxIndia has a very small subset of tweets and therefore, we do not find a wide range of complaints. The maximum number of complaints in @IncomeTaxIndia belongs to PAN card related issues.

2.7.1.2 Infographic Visualization of Complaints and Grievances Reports

Dashboards are a medium of communication and a data visualization tool which provides a powerful means to present information Key Performance Indicators (KPI) and metrics important to the organization

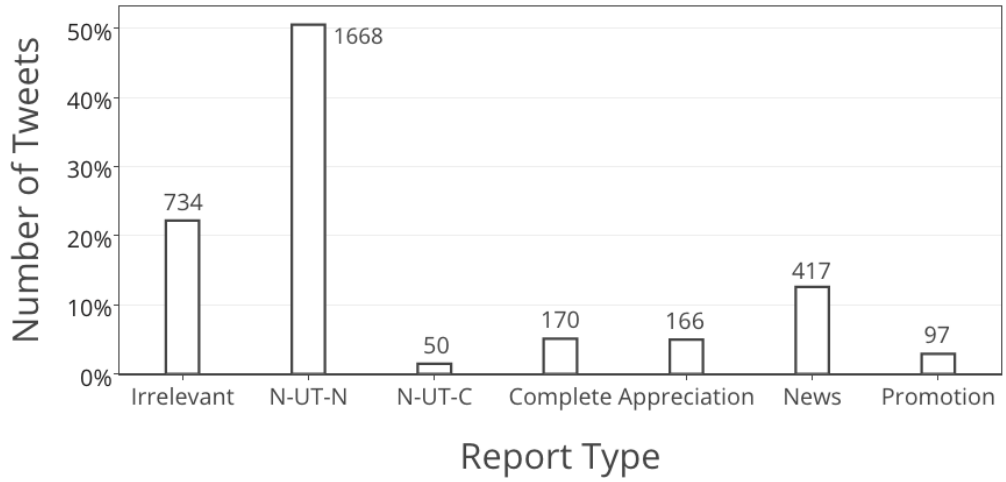


Figure 2.18: Experimental results and count of distinct tweets identified as appreciation, news, promotional, complete reports, irrelevant tweets, nearly-useful tweets (Convertible- N-UT-C and Non-Convertible - N-UT-N)

or the decision maker [115]. We design a dashboard based on real and synthetic data (to demonstrate the potential) from Twitter tailored for our application to demonstrate its ability for extracting information through visual inspection. Figure 2.17 displays a dashboard showing the status of complaints and grievances received by the @RailMinIndia Twitter account. As shown in Figure 2.17, the dashboard provides a quick glance on the percentage or amount of complaints and grievances received for some the important trains and routes. Our objective and study consisting of proposing and presenting business intelligence dashboards as front-end information visualization tools to gain insights from the data mined from the tweets. Figure 2.17 is a combination of both an analytical dashboard and operational dashboard[18]. For example, in Figure 2.17, the number of complaints received for various routes and trains are dynamically updated on the dashboard and comes under frequently changing and current performance metrics (operational dashboard). On the other hand, the histogram in Figure 2.17 displaying the distribution of the number of complaints across categories for a particular period of time such as a month or quarter falls under the analytical dashboard.

2.7.2 Case Study 2

In this Section, we discuss the empirical analysis performed on the tweets and acquired experimental results. As discussed in Section 2.4, we collect the original data of complaint reports from Twitter and call it as our experimental dataset 1 (ED1) after executing the proposed micropost-enrichment algorithm. We further filter the tweets that are certainly not complaint tweets and call our remaining tweets as experimental dataset 2 (ED2) (refer to Section 2.6.1). Based on the features extracted and selected in Section 2.6.2, we divide our tweets into irrelevant tweets (IRT), useful tweets (UT) and nearly-useful tweets (N-UT). We refer UT and N-UT as experimental datasets ED3 and ED4 respectively. As discussed in Section 2.6.3.2.1 and shown in Figure 2.15, we convert possible nearly-useful tweets into useful tweets and combine them with ED3. As illustrated in Figure 2.15, we extract actionable information and insights from these useful tweets using visualization method. We plot the extracted features and their statistics on the map and demonstrate the results using front-end visualization.

[18]`https://www.klipfolio.com/resources/articles/operational-analytical-bi-dashboards`

Table 2.13: Examples of Tweets Present in Our Experimental Dataset and Classified into Non-Convertible Nearly-Useful Tweets. Table also Illustrates the Available and Missing Component in the Tweets

Tweet	Available	Missing
Mumbai contrast. Is this really the same city. Nitin Gadkari.	$<T_c>$	$<T_l, T_p>$
Nitin Gadkari even police vans do not follow traffic rules	$<T_p>$	$<T_l, T_c>$
Nitin Gadkari Sir NH 8 a black spot on your department? Seems you are feeling helpless because of Haryana Govt	$<T_l, T_c>$	$<T_p>$

Table 2.14: Confusion Matrix Results for the Rule-based Classifier for Identifying Irrelevant (IRT), Nearly-Useful Tweets (N-UT) and Useful Tweets (UT)

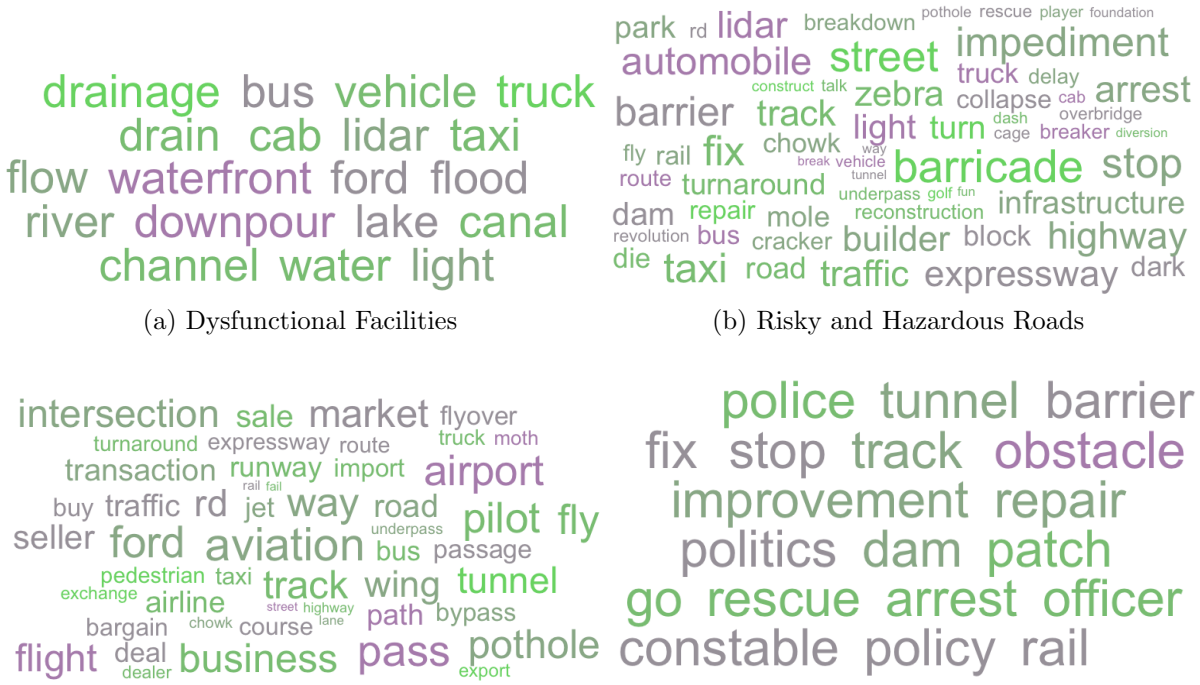
		Predicted			
		N-UT	IRT	UT	**Total**
Actual	N-UT	1088	131	59	1278
	IRT	376	569	17	962
	UT	254	34	94	382
	Total	1718	734	170	**2,622**

Figure 2.18 shows the distribution of original tweets classified into different categories while following the procedure of Non-complaint tweet classification and complaint report classification. Figure 2.18 reveals that using AISP classification method proposed in Section 2.6.1, we were able to classify a total of 20.5% (680 out of 3302) tweets as AISP distributed into 166 appreciation posts, 417 news and information sharing tweets while 97 reports are classified as promotional and advertisement tweets. Figure 2.18 reveals that a very small percentage of tweets (5%- 170 tweets out of 3302 reports) are identified as complete reports that contain all three important components of a killer road complaint. Whereas, the largest chunk of reported tweets is classified as incomplete or nearly-useful tweets (1718 reports out of 3302 tweets). Figure 2.18 also reveals that further only a very small percentage of tweets are convertible (N-UT-C) into complete or useful tweets (50 tweets out of 1718 nearly-useful tweets) while 97% (1668 out of 1718) of nearly-useful tweets have either landmark or concrete problem component missing from the tweets. Table 2.13 shows the example of tweets present in our dataset and classified as nearly-useful tweets which further cannot be enriched or converted (N-UT-N). Figure 2.18 reveals that due to the large percentage of reports with missing or incomplete information that is not possible to enrich, it is technically challenging to identify each and every complaint tweet efficiently.

In order to measure the performance of our classifier, we use standard metrics of Information Retrieval and compute the overall accuracy of the proposed approach. Table 2.14 shows the confusion matrix of the proposed classifier. Since the proposed approach is a multiclass classifier, we compute the performance of each class. Based on the results acquired by our rule-based classifier, we classify bad road related complaints with an overall accuracy of 67%. In addition to measuring the accuracy of our classifier, we also measure the overall recall value of the classification. Based on our results and the Table 2.14, we record a recall of 65%. As discussed in Section 2.6.2.2.2, the complaints reported to public agencies' accounts are user-generated content and lacks a standard format or terminology for complaining a report. Further, the excessive use of metaphor and sarcasm while reporting a complaint generates false alarms and impacts the overall accuracy of the classification.

Table 2.15: Distribution of Complaint Reports Classified and Labeled into 4 Different Categories. Further, the Table Reports the Number of Reports Classified into More than 1 Category.

	A	B	C	D
Dysfunctional Facilities (A)	402	345	282	44
Risky and Hazardous Roads (B)	-	1142	933	145
Poor Condition Amenities (C)	-	-	1042	92
Indiscipline and Carelessness (D)	-	-	-	252



(a) Dysfunctional Facilities

(b) Risky and Hazardous Roads

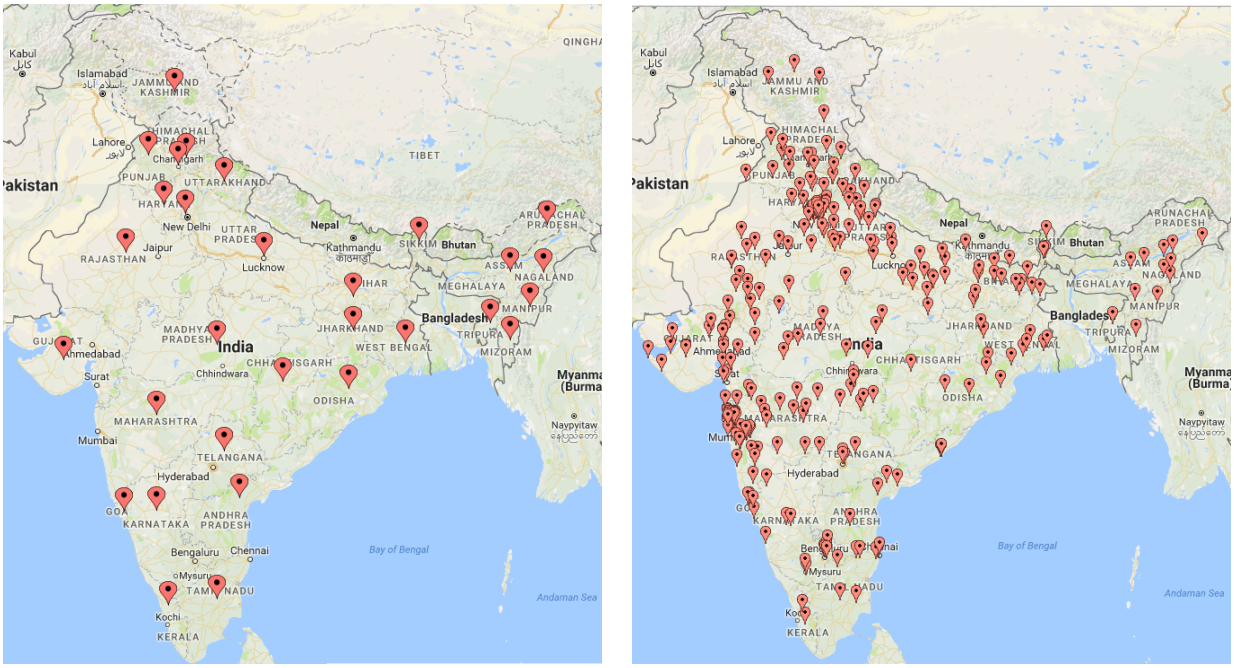
(c) Poor Condition Amenities

(d) Indiscipline and Carelessness

Figure 2.19: Word-cloud Presentation of Common Words used for Reporting the Issues Related to Different Categories of Complaints.

2.7.2.1 Characterization of Complaint Reports

As discussed in Section 2.6.2.2.2, based on the types of issues reported in killer road complaints, we categorize these complaints into four major categories: 1) Dysfunctional facilities, 2) Risky and Hazardous (accident prone), 3) Poor Conditions and 4) Indiscipline and Carelessness. We perform a feature-based characterization on the tweets classified into each of these categories. Based on our results and as illustrated in Figure 2.11, we find that several complaints belong to more than one categories. We, however, report the results only for the complaints that are classified into one or two categories. Table 2.15 shows the distribution of complaint reports (including nearly-useful tweets), the maximum number of complaints (1142 and 1042) are reporting about the dangerous and hazardous roads which are accident prone and poor conditions of amenities and roads respectively. While, the minimum number of complaints (252) are reported about the



(a) Distinct States (b) Distinct Places

Figure 2.20: Distribution of Distinct Geographical Locations (places and states) Identified in the Complaint Reports in our Experimental Dataset

risk caused by indiscipline or carelessness of driving. Table 2.15 also shows that further a large percentage of complaints (933) reported on the risky and accident prone roads are due to the poor condition of the roads. Figure 2.19 shows the word-cloud presentation of the terms commonly used in the complaints reporting the issues. Figure 2.19b shows that the complaints related to risky and accident prone roads contain the term and issues like *barricade, cracker, expressway, highway, chowk, zebra, street, turn, impediment, flyover, construction* and *builders*. Figure 2.19c reveals that *pothole, track, pedestrian, road, bypass, intersection, truck, traffic, passage, bus, underpass* are the commonly used terms in the tweets reporting about poor conditions of roads. Further, Figure 2.19d reveals that *barrier, stop, rescue, arrest, constable, policy, police, tunnel* are the common words used in the reports complaining about the carelessness of drivers or authorities.

In order to identify the source of maximum complaints on killer roads, we perform a characterization on the location feature of the tweets. Figure 2.20 shows the distribution of geographical locations identified in the complaints reports (complete and nearly-useful tweets). Figure 2.20a reveals that citizens post complaints to @MORTHIndia and @nitin_gadkari from all regions of India. Figure 2.20b shows the city-level distribution of complaint reports posted from all over the India. Figure 2.20b reveals that despite receiving complaints from every state in India, there are some states from where the maximum numbers are reported. The map presented in Figure 2.20b shows that the cities of Mumbai, Delhi, Haryana, Bihar and Uttar Pradesh report more complaints relative to the cities of Karnataka, Telangana, Odisha or North East regions of the country.

2.8 Conclusions and Future Work

Due to the immense popularity and wide reachability of the website, Twitter is being used by government and official public agencies to reach out to the public and resolve their complaints in a timely manner. It is also seen that public citizens use Twitter to report their complaints and grievances on various issues actively. However, due to the free-form nature of social media text and high velocity of data, automatic identification of these reports is a technically challenging problem. In this chapter, we conduct our experiments on open-source Twitter data and propose to use linguistic features for identifying complaint report tweets. Based on the type of grievances reported to the public agencies (1.complaints posted for awareness and bringing the attention of the government to the reported issues, and 2. complaints seeking for immediate action and response), we perform two case studies on Twitter. In the first case study, we analyze four official Twitter handle of Indian Government (@RailMinIndia, @IncomeTaxIndia, @dtpTraffic and @DelhiPolice) and investigate the efficiency of linguistic and computational features for identifying complaint tweets. To evaluate the performance of our proposed approach, we execute our SVM classifier for three kernels (linear, polynomial and RBF). Our results reveal that linear kernel one-class SVM outperforms RBF with a margin of 20% in the precision rate while polynomial and linear kernels produce the similar results with a difference of 1% to 2% of performance. However, the rate of misclassification in the polynomial kernel is higher than the linear kernel function. The linear kernelized SVM outperforms polynomial and RBF kernlized SVMs due to the sparse nature of data and high-dimensional feature space. Furthermore, linear kernel SVM avoids overfitting of data and improves the accuracy of classification. We further boost the accuracy of our proposed approach by combining three kernels into a cascaded and parallel manner. Our result shows that parallel ensemble classifier outperforms cascaded ensemble SVM. Using parallel ensemble technique, we improve the precision of complaint tweet classification by 20%.

In the second case study, we present an approach for automatic identification of complaints reported in road irregularities and poor road conditions. To conduct our experiments, we analyze the tweets posted to the official Twitter handles of ministry of road, transport and highways, Government of India (@MORTHIndia and @nitin_gadkari). We propose various linguistic features that demonstrate the key components (problem reported in the complaint, landmark or pinpoint location, city or location information) of a killer road complaint. Based on the available components in the tweets, we classify them into three categories: useful tweets, nearly-useful tweets, and irrelevant tweets. We further propose a mechanism to enrich the nearly-useful tweets and convert them into useful tweets. Our results show that a significant percentage of complaints posted on Twitter are incomplete and lack the major components in the report making them less likely to get addressed and resolved. We perform a characterization on complaint reports, and our results reveal that a maximum number of complaints are reported about the dangerous and accident prone roads while most of them are due to the poor condition of amenities. Further, the complaints are reported from all over the India while the maximum complaints are reported from Maharashtra, New Delhi, Uttar Pradesh, and Bihar states. In addition to extracting features for automatic classification of complaints and grievances report we also propose features that are strong indicators of a tweet to certainly not to be a complaint report such as appreciation, information sharing, and promotional tweets. In this chapter, we also address the challenge of free-form text in tweets and capture the dependencies between noisy text and semantics. We proposed a generalized micropost-enrichment algorithm that performs semantic and syntactic enrichment on tweets for improving the accuracy of linguistic features extraction.

Future work includes addressing the limitations of present study by improving the accuracy of location identification. Further, there are several complaints which contain humor, sarcasm, and ambiguity. Such tweets do not provide any information about the issue reported in the complaint. Our future work includes the use of sensemaking for extracting information from such ambiguous posts. The future work also includes creating a front-end data visualization tool and showing the status of complaints and grievances in the form of a dashboard.

Chapter 3

A Collision of Beliefs: Investigating the Dynamics of Religious Conflicts by Mining Public Opinions on Social Media

3.1 Introduction

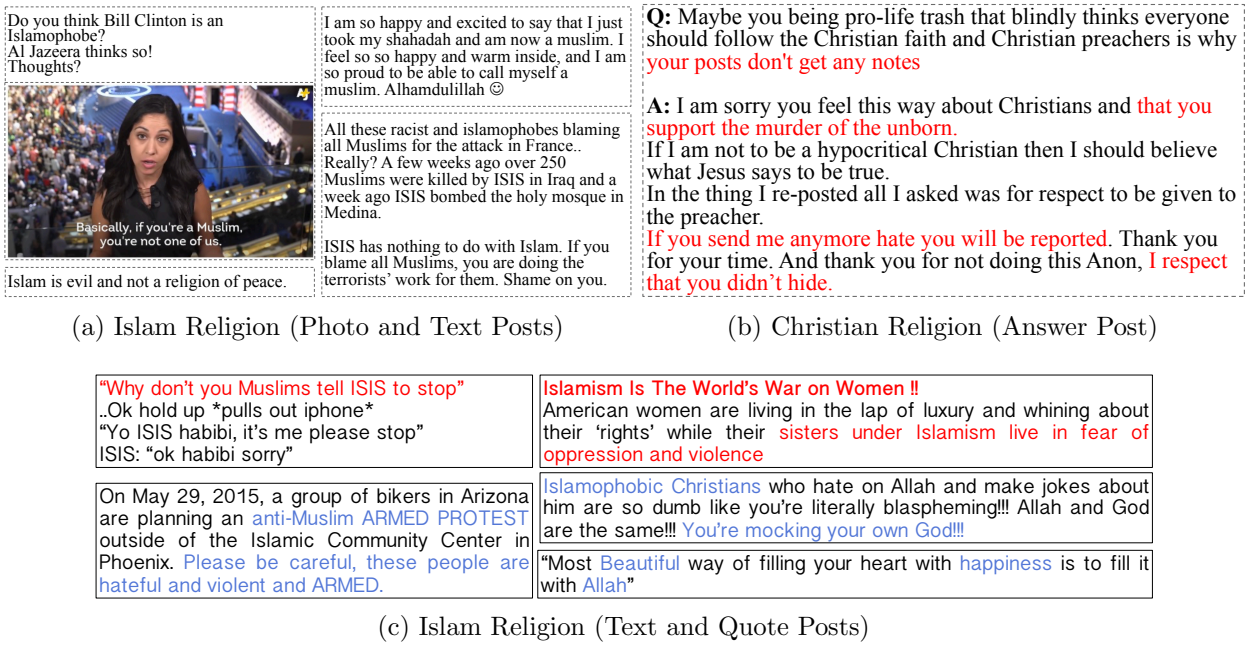
Research shows that the unexpected emergence of religion and faith in society, has led to the discrimination and violence against rival religious groups [116] and civil protests [1,2]. It is seen that people use various platforms (chat groups, online forums, blogs, social media) to share their beliefs and opinions about their religion [30]. Whereas, other like-minded people post extremist and hateful views towards other religions [117]. These groups of individuals take the leverage of freedom of speech and social media to post their sentiments and beliefs about a variety of sensitive topics including religion and race [117]. Despite several guidelines of social media platforms[3] and constraints of freedom of speech [118], people post racist and harsh comments against other religions that can hurt the religious sentiments of an individual or a community [119]. Figure 3.1 shows examples of several online posts indicating the conflicts in the context of Islamic and Christian religious beliefs and sentiments of authors. Figure 3.1 reveals that while some users posted defensive and promotional content about Islam religion; other users posted negative comments and insulting the beliefs of people believing in Islam. Further, some users only make posts to share information on real time incidents or news and not presenting any sentiment or argument for religion. As seen in the real world, many young age people and students get influenced by social media messages and join religious wars and radical groups [30]. Therefore, monitoring such content on social media and identifying religious conflicts within society, understanding the cause of such conflicts and arguments have become a major problem for the government, social scientist, and law enforcement agencies.

We conduct a literature survey in the area of political and religious conflict identification on social media. We find that over the past 3 decades, social science researchers have been conducting offline surveys

[1] http://www.hindustantimes.com/india/traitor-anti-national-and-intolerant-who-said-what-at-the-jnu-protest/story-zFhiONmAl93OETobVEjAHK.html

[2] http://timesofindia.indiatimes.com/india/JNU-students-protest-outside-Home-Ministry-detained/articleshow/54979642.cms

[3] https://www.tumblr.com/abuse/maliciousspeech



(a) Islam Religion (Photo and Text Posts) (b) Christian Religion (Answer Post)

(c) Islam Religion (Text and Quote Posts)

Figure 3.1: Several Concrete Examples of Different Tumblr Posts Showing Differences and Conflicts among Bloggers on Islam (Multiple Posts) and Christian (Single Post) Religions

for identifying religious conflicts within society. Whereas, the area of identifying such conflicts by using computer science applications is relatively unexplored. Based on our analysis, we divide our literature survey into following four lines of research:

1. **Offline Data and Manual Analysis:** Swinyard et. al. [120] and Wilt et. al. [121] conducted surveys to examine the relationship between religious and spiritual beliefs and emotions of people such as happiness and anxiety. Yang et. al. [122] present a study on the impact of low coverage of HindRAF event in media causing the religious conflicts among citizens of Malaysia.

2. **Offline Data and Automated Analysis:** Vüllers et. al. [123] present a study on the religious factors of 130 developing countries. Their analysis reveals that the clashes between religious groups and attacks by religious actors are the primary cause of religious conflict within state and community. Basedau et. al. [124] used logistic regression approach on the same dataset to identify several discriminatory religious factors that cause conflicts, religious violence and grievances.

3. **Online Data and Manual Analysis:** In addition to the social science researchers, various non-profit organizations like Pew Research Center[4], Berkley Center for Religion, Peace and World Affairs[5] and United States Institute of Peace[6] conduct online polls, offline statistical and text analysis on blogs and social media data. The aim of their studies is to identify the religious beliefs and issues within local and global regions. Some of the recent studies of Pew Research Center include the global trend and projection of population growth of various religions, gender gap in religious commitment of Muslim

[4] http://www.pewresearch.org/topics/religion-and-society/
[5] https://berkleycenter.georgetown.edu
[6] http://www.usip.org/about-usip

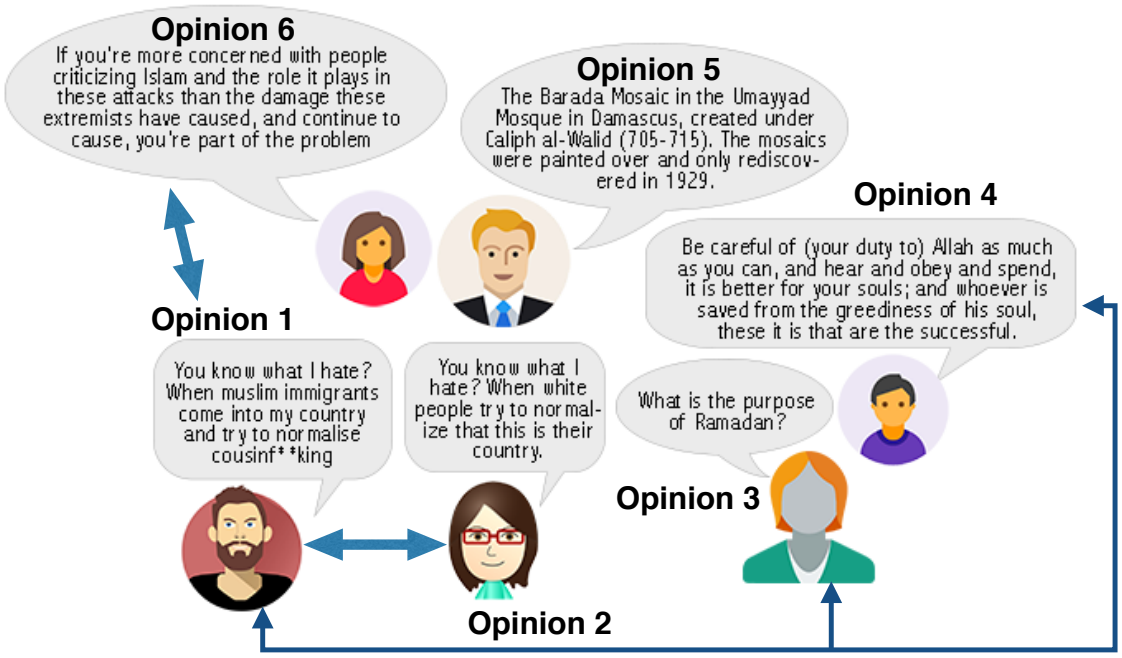


Figure 3.2: Demonstrating the Contrast in Public Opinions (extracted from Tumblr website) Reflecting the Conflicts in Islamic Religious Beliefs and Sentiments of People.

and Christian communities, increment and decrement rate of government restrictions on religion and social hostilities.

4. **Online Data and Automated Analysis:** Chesnevar et. al. [125] propose an opinion tree using IR and argumentation technique for identifying conflicts and confronting opinions in E-Government contexts. They conduct their analysis on Twitter messages and identify the polarity (positive, negative and neutral) of contrasting arguments.

3.1.1 Motivation

Current state-of-the-art reveals that various social science researchers have been conducting offline surveys for identifying religious conflicts within society [120] [123]. However, the immense amount of data available on social media in the form of comments, communications, discussions has been largely ignored in existing works and raises three major limitations:

1. The offline surveys are conducted among selected groups of people. For example, volunteers are from a specific region, belonging to a particular age-group, or performing specific activities such as going or not going to Church. Hence, the offline surveys lack the generalized claim of conflict of beliefs.
2. The offline surveys or in-person questionnaire lack the subjectivity and real opinions of the people. Despite conducting an open survey the opinions are highly likely to be influenced by others and resulting in inaccurate statistics.
3. In offline polls, the identity of people volunteering for surveys is not anonymized.

The work presented in this chapter is motivated by the prior literature and a need to develop an automatic solution to identify religious conflicts among social media users. However, automatic identification of religious beliefs and faith by mining user-generated data is a technically challenging problem. As discussed in Chapter 1, the presence of noisy content such as misspelled words, short text, acronyms, multilingual text and incorrect grammar decreases the accuracy of linguistic features and Natural Language Processing tools [30]. Further, the presence of ambiguity in posts and the intent of author makes it difficult even for human annotation [117]. We conduct our experiments on Tumblr micro-blogging website since Tumblr overcomes the limitations of offline surveys carried out in previous literature. Tumblr is the second most popular micro-blogging service facilitating its' users to make posts in 8 different multimedia formats (image, audio, text, video, URL, quote, chat and ask). Unlike Twitter, despite being a micro-blogging platform, Tumblr has no character limit for captions, body content or tags of a post. Due to no restriction on content length bloggers are allowed to write longer posts and can express their opinions, emotions, and thoughts on a topic in open and descriptive manner which is not there in in-person or offline surveys. Tumblr facilitates blogger to send anonymous messages which further gives them leverage to express their opinions without revealing their identity. Furthermore, Tumblr provides a community space to bloggers sharing similar interest without caring about the real identity or credibility of the blogger [50].

3.2 Research Contributions

In contrast to the existing studies, our work makes the following novel and unique technical contributions:

1. While, prior literature conducts offline surveys for religious conflict identification, we present the first study on automated identification of religious beliefs, opinions, and faith in global public communities. We propose to use the online social media data and user-generated posts for identifying the religious sentiments of people.
2. We address the challenge of region-specific surveys by mining multilingual posts on social media. We use English as our base language and counter the challenge of multilingual scripts by translating all non-English posts into the base language.
3. We propose to use topic modeling based features that can automatically classify religion based posts. We further introduce computational linguistic features for mining religious sentiments and identify contrast in the dynamics of conflicts.
4. We investigate the efficacy of multiclass semi-supervised classifier across various dimensionality reduction techniques for classifying social media posts into multiple dimensions of conflicts. To validate our approach, we conduct our experiments on an open source data collected from Tumblr website. We publish the largest database of Tumblr posts associated with tags and keywords related to religious topics and make it publicly available to the research community.

3.3 Experimental Setup

3.3.1 Experimental Dataset Collection:

Figure 3.3 shows the high level architecture for the experimental setup. In order to collect our dataset, we create a list of tags that are commonly used in religious post and call them as seed tags S_{tag} (hinduism, islam, muslim, religion, isis, jihad, jews, christian, islamophobia and jews). We use Tumblr search API[7] and

[7] `https://www.tumblr.com/docs/en/api/v2`

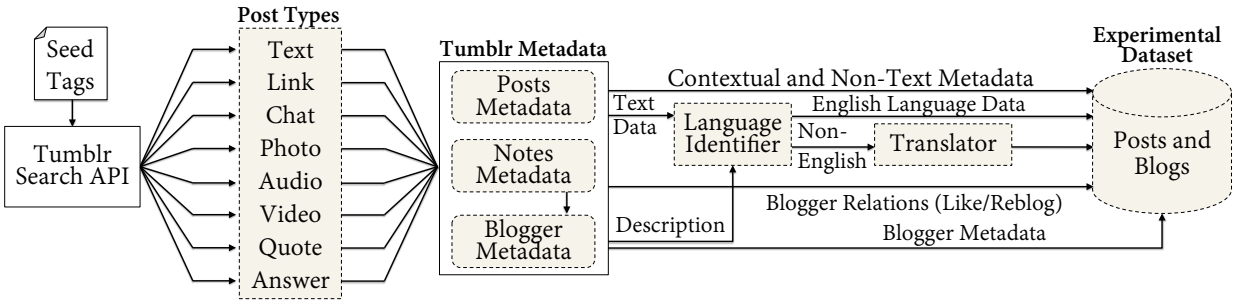


Figure 3.3: A General Research Framework of Experimental Dataset Collection and Enhancement- Primarily Consisting of Tumblr Posts Collection, Metadata Extraction, Non-English Posts Translations and Combining Metadata of Posts and Bloggers

extract all available posts that contains any of these tags. Tumblr API allows us to extract only 20 most recent posts. Therefore, we use an iterative process and extract the posts created before the timestamp of 20^{th} post in search results. We extract all eight types of posts for these tags published since 2007. Using Tumblr Search API, we were able to extract a total of 107,586 posts (Religion: 7,019, Islam: 6,143, Muslim: 31,113, ISIS: 28,768, Islamophobia: 2,927, Christian: 923, Jews: 17,785, Judaism: 4,515, Hinduism: 1,826, Jihad: 6,567). Figure 3.4a shows the statistics of total number of posts collected in each category of Tumblr posts. Figure 3.4a reveals that maximum number of posts consisting of religious tags are either posted as photo (49,072) or text (34,902) posts. Similarly, URL or link types of posts (10,062) are relatively higher in comparison to chat (507), audio (390) and answer/ask box (1,077) categories.

Since, each category in Tumblr has different and unique attributes, we extract the type of each post and acquire related metadata accordingly. Since the aim of this study is to build a multiclass text classifier, we keep only the textual metadata of each post. For example, the caption of photo and video posts, phrases in chat posts and question-answer in an answer post. Table 3.1 shows the list of all metadata extracted from various types of Tumblr posts. Further, Table 3.1 reveals the metadata used for our experiments and filtered due to the low irrelevance. We merge all the posts collected for all 10 tags and remove all duplicate entries from our dataset. Figure 3.4a shows the statistics of number of unique posts collected in each category. Figure 3.4a reveals that there is only slight difference in originally collected posts and unique posts. It happens because while re-blogging a post in Tumblr, users are least likely to add new tags. While Tumblr Search API extracts only the posts that contain a search term in their associated tags. Since the aim of this study is to identify various linguistic features from textual metadata of posts and many religious posts are made in different regional languages. We determine the language of the textual post (text, chat, quote, and answer) and textual metadata of other posts (audio, video, URL, and photo). We use Yandex language and Translate API[8] to translate all non-English content to the English language. The Yandex language and Translate API provides access to the Yandex online machine translation service[9]. The API supports more than 70 languages and translate the input text irrespective of the size. To translate a word, sentence or paragraph, it requires providing the source and targeted language. We use Yandex API to translate Tumblr posts present in our experimental dataset. For the posts written in multilingual scripts, we use source language based on the maximum content written in one language. We further remove the posts that contain no text. For example, a photo post with no caption, a post consisting of only external URL or only emojis. Emojis are the 'picture characters' (😊 😝 😇 😘) commonly used to express emotions in a text-based communication in smartphone chats and social media sharing [126]. For video posts consisting of no text, we extract the title of videos by parsing the URL mentioned in the post. Figure 3.4a shows the statistics of the

[8] `https://tech.yandex.com/translate/`

[9] `https://translate.yandex.com`

Table 3.1: List of Tumblr Posts and Bloggers Metadata Extracted from Tumblr Website. Further, Showing the Metadata Used for Our Experiments and Discarded due to the Low Relevance.

Post	
post_id	unique id of a post
blogger	username of the blogger
timestamp	timestamp of the post
date	date and time
state	published, drafted, queued
slug	url of the post
format	html or markdown
short_url	compressed and direct url of the post
note_count	number of notes on a post
type	unique type of the post
reblogged_key	link to the source post
reblogged_from	id of parent blogger

AnswerPost	
post_id	unique id of the post
name	name of the blogger or anonymous
~~url~~	~~null for anonymous~~
question	question that is asked
answer	answer posted by the blogger

AudioPost	
~~post_id~~	~~unique id of the post~~
~~audio_url~~	~~url of the audio file~~
~~track_name~~	~~name of the track~~
~~artist_name~~	~~name of the artist~~
~~album_name~~	~~name of the album~~

ChatPost	
post	unique id of the post
label name	names used in the post
phrase	content posted in chat

UrlPost	
post_id	unique id of the post
title	title of the post- not url
~~url~~	~~source url~~
description	description of the post

Tag	
post_id	unique id of the post
tag	tag associated with the post

Notes	
post_id	unique id of the post
blogger	author of the post
note_type	like or reblog
note_by	who liked or reblogged the post
note_timestamp	timestamp of note

Blogger	
blogger	id of the blogger
blogger_name	name of the blogger
~~last_updated~~	~~timestamp of last activity~~
~~blogger_ask~~	~~if questions are allowed~~
number_posts	total number of posts
blogger_title	title of blogger
blogger_description	description of blogger

PhotoPost	
post_id	unique id of the post
caption	caption of post
~~photo_url~~	~~url of all photos shared in post~~

VideoPost	
post_id	unique id of the post
caption	caption of the post
~~source~~	~~source of video shared in the post~~

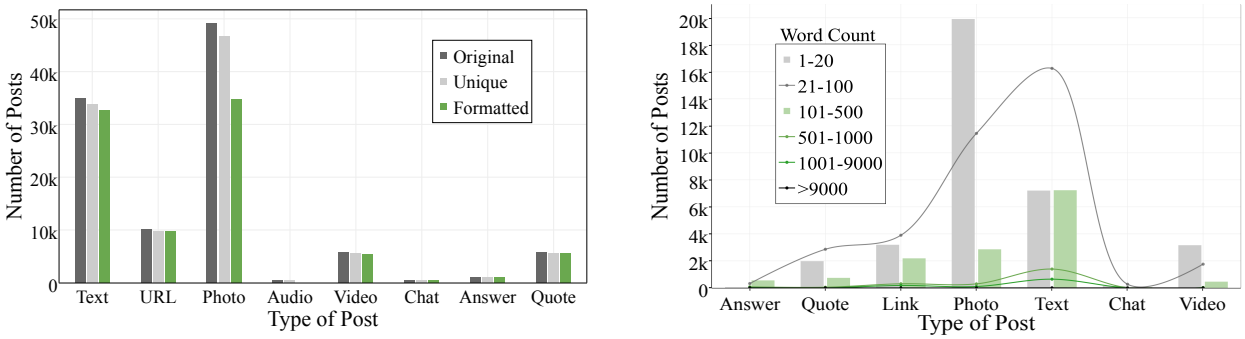
QuotePost	
post_id	unique id of the post
~~source~~	~~source of the quote- if any~~
content	content of the quote

TextPost	
post_id	unique id of the post
body	content of the post
title	title of the post

number of posts obtained in each category after cleaning the data. We, however, discard the audio posts due to the very short text present in the metadata such as artist name and album name. After pre-processing of the raw data, we were able to acquire a total of 89,803 posts calling them as our experimental dataset. We publish our pre-processed and enriched data on Mendeley and make it publicly available for the research community for benchmarking and comparison [65].

3.3.2 Data Annotation

To create the ground truth for our dataset and a training dataset, we use 89,803 pre-processed posts for further annotation which spans only 83.4% of the collected data. In order to remove the bias from our annotation, we hired a group of Tumblr users who had an experience of 2 to 3 years of using Tumblr website.



(a) Number of Posts Extracted and Formatted for Seed Tags

(b) Number of Posts Consisting of Different Word-Length (Varying up to 13,318)

Figure 3.4: Various Data Statistics of Number of Posts for 8 Types of Categories Available on Tumblr

We published a post on Tumblr and asked bloggers to volunteer for data annotation. In a span of one week, 34 bloggers replied and agreed to annotate an average of 30 posts. We declined 4 bloggers who joined Tumblr recently. Among 30 bloggers, only 23 bloggers reverted with 690 annotated posts among which 6 posts were sampled more than once. Due to a large amount of Tumblr posts and challenge of creating ground truth [30]; we used only these 684 posts for creating our training dataset.

3.3.3 Experimental Dataset Characterization

Figure 3.4b shows the distribution of posts in each category based on the number of word counts. Despite being a micro-blogging website, Tumblr has no word limit for text content (posts and tags). However, due to the presence of micropost, short-text, and very long text, it is challenging to extract the similar features for each post present in our experimental dataset. Therefore, we divide our dataset into 6 categories and compute the number of posts in each category. Figure 3.4b reveals that maximum number of posts have a content of 21 to 100 word length while only 4 posts have word length more than 10000. Among which Photo and Texts have the maximum number of posts (19908 and 7206 respectively). Figure 3.4b also shows only text posts have a reasonable number of posts (1394) that have more than 500 words in the textual content.

3.3.3.1 Evolution of Religion Based Posts

We also study the evolvement of various topics and tags on Tumblr and their popularity on the website. We identify the popular tags P_{tag} (different than the seed tags) that are commonly used in religion based posts on Tumblr and compute the number of posts consisting of these tags. For example, Allah, jihad, Islamic state, God, faith, and racism. We also extract the number of notes (likes and reblogs) on each post in our experimental dataset. We compute the number of posts in each month having more than an average number of notes (50) in our experimental dataset. Figure 3.5 shows the timeline based graph of the monthly distribution of the total number of posts in our dataset, the number of posts consisting of other popular tags and more than 50 notes. Figure 3.5 reveals that the number of posts consisting of P_{tag} follows the same patterns as the number of posts consisting of S_{tag}. We observe that maximum of these tags are related to Islam religion. For example, among seed tags, Muslim keyword has the maximum number of posts (35254) on our experimental dataset. While among other popular tags, Allah, Quran and Syria key terms are associated with 9223, 6215 and 5523 posts respectively.

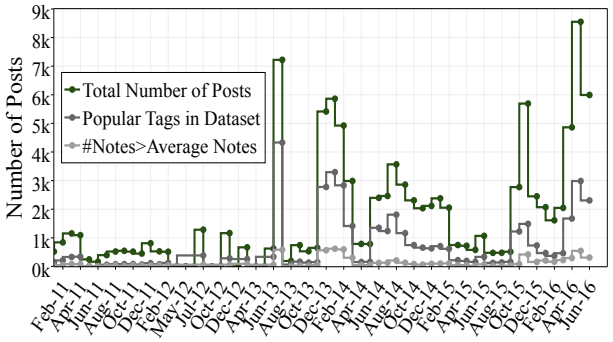


Figure 3.5: A Timeline Based Review of Number of Posts Extracted, Consisting of Popular Tags and More than Average Notes

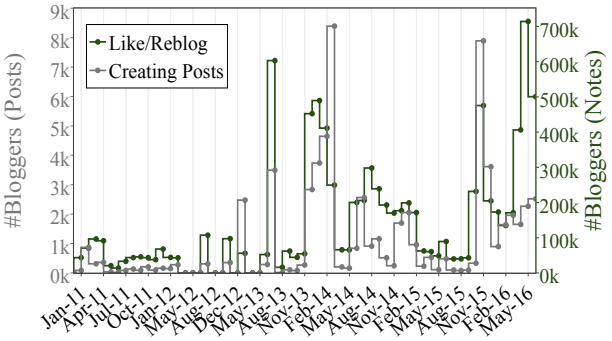


Figure 3.6: Number of Bloggers Creating New Posts and Participating in Community by Liking and Re-blogging These Posts

3.3.3.2 Community Participation

Experimental dataset reveals that Tumblr is actively being used as an active platform for posting content related to religion/spiritual topics. However, it is also important to analyze the involvement of bloggers posting about such topics. Figure 3.6 shows a timeline based graph of community participation on Tumblr posting about various topics related to religion. The x-axis shows the monthly timeline since 2011 while Y-axis indicates the distribution of the number of bloggers participating in the community. Y-axis on right of the Figure 3.6 shows the number of bloggers creating new and distinct posts while Y-axis on the left side of the Figure 3.6 indicates the number of distinct bloggers liking or reblogging these posts. We observe that over the past six years a majority of bloggers have been actively posting religion based content on Tumblr website. While in last three years (2014-2016) number of distinct bloggers has reached up to approximately 85K (89, 803). Figure 3.6 also reveals that not only the users who create new posts but also the users liking and further re-blogging these posts are also increasing rapidly. The number of distinct bloggers liking and re-blogging these posts shows the huge community participation of Tumblr bloggers. Figure 3.6 shows that in 2016, these numbers reach up to 0.7 million (recorded in May 2016).

3.4 Dimensions of Conflicts

We conduct a survey among 50 people (10 groups of 5 people) having different religious, spiritual and race beliefs. We asked them questions regarding their activity on social media websites and if they share the religious and race based posts. We showed them recent incidents related to religious conflicts and war and further asked them the type of sentiments they would like to share on social media (Facebook, Twitter, and Tumblr). Based on the reviews and types of emotions of these 10 groups, we find 9 dimensions that can lead to conflicts and disagreements among users on social media websites. We further conduct another survey among three different group of people. We randomly selected 10 graduate students of our department and 30 Tumblr bloggers (followers on authors' personal Tumblr account) for our survey. We also conduct our survey among 20 random people (friends, people meeting in non-technical gatherings and public citizens) between the age group of 20 to 35 years irrespective of their profession, education, and religion. For the set of graduate students and randomly selected people, we survey only those people who actively use social media platforms (Facebook and Twitter). We conducted a small questionnaire consisting of questions related to their activities on social media platforms *e.g. how frequently they make a religion based posts on social*

Table 3.2: Concrete Examples of 11 Dimensions and 3 Polarities of Religious Beliefs and Sentiments in Tumblr Posts Created About Christian Religion and Community

Type	Post Content
C1	Pray for abortion access. People deserve easy access to abortion services.
C2	In a show of solidarity, Muslims are standing with Christians and giving up guilty pleasures for lent.
C3	Doesn't the Bible teach us not to take a life of another? To turn the other cheek and not respond with violence? Isn't better to die and be in heaven then kill and stay on earth?
C4	I'm still over the moon about God. I'm in total awe that He not only hears me, but actually listens and does something about it. I feel so loved and acknowledged.
C5	If you're a Christian and voted for Trump I wanna ask you a question. What does it feel like to go against everything God wanted for us?
C6	Jesus himself could crawl out of his grave, take me by the hand, and point me to salvation and heaven. I would say no. I would seriously 100% rather die as a Jew then live for even a millisecond as a Christian. So stop trying to convert me to Christianity because it is not going to happen.
C7	Burn churches not calories. Christianity is stupid!- Well I am not the only one that feels the same way.
C8	So this dude that was running in local elections for council said women who have abortions are worse than ISIS
C9	I feel like a bad Christian. I have so much hate in my heart after this election, at Drumpf, at his voters, at my country. I know I should turn the other cheek and love radically and protest without hating but I'm so angry. I feel like I can't let that hate go, not so soon. But I need to and I'm furious at myself
C10	Imagine the peace we'd all have without religion. Wouldn't it be a better world?
C11	When Christ has a cold he sneesus

media or react to other religious posts. We created a set of 30 posts about different religions and asked for their opinions. Based on our survey, we decided 11 dimensions of public opinions that can be used to define the contrast of conflict among people: Information Sharing (IS), Query, Not a religion based post (N/A), Disbelief, Defensive, Annoyance, Insult, Disappointment, Sarcasm, Ashamed and Disgust. Table 3.2 shows the examples of 11 Tumblr posts created about Christian religion and community reflecting the different dimensions of public opinions about the community. We discuss each of these categories below:

1. **Not a Religious Posts (C1):** Based on our observation on Tumblr website, we find that many posts contain religion based tags and terms but have no relation to the religious beliefs.
2. **Information Sharing (C2):** In information sharing posts, blogger are only willing to share some information related to a religion or any incidents caused by religious activists. These users do not depict any emotions in their posts and only share such posts for awareness.
3. **Query (C3):** Based on users' answers and our observation on Tumblr website, we find that many bloggers ask questions related to various religions. These users seek for the information and do not depict any positive or negative emotion but a curiosity towards that religion.
4. **Defensive (C4):** In defensive posts, users share positive, sensitive and justifying comments about a particular religion. We observe that on Tumblr, users make such posts by sharing their beliefs in the form of quotes. However, it is challenging to identify a defensive post if the user is comparing two religions and sometimes it can be misclassified as a promotion post due to the presence of persuasive content [117].

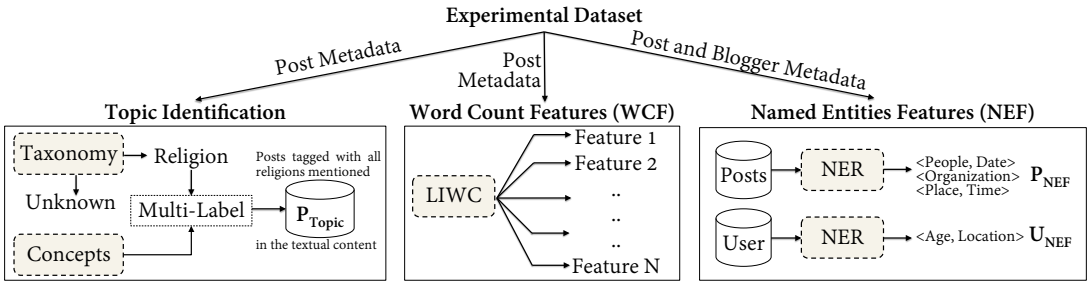


Figure 3.7: A General Research Framework for Extracting Linguistic Features from Tumblr Posts- Primarily Consisting of Topic Modeling, Linguistic Word Count and Named Entity Based Features

5. **Disappointment (C5):** Similar to annoyance and disagreement posts, in these posts, users share their difference in opinions. However, such posts contain higher rate of sadness and a lower rate of anger. These posts target a community and users show their disappointment towards a religion or race for certain actions.
6. **Annoyance (C6):** These posts on Tumblr depict the frustration of the blogger towards a religion or community. These posts consist of an informal and personal writing rather than a formal structure. We observe that majority of such posts are published after influencing from some religious incident.
7. **Insult (C7):** The nature of these posts is similar to sarcastic posts. However, the aim of these post is not to make fun or jokes but to post rude and harsh comments that can hurt the sentiments of targeted community.
8. **Disgust (C8):** Based on our survey and observation, these posts have content similar to annoyance and disappointment posts. However, the emotional range of these posts is higher than the disappointment and lower than the annoyance posts. A majority of these posts consist of factual content and written quoting real word incidents.
9. **Ashamed (C9):** These posts are created by the people of a religious group posting negative comments about their community. In the case of absence of use of correct pronouns, identification of such posts is technically challenging since these posts have similar features to disappointment (C5) and annoyance (C6) posts.
10. **Disbelief (C10):** In such posts, users shows their disinterest towards all religion, sometimes making their posts ambiguous with the posts having higher negative polarity.
11. **Sarcasm (C11):** Bloggers on social media misuse the freedom of expression by sharing funny comments about a religion and community (referred as trolls and memes). We observe that majority of sarcastic posts on Tumblr are published in the form of photos and reaction GIF images.

3.5 Constructing Feature Vectors

3.5.1 Feature Identification

In this Section, we discuss various linguistic features extracted from Tumblr posts. Figure 3.7 shows the high-level framework for feature extraction from Tumblr posts and bloggers metadata, primarily consisting

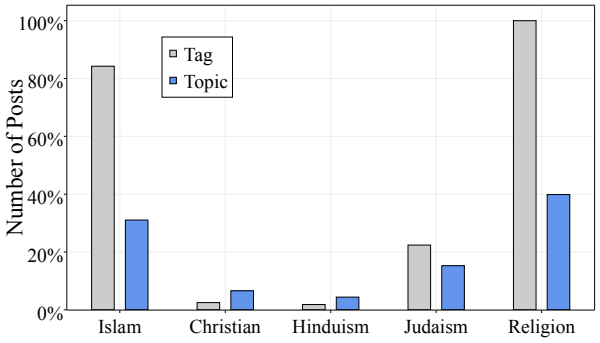


Figure 3.8: Relative Percentage of Number of Posts Consisting of a Religion Based Tag and Topic

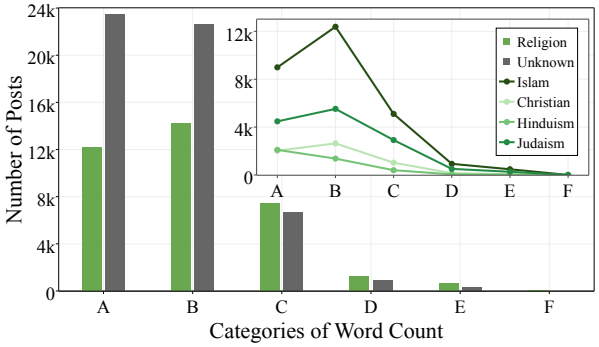


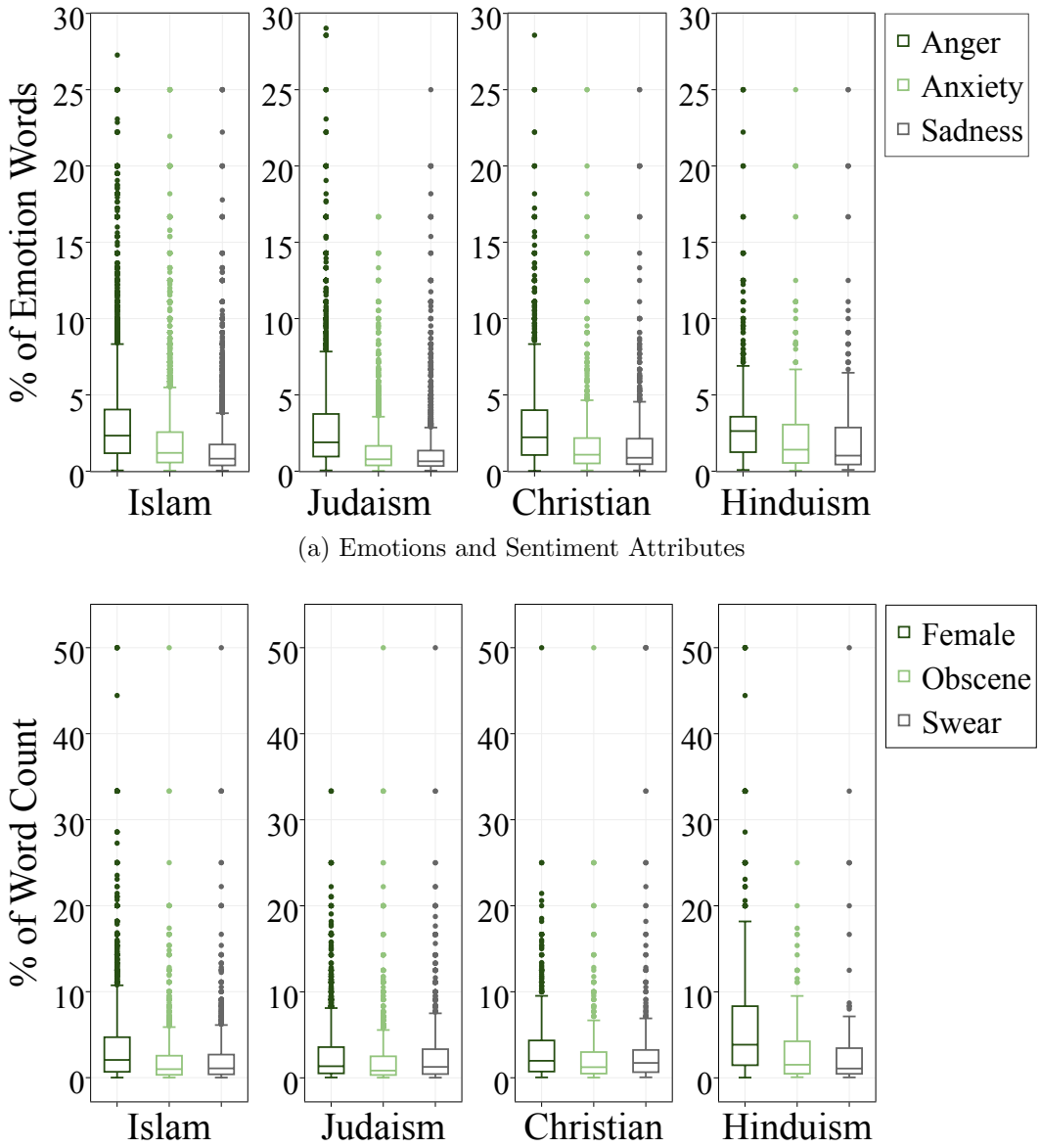
Figure 3.9: Word Length Based Distribution of Religion and Unknown Posts

of three phases: topic modeling, word count based features and named entity recognizer. As shown in the Figure 3.7, we extract the linguistic and textual metadata of Tumblr posts and bloggers and use them for extracting the features for further analysis. As discussed in Section 3.3.1, for different types of Tumblr posts, we extract different available data. For example, we extract caption of video and photo posts while title and description of text and quote posts. We discuss each of the linguistic features extracted from our experimental dataset in the subsequent subsections:

3.5.1.1 Topic Modeling

During our survey for identifying the dimensions of conflict, we observe that many users add religion based tags in their posts while the content of the post is irrelevant to any religion or community. Since our experimental dataset is collected using a keyword flagging based approach, we identify the topic of each post to filter the irrelevant posts. To perform the topic modeling on our experimental dataset, we use AlchemyLanguage [114] and extract the taxonomy of each post. Figure 3.8 shows the statistics of the number of posts consisting of religion based tags and actually discussing those religions. Figure 3.8 reveals that among all posts (85% of the experimental dataset) consisting of seed tags related to Islam religion (islam, muslim, islamophobia, isis and jihad), only 31% of the posts are about Islam religion. Similarly, among $20,106$ posts (22% of the experimental dataset) consisting of judaism and jews tags, only 15% ($13,695$) posts belong to Judaism religion. For each post, we assign a binary value where 1 denotes the topic (religion based post), and 0 denotes the non-topic (not a religion specific post). We further extract the name of the religion being discussed in the post since a post can have content about more than one religion. For example, in the following post *"KKK burns black Churches even tho they claim Christianity as their religion and ISIS blows up mosques even tho they claim Islam as their religion."*; author mention about both Islam and Christian religion. In our experimental dataset, we find that only 40% of posts ($35,799$) belong to a religious topic (Islam, Hinduism, Christian and Judaism). While, the remaining 60% of posts ($54,004$) only contains religious tags but do not contain the content related to a religious group or community.

Since, the performance of AlchemyLanguage API [114] varies for different size of text input; we divide our dataset into 6 categories based on the word count (WC) of each post. A: 1-20, B: 21-100, C: 101-500, D: 501-1000, E: 1001-9000 and F: WC >9000. Figure 3.9 shows the WC based distribution of posts classified as topic (any religion) and unknown. In comparison to Figure 3.4b, Figure 3.9 reveals that category A and B have the maximum number of posts in our dataset. But due to the presence of short text only 33% of the posts are classified as the religion based post while the number of relevant posts increases as the length of content increases. We see a similar pattern for C, D and E categories of posts. Since the number of posts



(a) Emotions and Sentiment Attributes

(b) Mention of Linguistic Attributes

Figure 3.10: Distribution of LIWC Features for Tumblr Posts Identified as Religion Posts

having more than 9000 terms is only four, values of topic and unknown posts are negligible. The statistics reveal all four posts to be the topic related posts.

3.5.1.2 Linguistic Inquiry and Word Count

In order to compute the correlation between various linguistic features and sentiments, we use an open source API by LIWC- Linguistic Inquiry and Word Count [127]. LIWC is a linguistic analysis based API

Table 3.3: Concrete Examples of Tumblr Posts Showing the Presence of Named Entities for Mapping them with Real Time Incidents.

ISIS is honoring **President Obama**. He is the founder of **ISIS**. He is the founder of **ISIS**, OK? He's the founder. He founded **ISIS**. And I would say the co-founder would be crooked **Hillary Clinton**.
real talk: why hasn't **donald trump** been cursed to have snakes and toads fall from his mouth whenever he speaks
I honestly CANNOT BELIEVE how popular that '**Hillary** is an **advocate** for rapists' meme is when it takes literally one google search to find out that it's completely inaccurate. I CANNOT BELIEVE there are some people who have let this meme convince them to vote for a man WHO IS AN ACTUAL RAPIST #**DonaldTrump**. I AM STEAMING.
7/27/2016 - Protesting the **Trump** rally in downtown **Toledo**. #ToledoTrumpsHate
If terrorism has no religion, than name a **Christian** suicide bomber. Name the last **Buddhist** bombing. A confusion shooting. Name someone who yelled ? deus vault? before opening fire on gays. Play the audio of someone saying ? **God** is good all the time, and all the time **God** is good? right before flying a plane into a tower. Any way you spin it, **Islam** is not a religion of peace. It brings with it death and violence wherever it goes. If you think **Islam** is a religion of peace, than turn on the news and open a **Quran**.

that captures the hidden and real-world behavior of words used in daily communications. As the name suggests, LIWC counts the words present in a text that defines a psychologically meaning. For example, attentional focus, emotions, social relationships, thinking styles, personalities and individual differences. We use LIWC to extract several features from a social media text that reflect the psychological behavior of the author. We obtain a total of 45 features grouped into 14 categories of linguistic dimensions. To identify the sentiments and emotions of the bloggers, we compute the relative percentage of the emotions e.g. *sadness, anxiety, anger, happiness*. We measure the authenticity and personality traits of authors by computing the summary of language variables in a post e.g. *analytical thinking and authenticity*. Further, to identify the personal beliefs and relation with the real world incidents, we compute the percentage of *sexual* terms, mention of *family, friends, male* and *female* references in a post. In order to identify the level of aggression and certainty of a post, we compute the percentage of use of informal language such as *swear words, slangs* and *fillers*. Apart from these features, we also compute the presence of various other linguistic dimensions such as *pronouns, negations, interrogatives words, cognitive process, perceptual process, power, time orientations* (mention of past, present or future incidents), *time* and *personal concerns* (work, religion, and death). LIWC computes the relative percentage of these features for each post. For example, if a post contains a total of 200 words and 30 terms reflecting swear words and slang language in the content then the score of swear attribute will be 15%.

Figure 3.10a reveals that on an average the maximum number of the key terms showing anxiety in their posts are approximate 10%. Whereas, the relative percentage of phrases showing anger in the post are comparatively higher in Hinduism posts. On a contradictory, in Islam and Christian religion based posts, the percentage of anger terms is lesser than Hinduism religion based posts while they have a large number of outliers. Similarly, despite having a low median value in sadness terms, the number of outliers are reasonably higher in Islam and Judaism posts. Further, Figure 3.10b shows that the large percentage of obscene and swear terms in religion based posts shows the conflicts among bloggers. Figure 3.10b also reveals that a lot of discussion in religion based posts is about women. For example, in all religion based posts in our experimental dataset contains an approximate of 10% women related key terms. Our analysis also reveals that Islam, Christian, and Judaism religion posts have a similar number of outliers mentioning female words, obscene and swear words showing the conflicts among these communities.

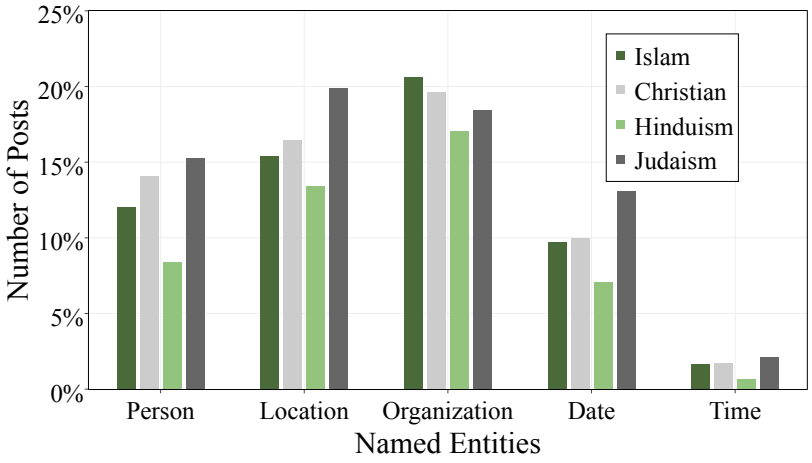


Figure 3.11: Religion Based Distribution of Posts Consisting of Featured Named Entities

3.5.1.3 Named Entities Based Features

As discussed in Section 3.4 and shown in Figure 3.5, majority of religious conflicting posts are created after an incident influencing the beliefs and sentiments of users. We observe that such posts contain the mention of various people, personalities, and organizations. For example, the name of prophets, political parties, religious communities and terrorist organizations. We use Stanford Named Entity Recognizer API[10] and extract five named entities for each post. We extract the person names, location, organization, date and time in a post to check the credibility of a post and improve the efficiency of computational linguistic features. Table 3.3 shows the examples of three posts created after the attack happened in Paris. Examples reveal the conflicts of sentiments and beliefs between two different communities. Further, we extract the age and location from the user profile and enhance the blogger data for making it publicly available.

Figure 3.11 shows the distribution of five named entities in various posts related to Islamic, Christian, Hinduism and Judaism religions. Figure 3.11 reveals that a very small percentage of posts contain "time" information while the "date" entity is mentioned in 10 to 15% of the posts. It is probably associated with the fact that the people discuss several incidents that happen around the religious beliefs. Figure 3.11 also reveals that despite the subjectivity in Tumblr posts, due to the variation in word count of each post, only 10 to 20% of the posts contain the person, location, and organization names in the post. Further, the English translated post might change the syntax of the content leading to the inaccuracy in named entity recognition.

3.5.2 Features Selection

3.5.2.1 Using All Features (FS1)

In the first iteration, we use all 45 linguistic, sentiment and text-based features extracted using LIWC. We train our model on available features and investigate the efficacy of classification of Tumblr posts into 11 dimensions of conflicts.

[10] `http://nlp.stanford.edu/software/CRF-NER.shtml`

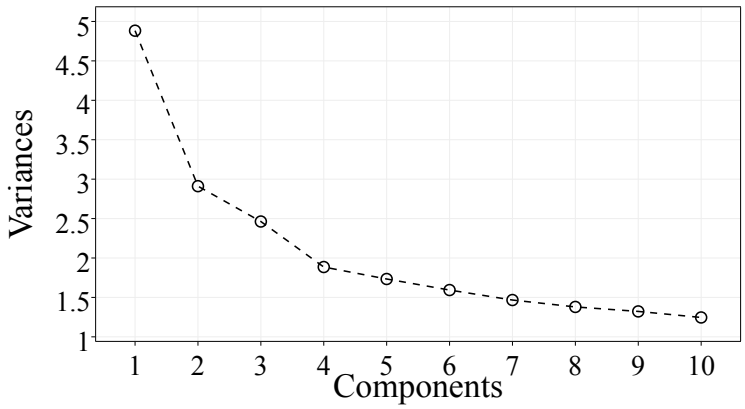


Figure 3.12: Relationship Between Variance and Components in Principal Component Analysis

3.5.2.2 Principal Component Analysis (FS2)

In the second iteration, we use Principal Component Analysis (PCA) to project our high dimensional data in small dimensions. PCA is a dimensionality reduction technique that reduces the number of original dimensions (or feature vectors) by transforming them to an artificial set but retaining the same information as original data [128]. We compute the covariance metrics among all feature vectors and identify the components characterizing the complete data. For $n = 45$ vectors in our experimental data, we get n eigenvectors. We select the first $p = 10$ eigenvectors having maximum eigenvalues and discard the ones with less significance. We project our dataset into 10 dimensions and form our feature vector $FS2$ by taking the eigenvectors of 10 components. Figure 3.12 shows the distribution of variances for all 10 components selected after dimensionality reduction of the data.

3.5.2.3 Attribute Selection Correlation (FS3)

In the third iteration, we use Correlation Attribute Evaluation technique to identify a set of discriminatory attributes. Unlike PCA, Attribute Selection Correlation is a feature selection technique (and not a dimension reduction technique) that identifies the subset of primary attributes which has the greatest impact towards the targeted classification. Attributes Selection Technique computes the correlation between all attributes and filters a subset of features consisting of irrelevant or less coherent information [129]. Therefore, the output or final set of attributes represent relevant and enriched information unlike the initial set of attributes. To select the best subset of attributes, we measure the Pearson's correlation between each attribute (feature vector) and the class. We create a correlation matrix of 45 attributes and class for each record in the dataset and compute the overall correlation by computing the weighted average of the attribute. Based on the correlation between each attribute vector and class, we create a set of 10 features having moderately higher positive and negative correlation and drop the features having correlation closer to zero. For our experimental dataset, FS3 returns the following 10 features: mention of past and present tense, pronouns, male references, perceptual process, negative emotions, clout, the presence of negation, swear words and anger.

3.6 Proposed Solution Approach

3.6.1 Classes and Membership Groups

Based on the polarity of opinions in religious posts and their importance in defining the dynamic of conflicts, we split the dimensions of conflicts into six classes: information sharing, query, N/A, sarcasm, defensive and disagreement. We further divide the 'disagreement' class into six subclasses reflecting a higher range of negative emotions: disappointment, annoyance, insult, disgust, ashamed and disbelief. For a given data point y_m, to identify the polarity of the post, we first classify the post into six classes and assign a label o_m. If the post is identified as a disagreement or negative post, we further classify into six subclasses identifying the low-level details of negative emotions in a given post.

3.6.2 Classification Approach

Due to the constraint of lack of ground truth and only a very small chunk of available labeled data (2%), we use semi-supervised classification method to classify the unlabelled posts over unsupervised method. Semi-supervised classification approach uses both annotated and unlabeled data to learn the model iteratively in a snowball manner. We use 684 posts annotated by Tumblr bloggers and use them to train our model in the first iteration of the semi-supervised classifier. We conduct our experiments on $35,799$ posts identified as the topic related (discussing any religious group or community). Given a labeled data (X_N, C_N) where the data points are denoted by $X_N = (x_1, x_2, x_3...x_n)$ and their labels are denoted by $C_N = (c_1, c_2, c_3...c_n)$. The unlabelled data points $Y_M = (y_1, y_2, y_3...y_m)$ and their unknown labels $O_M = (o_1, o_2, o_3...o_m)$ are denoted as (Y_M, O_M).

We use the *'R'* statistical language to perform classification using *"upclass"* package [130]. "Upclass"[11] is a semi-supervised classification method and an adaptive version of the model-based classification method proposed in Dean et. al. [131]. Upclass uses an iterative method that initiates by using model-based classification method and uses Expectation- Maximization (EM) algorithm in further iteration until convergences. In the first iteration of classification, a set of 14 models is applied on the dataset considering different constraints (E- equal, V- variable, I- identity) upon covariance structure- volume, shape, and orientation of the cluster. For example, in EEE model each cluster has equal volume, same shape and same orientation along the axis. The clustering is performed in multiple iterations by estimating group membership on unlabeled data based on the maximum likelihood of EM algorithm. In order to perform the model based discriminatory analysis on unlabeled data points, the model of data (combination of E, I, V constraints) must be known. If the model is null, then Upclass fits every model to the data and identifies the best-fitted model of given data and attributes. To identify the best-fitted model, Upclass calculates the Bayesian information criterion (BIC) value for each model. $BIC = 2\log(l) - p\log(n)$; where l is the likelihood of the data, p is the number of model parameters and n is the number of data points. The model with the highest BIC value is selected as the best-fitted model for the data.

3.7 Empirical Analysis and Evaluation Results

In this Section, we present the classification results of Upclass semi-supervised method applied for both classes and sub-classes identification. We apply 3 iterations of Upclass supervised classification methods on all $35,799$ posts for each feature vectors model (FS1, FS2 and FS3) discussed in Section 3.5.2. If a post is labeled as "Disagreement or Negative", we further train our model on the posts labeled under the six

[11] `http://CRAN.R-project.org/package=upclass`

Table 3.4: Classification Results for Feature Selection Techniques of Different Membership Groups and Observations

	FS1		FS2		FS3	
Attribute	Class	Sub-Class	Class	Sub-Class	Class	Sub-Class
Converged	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Iteration	272	491	604	252	207	110
Dimension	45	45	10	10	10	10
Model Name	VEV	EEV	VVV	VVV	VEV	VVI

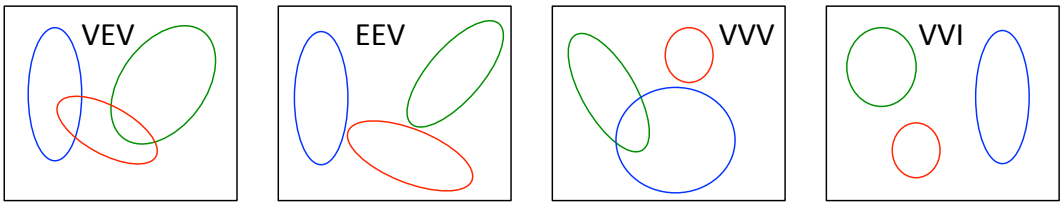
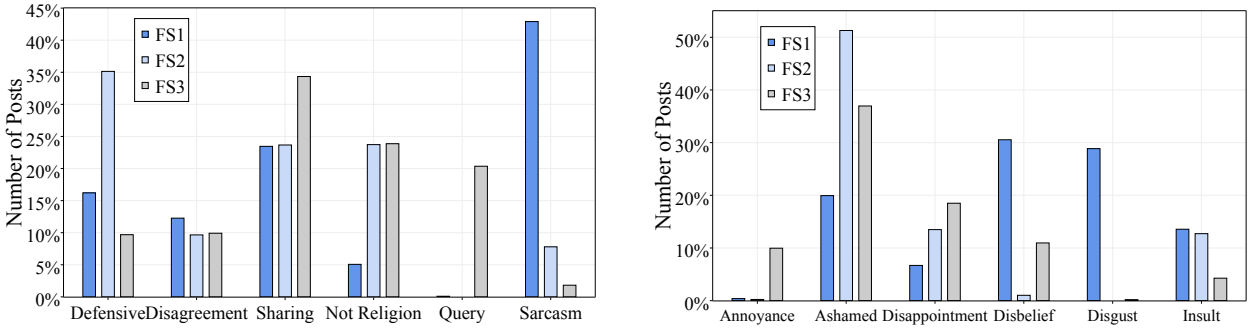


Figure 3.13: Visualization of Volume, Shape and Orientation Constraints for Best-Fitted Models for Classification. V= Variation, E= Equal and I=Identical

subclasses of disagreement and classify unknown data points into one of the six sub-groups using Upclass semi-supervised classification method. Table 3.4 shows the experimental results of classification performed using each feature vector model for each membership groups. Table 3.4 reveals that the classification model converges for each set of feature vectors. During the first step of classification both FS1 and FS3 run the similar number of iterations whereas, FS2 execute approximately 2.5 times of their iterations. Further, for FS1 and FS3, VEV is selected as the best-fitted model while for FS2 attributes, VVV showing the non-linear distribution of labels (different orientation of each cluster against the axis). Figure 3.13 shows the visual representation of clusters created using different models (considering the constraints on covariance structure). Table 3.4 also reveals that during the second set of classification (sub-groups of disagreement class), Upclass method takes a different number of iteration for each feature vector model. Further, for each feature vector, a different discriminant model is selected. Table 3.4 shows that in FS2 Upclass uses the same model i.e. VVV for identifying classes as well as sub-classes. Whereas, using all attributes at once, it classifies all observations into equal parts and creates clusters of identical shapes varying in the orientation against the axis. While using the features selected using Pearson's correlation technique, it classifies the observations in a linear manner- varying the shape and volume of the clusters when all data points aligned towards an axis. Unlike, FS1 and FS3, while using principal components as feature vectors, the clusters are created in a non-linear manner- varying in size, shape, and orientation of data points.

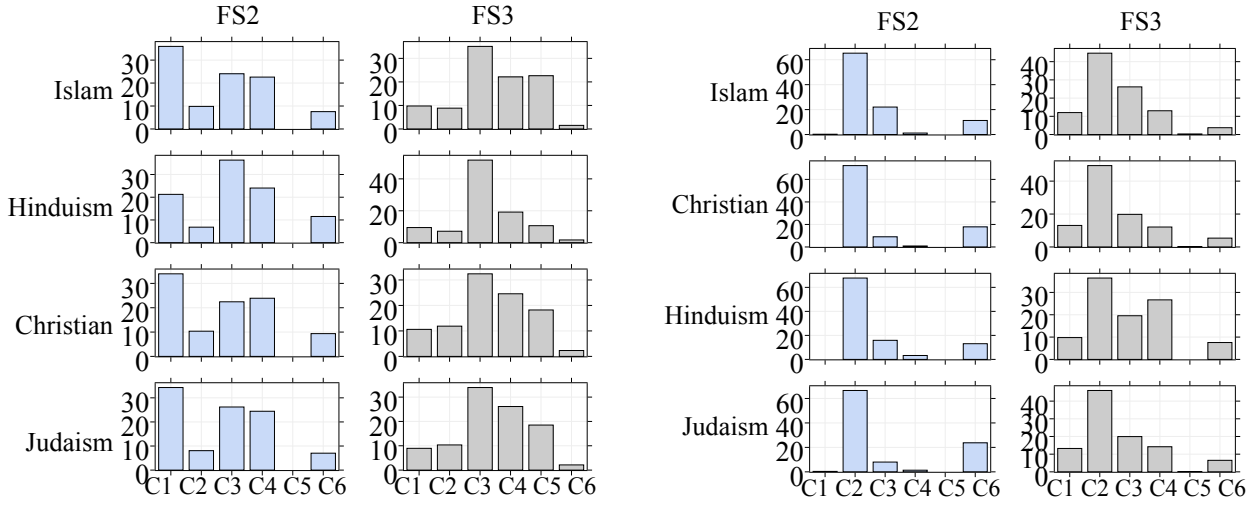
Figure 3.14 shows the distribution of Tumblr posts classified into different groups of classes and sub-classes based on the polarity of opinions. Figure 3.14a shows the relative percentage of posts classified into each of the defined labels. Figure 3.14a reveals that while using all attributes as features, maximum number of posts (43%) are labeled as sarcasm posts which is below 10% while using dimensionality reduction techniques. While only a very small percentage of posts (5%) are classified as non-religion based posts which is significantly higher for both FS2 and FS3 (approximately 25%). The classification results shows that except FS3, using FS1 and FS2 feature vectors, the classifier does not have sufficient examples for labelling query posts. The graph in Figure 3.14a shows that for each feature selection method, the classifier classifies 10% to 12% posts as disagreement/negative that are further classified into sub-classes. Figure 3.14b shows the relative percentage of these 10% to 12% posts further classified into sub-classes of extreme negative emotions.



(a) Distribution of Polarity Based Classes Identified for All Observation Posts in Experiemental Dataset

(b) Distribution of Sub-Classes Identified for All Posts Labelled as "Disagreement/Negative"

Figure 3.14: Classification Results of Upclass Semi-Supervised Method for Unknown Posts Categorized into Polarity Based Classes and Extreme Emotions Based Sub-Classes



(a) Distribution of Polarity Based Classification of Tumblr Posts

(b) Distribution of Sub-Classification Results of Tumblr Posts

Figure 3.15: Distribution of Classification Results of Tumblr Posts Specific to A Religion.

Figure 3.14b reveals that while taking all attributes into account, a very small percentage ($\sim$ negligible) of posts are classified as "Annoyance" posts while the distribution of other classes are significantly higher. While the distribution of posts for FS2 and FS3 is varying for each category- as reflected in best-fitted model selected for classification (refer to Table 3.4). The variation in distribution of all posts in different categories shows the dynamics of public opinions on religious posts. The size of each cluster (number of posts grouped in a class) for different combinations of attribute selection techniques and classification method shows the presence of religious conflicts among users on Tumblr.

To address the challenge of determining beliefs in an ambiguous post (discussing more than one religion), we identify the name of religions being discussed in each post. We classify each post into classes (polarity based groups) and sub-classes (extreme negative emotions based groups) and further examine the results of classification for identifying religion specific conflicts. Due to the large volume size of "Sarcasm" cluster and no post classified as "Query" post, we discard the FS1 technique for identifying the conflicts among individual religious groups. Figure 3.15 shows the classification results and distribution of Tumblr posts classified into various dimensions of conflicts. For Figure 3.15a, $C1$, $C2$, $C3$, $C4$, $C5$ and $C6$ denote defensive, disagreement, sharing, not religion, query and sarcasm dimensions respectively. Similarly, for Figure 3.15b, $C1$, $C2$, $C3$, $C4$, $C5$ and $C6$ denote annoyance, ashamed, disappointment, disbelief, disgust and insult.

As shown in Table 3.4, while using principal component analysis feature vectors, the semi-supervised classification method selects VVV as the best-fitted model. Figure 3.15a also reveals that for FS2 feature vectors the volume of all clusters is different making some clusters too large or too small. Further, Figure 3.15b reveals that maximum number (more than 60%) of disagreement posts belong to "ashamed" category. Whereas, while using Pearson's correlation selection method, the posts are grouped into all classes. During the first phase of classification, there is a variation in the volume of observations in each cluster while the shape of each cluster is the same. Whereas, in the second phase of classification, the volume of observation and shape of each cluster varies. The classification of Tumblr posts for each religion into several dimensions of conflict shows that a lot of discussion about religious topics happen on social media where users have different opinions, beliefs, and sentiments about these religions. Our results show that various linguistic features such as emotions, social presence, summary of language variables and other linguistic dimensions of user-generated data can be used to identify the conflicts within religious faith and beliefs. Furthermore, Tumblr is a rich source of collecting public opinion posted in a detailed and open manner which is useful to study the low-level details of religious beliefs and overcome the challenges of offline data and surveys.

3.8 Conclusions and Future Work

With the unexpected emergence and rapidly growing influence of religious faith and beliefs in political and social activities leading to the discrimination and violence against other rivalry communities. Therefore, the identification of collision in various religion and race communities has come out to be one of the major problems for the government, local forces, and law enforcement agencies. We create a hypothesis that due to the popularity of websites and subjectivity in content, mining social media posts can fill the gaps of traditional and offline surveys. We conduct our experiments on an open source dataset consisting of the largest collection of Tumblr posts associated with religion based tags. We conduct a survey among three different groups of people (graduate students, Tumblr bloggers, and individuals from society) and define 11 dimensions of public opinions that can identify the contrast of conflicts. We perform topic modeling on our data and classify the posts that belong to certain religions. We further investigate the feasibility and efficiency of linguistic features and different dimensionality reduction techniques and compare their results of classifying Tumblr posts into different dimensions of conflicts. Due to the small size of labeled data, we use Upclass- a semi-supervised classification method to train our model and classify unlabeled observations. Based on our results, we conclude that despite the presence of noise and ambiguity in content, computational linguistic features are efficient and base methods to identifying the dynamics of religious conflicts. Furthermore, identifying the topic prior to the linguistic features extraction can be used to disambiguate the sentiments of author while discussing more than one religion in a single post.

Future work includes the improvement in linguistic features and making them generalized and efficient for classifying very short and short text posts. Furthermore, future work includes the identification of age and location of bloggers for identifying the collision of religious beliefs and sentiments in different age groups or various regions across the world.

Chapter 4

Detecting Extremist Content, Users and Hidden Communities on Social Media

4.1 Introduction

As discussed in Chapter 3, various people on the Internet takes the leverage of freedom of speech and features of social networking websites to express their opinions and beliefs about a variety of sensitive topics such as religion, race, and anti-national comments. Many like-minded people and a group of individuals use these features to outburst their opinions in the form of hateful and insulting messages against an individual or community. In addition to such people, several organizations use the Internet and the web 2.0 as a medium for not expressing their opinions but promoting their ideology referred as radicalized and extremist groups. Recent research demonstrates that the Internet and social media platforms are increasingly used by such extremists groups for online radicalization [132] [133]. Online radicalization consists of using the web as a medium by extremists and terrorists for conducting malicious activities such as promoting extreme social, religious and political beliefs, recruitment of youth, propagating hatred, promoting their ideologies, and forming communities sharing a common agenda [50] [59] [134]. Social media platforms (such as YouTube, Twitter, and Tumblr) are exploited for conducting extremist activities due to the low publication barrier, wide reachability to a large number of people across countries, and anonymity. Further, information shared on social media has access to more individuals to carry out "lone wolf" operations against Western targets. Creating channels and video clips on how to make a bomb or a speech on racist propaganda and posting it online on a popular video sharing website like YouTube fall under online radicalization [48] [135] [136]. Using a widely used micro-blogging platform such as Twitter for creating virtual communities and recruiting young people in their radical groups are examples of using Internet and online social media for radicalization [137]. Similarly, posting hate-promoting blogs and comments on online discussion forums to disseminate extreme religious beliefs is a major online radicalization concern for government and law enforcement agencies [138]. Figure 4.1 shows the concrete examples of various hate promoting posts and their associated tags created on Tumblr website. Figure 4.1 also illustrates the example of an anti-India video posted on YouTube website while the username of uploader also shows extremism behavior of the uploader. Figure 4.1 further shows examples of two Twitter handles posting extremist content very frequently and having a very large number of followers[1]. According to SwarmCast journal article and Shumukh-al-Islam posts, these accounts are stated

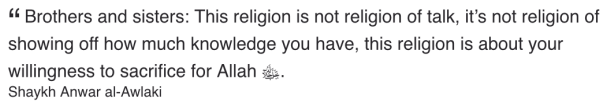
[1] `http://www.terrorismanalysts.com/pt/index.php/pot/article/view/426/html`

Figure 4.1: Concrete Examples of Various Extremist and Online Radicalized Posts Created on Tumblr and YouTube websites. Further showing the Example Snapshots of Twitter Accounts Labeled as Jihadi and Extremist Groups.

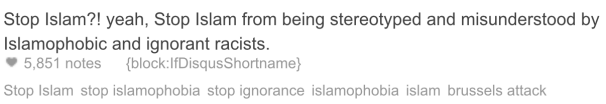
as the two most important and active Twitter accounts of Jihadi and Mujahideen groups[2]. According to an article published in New York Post[3], in 2012, Twitter's digital hate was grown at an alarming speed. The website recorded an incredible 30% surge in such kind of traffic while there were about 20,000 hate-filled hashtags and handles (anti-Semites, racists, and terrorist groups micro-blogs).

Online radicalization has a major impact on society that contributes to the crime against humanity and mainstream morality. The presence of such content in large amount on social media is a concern for website moderators (to uphold the reputation of the website), government and law enforcement agencies (locating such users and communities to stop hate promotion and maintaining peace in the country). However, due to the dynamic nature of social media platforms, the manual identification of such content, users and communities are overwhelmingly impractical. Further, due to the free-form nature of social media and user-generated

[2]http://jihadintel.meforum.org/identifier/149/shumukh-al-islam-forum
[3]http://nypost.com/2013/10/26/hate-speech-running-rampant-on-twitter/



(a) Topic- Religion, Intent-Yes



(b) Topic- Religion, Intent- No

Figure 4.2: Concrete Examples of Tumblr Posts Showing Different Intent of Bloggers Creating Religion Based Posts

Ali Shukri Amin, (17 yrs. Old) Twitter @AmeerkiWitness

Jake Bilardi, Radicalized Jihadist (18 yrs. old, 10th std. student)

3 men from Maharashtra joined ISIS in Iraq (age group 20-25)

Figure 4.3: Real-World Examples of Youth Getting Influenced by Extremist Content on Online Social Media

data, automatic identification of such content and communities is technically challenging. In addition to NLP and text analytics based challenges, another problem that we encounter is ambiguity in the posts. Social media posts are user-generated data, and hence the classification of a published content as a radicalized post majorly depends on the intentions of the narrative. Figure 4.2a and 4.2b shows examples of two Tumblr posts where bloggers mention about the Islam religion. In Figure 4.2a, the intention of author is to provoke his Muslim followers for Jihad and develop a willingness to sacrifice themselves for their religion, whereas in Figure 4.2b, the intention of author is to bring awareness that Islamophobic and other hate groups should stop misunderstanding Islam religion.

Further, it is technically challenging to identify the intent of a post when a naive post has similar terms as a radicalized or racist post. For example, a post P1: *"All types of Jihad is to establish peace for all & Sharia also promote peace, so there is no need to fix anything @simafaysal @profdstone"*- posted by an author with screen name 'Prisoner' and an another post P2: *"This settles it? 'Jihad is to establish peace' 'there is no need to fix anything' spoken like a true 'prisoner'"* have similar content. Here, the intention of P1 is to show his support for Jihad and terrorism while the intention of P2 is to make a sarcastic comment on P1 and author's belief. Further, despite having hateful remarks in a post, the intention of the author can still be naive. For example, in January 2016, Saudi Arabia released an official video on 'how to properly beat Muslim women' with the intention of targeting women communities. Recently, as the video got published worldwide, users at microblogging websites shared that video and posted hateful comments to oppose the video with no racist intent. Whereas, some users posted comments opposing the video and targeting whole Muslim community with racist intentions bringing ambiguity in their posts. Usage of gif images for expressions their reactions and opinions, facility of sharing content from external sources such as news websites or blogs makes it difficult to identify the intentions of the uploaders. The work presented in this chapter is motivated by the following facts:

1. Despite the non-radicalized or non-hate promoting intent of the author, but due to the severe impact of presence of such posts on social media, it is important for the government to constantly monitor the social networking websites, keeping track of this information in real time and identify the malicious

content which might lead to the religious violence. Figure 4.3 shows the concrete and real-world examples of young age kids and youth getting influenced by extremist and hate promoting content on social media. While, Jake[4], an 18 years old kid became a Jihadist and joined radicalized groups; Ali Amin[5], a 17 years old boy was arrested and sentenced to prison for raising funds for ISIS group via a Twitter account.

2. In addition to the presence of extreme, outrageous and grievances comments posted by naive users, there are many users who are either affected by an incident, brainwashed or manipulated by other extremist people and are intend to post malicious content on the websites frequently. Due to the presence of several such users constantly posting extremist content on social media, the identification and removal of only the content are not sufficient. One of the key motivations of the study presented in this chapter is to identify the (lone-wolf) users publishing hate promoting content, participating in comments, arguing, discussing sensitive topics and making provoking comments.
3. In addition to such lone-wolf users, there are several groups of people and individuals posting extremist content in an organized manner with a common agenda and propaganda. The work presented in this study is also motivated by the need to develop the solutions to identify the relationship between many such users and locate their hidden virtual communities spreading hate and recruiting people in their groups. Further, the motivation of this study is to find the closely-related users and playing a major role in the virtual community to run the organization on social media platforms.

4.2 Related Work

We conduct a literature survey in the area of mining social media platform for identifying hate promoting and extremist content, users and communities. We discuss only closely related studies to the work presented in this chapter. Based on the application and usage scenarios presented in section 4.1, we divide our literature survey into following two lines of research:

4.2.1 Hate Promoting Content, Users and Communities Detection on Social Media

We conduct a literature survey on the topic of hate and extremist content and community detection on Web 2.0. We characterize the existing literature across the data sources and the objective of the study presented in the chapter. The prior literature shows the studies conducted across a diverse range of content such as terrorism, extremist groups, anti-black communities, US domestic, middle eastern, jihad and anti-Islam. Ting et. al. [139] present a keyword-based flagging approach to identify hate groups on Facebook website. They further use social network analysis to identify the relationship among defined groups. Goodwin [140] perform a quantitative analysis for analyzing various counter-Jihad, Islam and Muslim communities on Web 2.0. He further presents an in-depth analysis of their activities, supporters, and reasons behind the emergence of these groups on social media. Sureka et. al. [141] propose to use video comments as a feature to identify the hate promoting videos on YouTube. They also use social network analysis to discover users and their hidden communities on YouTube website. Chen et. al [142] [143] [144] [145] present a text analytics based framework for identifying presence of extremist content on YouTube and discussion forums. They further analyze several Jihadi groups on Web 2.0, SecondLife- an online gaming website. Reid et. al

[4] http://www.news.com.au/national/freshfaced-westerners-are-being-lulled-into-terrorism-by-isis-propaganda/news-story/8448148e3a0c33c01b95db4dc1bab492

[5] http://www.ibtimes.com/who-ali-shukri-amin-virginia-isis-teenager-behind-pro-islamic-state-twitter-sentenced-2073208

[146] present a hyperlink study on various blogs, discussion forums, and video sharing websites to identify online communities of extremist groups. Salem et. al [147] propose a multimedia and content-based analysis approach to detect Jihadi extremist videos and the characteristics to identify the message given in the video. Mahmood S. [148] presents a keyword-based Google Search method to identify terrorist groups on social network websites. He analyzes the sentiments and opinions of users following several terrorist groups on OSM websites and proposes a counter-terrorism mechanism to identify those users who are highly likely to commit a violent act of terror. He also discusses honeypots and counter-propaganda techniques that can be used to rehabilitate radicalized users back to normal users. The disadvantage of keyword based flagging approach is that it generates a large number of false alarms. David et al. [149] present a keyword based search to detect several criminal organizations and gangs on Twitter & Facebook. They analyze the presence of organized crime and how these gangs use social media platforms to recruit new members, broadcast their messages and coordinate their illegal activities on the web 2.0. They perform a qualitative analysis on 28 groups and compare their organized crime between 2010 and 2011 on Facebook. O'Callaghan et. al. [133] describe an approach to identify extreme right communities on multiple social networking websites. They use Twitter as a possible gateway to locate these communities in a wider network and track the dynamic communities. They perform a case study using two different datasets to investigate English and German language communities. They implement a heterogeneous network within a homogeneous network and use four different social networking platforms (Twitter accounts, Facebook profiles, YouTube channels and all other websites) as extreme right entities or peers and edges are the possible interactions among these accounts.

4.2.2 Intent-Based Classification of Radicalized Posts on Social Media

In this Section, we discuss the literature survey conducted in the area of intent mining on social media platforms. Prior research shows that while there has been a lot of work in the field of mining social media data for detecting intent of buyers and consumers, the intent-based identification of racist or radicalized post is relatively an unexplored area. Further, the existing studies propose to use the keyword based techniques for identifying the intent of narrative which impedes the accuracy of classification due to misleading, sarcasm and ambiguous posts. We discuss the closely related work to intent identification on social media in following subsections:

4.2.2.1 Commercial Intent Classification

Wang et al. [150] present a graph-based semi-supervised learning technique to classify intent tweets. They combine keyword based flagging (referred as an intent keyword) and graph regularization method for classifying tweets into six categories. Purohit et al. [151] present a hybrid approach of combining knowledge-guided patterns and bag-of-tokens model for intent classification of short text. They conduct a study on Twitter for crisis events dataset and address the problem of ambiguity and sparsity to classify the intent of narrative. Ding et al. [152] present a transfer learning based convolutional neural network model for identifying users' buying or consumption intentions from Sina Weibo- a Chinese microblogging service. Geetha et al. [153] present a lexicon (sentiment Wordnet dictionary) based bootstrapping method to measure the polarity of opinion in a short text data. They conduct a study on Twitter data and compare their results for movie reviews, election results, and product reviews. Wang et al. [138] present a graph-based ranking model to identify the commercial intent from trending topics on microblogging platforms.

4.2.2.2 Racism/Radicalization Intent Classification

Smith et al. [154] conduct a quantitative content analysis on public documents to distinguish radical groups from non-radical groups. Prentice et al. [49] conduct a quantitative text analysis on 50 documents

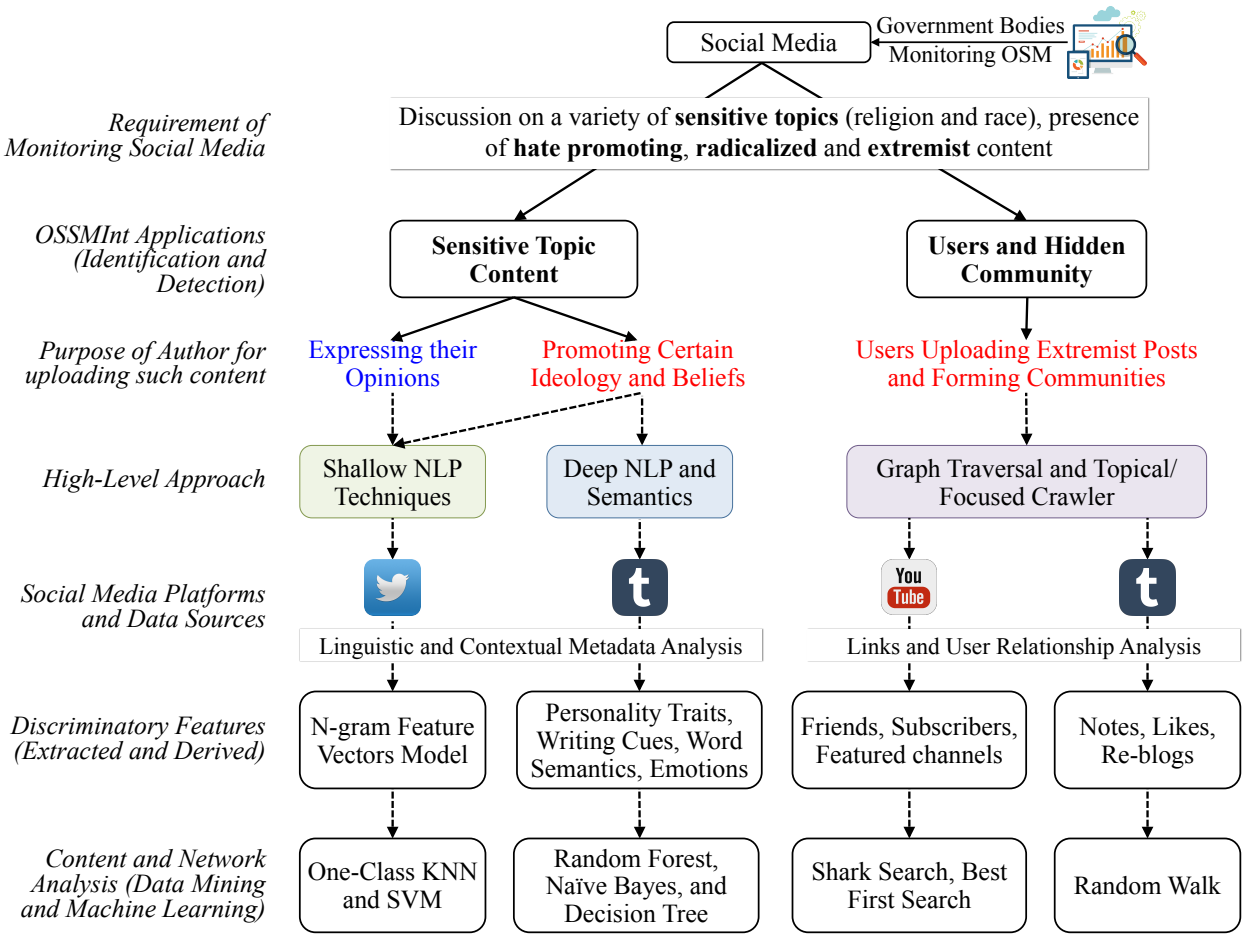


Figure 4.4: A High-Level Block Diagram Demonstrating the Case-Studies Conducted on Various Platforms for Identifying Extremist Content, Users, and Communities. Further Illustrating the Use of Proposed Features (linguistic, contextual and user relations) for Content and Network Analysis on Twitter, YouTube and Tumblr Websites.

originated from extremist websites. They present a 'Conduct and Composition Analysis' technique to classify the persuasion behavior of online extremist media varying for the documents posted before and after the Israeli activities in Gaza. Our literature survey reveals that there has been a lot of work in the area of commercial intention identification from free-form text whereas automatic detection of racist posts on social media platforms such as Tumblr is a relatively unexplored area.

4.3 Research Contributions

Figure 4.4 shows the high-level block diagram demonstrating the problem of online radicalization detection on social media. Motivated by the prior literature and the applications discussed in Section 4.1, we divide our problem of online radicalization detection on social media into three sub-problems (also divided into 4 case-studies):

1. Identification of all types of extremist content leading to the violence and manipulation of audience
2. Identification of the sensitive content posted with an agenda to promote an ideology and promote hate among the audiences
3. Locating the users uploading such content on social networking website and uncovering their hidden communities promoting such content in an organized manner.

Figure 4.4 reveals that due to the diversity in nature of the information that needs to be extracted, we propose three different types of approach. In order to identify all extremist and hateful posts, we use probabilistic and shallow NLP based techniques. Whereas, for the identification of the radicalized posts published with intent to promote hatred among social media users, we propose to use deep semantics and NLP based techniques. Further, we use links and users relations (specific for the domain to be analyzed) for identifying the extreme right communities on social network websites. In contrast to the existing literature, the specific and key contributions of the work presented in this chapter are the followings:

1. As demonstrated in Figure 4.4, we propose to use various n-gram based features such as the presence of religious words, war-related terms and several hashtags that are commonly used in extremist tweets. To the best of our knowledge, in contrast to the previous literature, we present the first study on identification hate and extremism promoting tweet using a one-class classifier framework. We conduct our experiments on a real-world Twitter dataset and investigate the efficacy of proposed approach across one-class SVM and KNN classifier and test its effectiveness for the given classification task.
2. In contrast to the prior studies, we present an application of focused or topical crawler-based approach for locating hate and extremism promoting channels on social media. While there has been a lot of work in the area of topical crawling of web pages, our work presents the first study on the application of focused crawler framework for navigating nodes and links on online social networking platforms.
3. We present our study on YouTube- the most popular video sharing website and Tumblr- the second most popular micro-blogging service. We conduct a series of experiments on real-world data downloaded from these websites to demonstrate the effectiveness of the proposed solution approach. We demonstrate the effectiveness of shark search, best first search and random walk graph traversal algorithm for navigating links between user channels We further use social network analysis based approach on the retrieved user profiles and their connections obtained from the focused crawler traversal to understand the presence of communities and central users.
4. We perform a characterization study of hate promoting (anti-religion, nationalism, terrorism, anti-India, politics) content posted on Twitter and YouTube. We analyze the multimedia content and available textual metadata for content characterization such as objective of posting the content and targeted audiences.
5. In contrast to the existing keyword spotting techniques for extremism detection, we present the first and novel study conducted on racist and radicalization detection based on the intent of narrative, unlike previous keyword spotting methods. We propose to use various natural language processing based features and investigate the efficiency of personality big five model, authors' emotions (unlike polarity based sentiments) and semantics for identifying racist intent based posts. We conduct our study on Tumblr- second most popular micro-blogging website and also publish the first ever semantically and sentimental enriched data of Tumblr posts and make our data publicly available for benchmarking and extension [63]

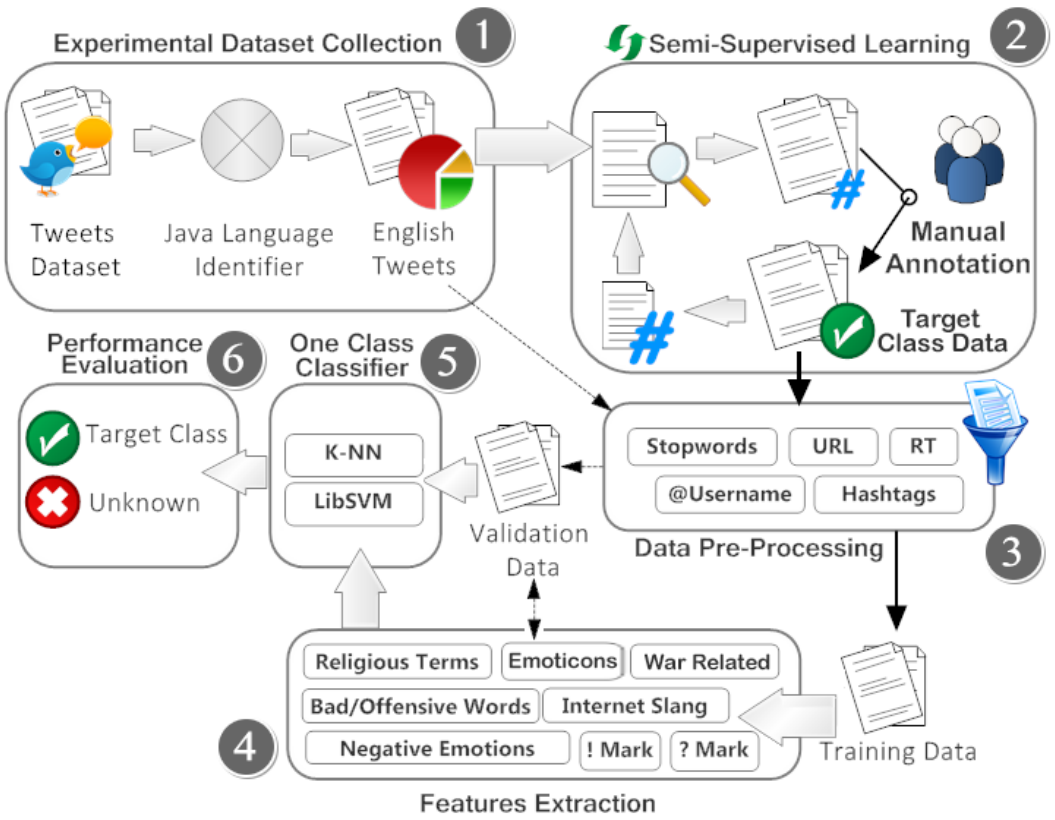


Figure 4.5: A General Research Framework for the Proposed Solution Approach Primarily Consisting of 6 Phases: Experimental dataset collection, semi-supervised learning, data pre-processing, features extraction, one-class classification, performance evaluation

4.4 Case Study 1: Identifying Hate and Extremism Promoting Content on Twitter

As discussed in Section 4.1, the automatic identification of hate and extremism promoting tweets is useful to intelligence and security informatics agents as well as Twitter moderators. Manual identification of such tweets and filtering information from raw data is practically impossible due to the large volumes of tweets (500 million) posted every day. Tweets consist of short text (maximum of 140 characters) and noise (incorrect grammar, spelling mistakes, slang, and abbreviations) as a result of which automatic classification of tweets is a technically challenging problem. The work presented in this case-study is motivated by the need to investigate solutions to address the problems encountered by security analysts for countering online radicalization on the largest micro-blogging platform on Internet. The research aim of the work presented in this case-study is the following:

1. To investigate techniques and propose linguistic & stylistic features and characteristics of hate and extremism promoting tweets for their automatic identification.

2. To conduct empirical analysis on a large real-world dataset and demonstrate the effectiveness of the

proposed approach.

3. To examine the relative influence of each proposed feature for the task of identifying hate and extremism promoting tweets. Furthermore, our objective is to compare and contrast the performance of various Machine Learning algorithms (KNN and LibSVM) for the purpose of recognizing extremism promoting tweets.

Algorithm 2: Training Dataset Collection

Data: Seed hashtags HT, Experimental dataset ED, Size of Training dataset S
Result: Training Dataset TD
Algorithm $TrainingDataset(HT, ED)$
  while $T_{h\&e}.size > S$ **do**
1 set of tweets $T \leftarrow ED.findTweets(HT)$
2 manual labeling of tweets as hate promoting or unknown.
3 $T_{h\&e} \leftarrow$positive labeled tweets
4 $TD \leftarrow T_{h\&e}$
5 Extended Hashtags $EHT \leftarrow ExtractHashtags(T_{h\&e})$
6 $TrainingDataset(EHT, ED)$
  return TD

Algorithm 3: Data Pre-processing

Data: Experimental Dataset ED, A Lexicon of English stopwords L_{st}
Result: Preprocessed Tweets T_p
Algorithm $FilterTweets(ED)$
1 **for** ***all*** *tweets* $t \in ED$ **do**
2 **for** ***all*** *stop words* $s \in L_{st}$ **do**
3 $t \leftarrow t.replaceAll(s, "")$
4 $t \leftarrow t.replaceFirst("RT", "")$
5 $t \leftarrow t.replaceAll("@username", "")$
6 $t \leftarrow t.replaceAll("URL", "")$
7 $t \leftarrow t.replaceAll("hashtags", "")$
8 $T_p.add(t)$

4.4.1 Research Framework

Figure 4.5 illustrates the proposed solution approach. The proposed method is a multi-step process primarily consists of six phases: experimental dataset collection, training dataset creation, data pre-processing, feature extraction, one- class classification and performance evaluation. The six phases are labeled in the solution framework. In phase 1, we download two publicly available datasets [155] [156] (refer to Section 4.4.3.1 on experimental dataset) and combine them to form a single experimental dataset. The dataset consists of tweets belonging to multiple languages. We use Java language detection library[6] to filter English and non-English tweets. We conduct experiments only on the English language tweets and discard non-English tweets. We notice that 85% of tweets are in English.

We require a training dataset to create a statistical model for one-class classification for the task of

Table 4.1: A Sample of Hate Promoting Tweets Leading to More Hashtags.

Seed #Tag	Tweet	Extended #tags
#Terrorism	Secret #recruitment British students #Muslim #extremists ? #islamophobia **#terrorism**	#islamophobia, #extremists
#Islamophobia	#NoJihad #Racism lowest form stupidity ! **#Islamophobia** height common sense ! #Quran	#NoJihad, #Racism
#Extremist	Engaging #AfPak Information War: Countering **#extremist** #propaganda with #mobile #technology	#propaganda
#Islam	**#Islam** evil according #GeertWilders one few islamophobic people Netherlands yet everywhere	#GeertWilders
#Terrorism	New: Al Qaeda Bomb Maker Video **#terrorism** #bomb #video #alqaeda #alquida	#bomb, #alqaeda, #alquida

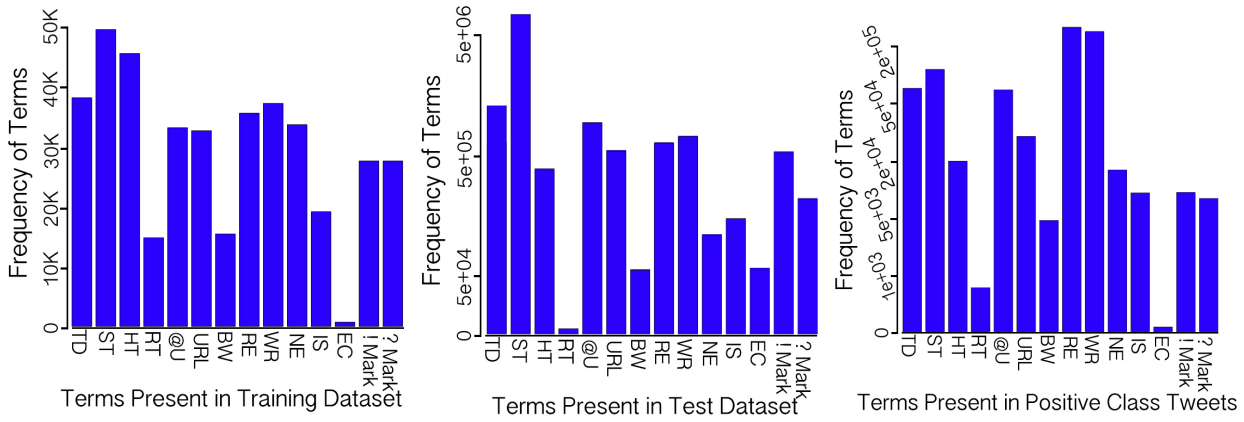
Table 4.2: A Sample of Keywords Present in Hate Promoting Tweets.

Hashtags	#islamophobia, #stealthjihad, #myjihad, #extremists, #NoJihad, #terrorism, #dreamact, #terrorist, #nativist, #GeertWilders, #alqaeda, #assassination
Religious	hijab, hizb, demon, jihad, god, maulana, kabba, azan, burka, prophet, koum, apostate, sikh, muhajir, immigrant, hijr, amen, hinduism, devil, atheist
War Related	LOC, Bomb, Blast, Attack, Holy war, Warfare, Tribute, Soldier, Jawan, Refugee, Enemies, Fighting, Patriot, Assassination, Expose, Army, Zindabad
-ve Emotions	endangered, enslaved, entangled, evaded, evasive, evicted, excessive, excluded, exhausted, exploited, exposed, fail, fake, hatred, regret, disgust, flaw, FALSE, fear, fed up, flaw, forced, forgetful, forgettable, forgotten, fragile, freak, frighten, frigid, frustrate, disgust, dishearten, disillusion
Emoticons	:), :-), :D, :-D, =], :], ;), =P, :P, :-P, :*, :(, :-(, =(, :-S, :S, :O, :-O, :\, :-\, \-o, :-}X, :-(, =), :-E, :-F, :-C, 3:*¿, :-(, :(, :-d, :-¿, :-@,)8-), 3:), O:), :'(
Internet Slangs	LOL, haha, ROFL, WT*, WTH, IMHO, OSM, AKA, BRB, 404, CC, TC, TT, Cya, Gr8, FAQ, FYI, Hw, L8r, N/A, W/O, B/W, BTW, NP, OMG, PLZ, OMG, PS, PLZ, MSG, RSVP, TTYL, TXT, U, U2, B4, W8, TY, THX, 403, W*F, GOSH, em, 4eva, ABT, DEPT, FF, ILY, TFS, WHOA
Bad Words	ahole, a**, ba****d, bit*h, crap, f**k, gay, damned, hells, jackoff, sh**, pe***, sexy, sl*t, XXX, b17ch, s.o.b., wh**e, screw, bulls**t, d-bag, jerk-off

identifying hate and extremism promoting tweets. We use a semi-supervised learning approach to creating our training dataset (refer to Algorithm 2). Hashtags are the strong indicators of the topic of the tweet. We create a list of seed hashtags such as #Terrorism, #Islamophobia, and #Extremist and identify tweets containing these hashtags. We manually analyze tweets containing such hashtags and identify hate and extremism promoting tweets. We extend the list of hashtags by extracting new hashtags (not already in the list) present in the positive class tweets. Table 4.1 illustrates a sample of some seed hashtags and their respective tweets leading us to new hashtags. We then identify tweets containing the new hashtags and manually analyze the tweets to identify hate promoting tweets. As a result of this, we extend the list of hashtags and our training dataset of size S. We repeat this process several times to collect training dataset. We make our experimental dataset publicly available so that our experiments can be replicated and used for benchmark purposes by other researchers[7]. We perform a random sampling on English tweets and use

[6] `https://code.google.com/p/language-detection/`

[7] `https://goo.gl/QQcDNY`



(a) Training Dataset (b) Testing Dataset (c) Positive Class Tweets in Testing Dataset

Figure 4.6: Frequency Distribution of Various Terms Present in the Training and Testing Dataset. TD= Tweet Dataset, ST=Stopwords, HT= Hashtags, RT= Retweets, @U= @username Mentioned, URL= Hyperlinks, BW= Bad Words, RE= Religious Terms, WR= War Related Terms, NE= Negative Emotions, IS= Internet Slangs, EC=Emoticons.

a sample as our testing (or validation) dataset. Algorithm 3 describes the data pre-processing done on the training and testing datasets. As shown in Algorithm 3, we remove the term 'RT' (Re-Tweet), @username (username of the direct mention of a user in the tweet), URL (short URL) and hashtags. After removing these terms, our problem becomes more challenging due to short text classification. In phase 4, we perform characterization and identification of various discriminatory features and compute the frequency (TF) of various terms. For example, religious, offensive, slang, negative emotions, punctuations and war-related terms. Table 4.2 shows a sample of these terms present in hate and extremism promoting tweets. Figure 4.6a and 4.6b shows the frequency of these terms present in the training and testing dataset. While Figure 4.6c shows statistics of only positive class tweets present in testing dataset. Figure 4.6 also illustrates the frequency of terms that have been preprocessed in phase 3. All statistics are computed in logarithmic scale. These graphs show that the frequency of religious and war-related terms is very high in hate promoting tweets. We convert our datasets (training and testing) into a matrix of feature space; where each entity represents a TF of respective column feature in a given tweet. In phase 5, we implement two independent one-class classifiers (KNN and LibSVM) to classify a tweet as hate promoting or unknown. Algorithm 4 & 5 describes the procedure of KNN and LibSVM classifiers respectively. In the last phase, we evaluate the performance of the two classifiers using standard confusion matrix. An accuracy of the classifier is computed in terms of precision, recall, and f-score.

4.4.2 Solution Implementation

A one-class classifier learns from a training data containing all instances of one class (positive or target class). The aim of the one-class classifier is to identify if a new instance belongs to the target class or is an outlier. As discussed in Section 2.6.1, conventional binary and multi-class classification algorithms classify the input data into pre-defined labels. However, the challenge arises when the input does not belong to any class or is irrelevant to the classification task. In one-class classification algorithms, the training data

Algorithm 4: One Class K-NN Algorithm

Data: Training Dataset D_{tr}, Test Dataset D_{te}, Neighbors K, Threshold th
Result: List of class labels for test dataset C_{te}
Algorithm $OneClassKNN(D_{tr},\ D_{te},\ th,\ K)$

```
1  for each instance I ∈ D_te do
2      N1 ← NearestNeighbor(I, D_tr)
3      D_1 ← Euclidean_Distance(N1, I)
4      for given value of K do
5          ND_i ← Euclidean_Distance((D_tr), N1)
6          D_2 ← Average(ND_1, ND_2, ....., ND_K)
7      if (D_1/D_2 > th) then
8          C_te.addClass(Unknown)
       else
9          C_te.addClass(TargetClass)
10 return C_te
```

contains the sample of the relevant class (referred to as the positive class or target class) [90]. While for the other class training data does not contain any instances. The one class-classification algorithm is trained on the samples of only positive class instances, and it classifies as many instances as possible from target class and minimizes outliers. Therefore, it requires identifying strong and discriminatory features to find the best separation of both target and unknown class instances. In work presented in this case-study, we implement two independent one-class classifiers i.e. K-Nearest Neighbor (KNN) and Support Vector Machine (SVM).

4.4.2.1 K-Nearest Neighbor Classifier

The conventional K-Nearest Neighbor (KNN) classifier is the simplest method for identifying the patterns in data. The kNN rule classifier trains the model based on the distances between each instance present in the dataset. For any unknown input data point, it computes the distance between new input and existing data points and identifies the K nearest neighbors. KNN classifies an unlabeled data input by the majority label among its neighbors. Therefore, the performance of KNN depends crucially on the distance metric used to identify nearest neighbors [157]. The single-class kNN classifier has a number of parameters that may be adjusted and customized to change the performance of the algorithm; the number of k nearest neighbours can be changed; the threshold value of accepting outliers or rejecting target classes can be changed; furthermore, the distance metric can also be changed. We use one-class KNN classification algorithm [158] for classifying extremist content since our data is projected as large data points in a low dimensional space. The proposed method (Algorithm 4) follows the standard one class KNN algorithm to classify a tweet into one of the classes: hate promoting or unknown. Inputs to this algorithm are pre-processed training dataset D_{tr}, testing dataset D_{te}, the number of nearest neighbors K and a threshold measure th for accepting outliers. Each tweet in testing dataset is an arbitrary instance I that is represented by a feature vector $(f_1(I), f_2(I), ..., f_m(I))$, where $f_i(I)$ is an instance value for a given feature, and m is the number of discriminatory features. In steps 2 and 3, we compute Euclidean distance [159] between an instance I of testing data and all the instances of training datasets.

$$D = \sqrt[2]{\sum_{i=1}^{n}(f_i(I) \sim f_i(J))^2}, \text{ where } J \in D_{tr} \tag{4.1}$$

We create a distance matrix of size $n * 1$ for every instance $I \in D_{te}$, where n is the size of the training dataset. Equation 4.1 shows the formula for computing the euclidean distance between two instances. Based upon this distance matrix, we find the nearest neighbor N_1 of I in training data. In steps 4 to 6, we find K nearest neighbors of N_1 in training dataset D_{tr}. Due to the large size of the testing dataset, we use $K = 100$.

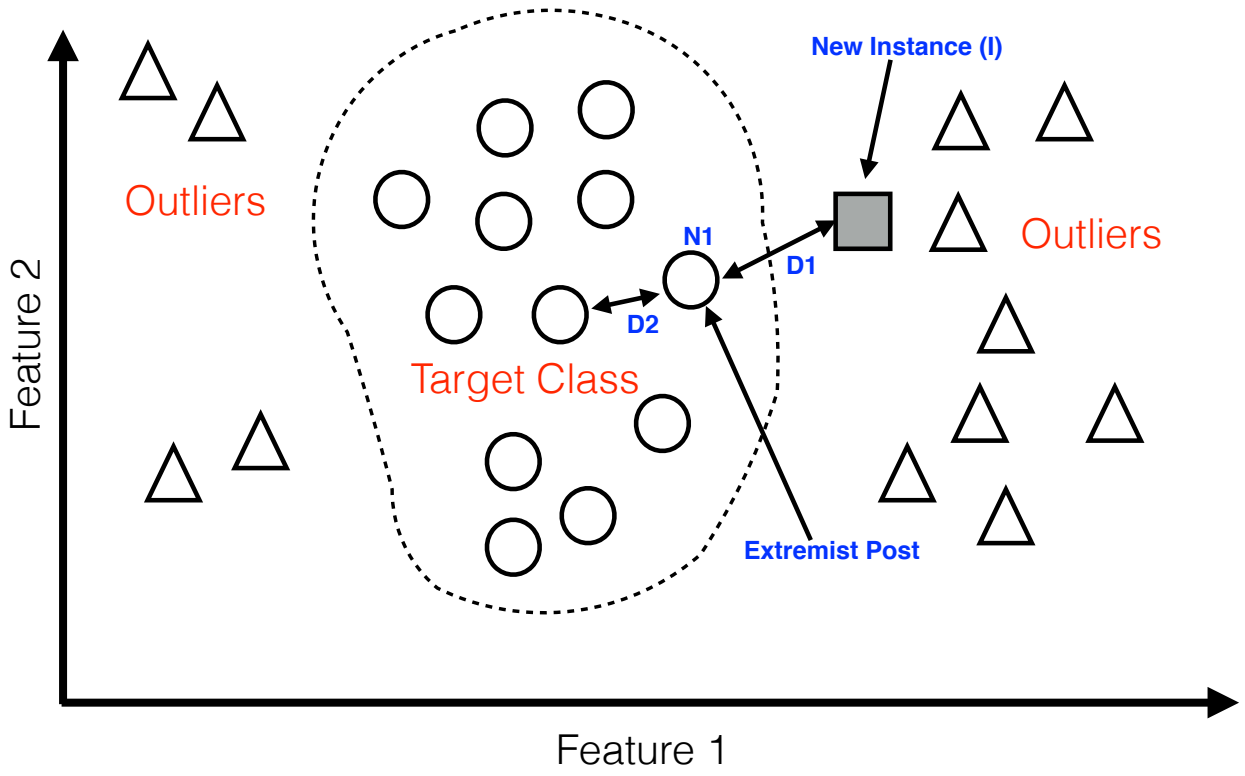


Figure 4.7: One class K-NN Classification for Classifying Extremist Tweets from Unknown or Outliers.

In step 6, we take an average of all K distances and name it as D_2. Steps 7 to 9 perform unary classification. If the ratio of distances D_1 and D_2 comes out to be lower than threshold measure th, then instance I belongs to the target class otherwise, it is classified as unknown. The intuition of such computation is to measure the local density of new instance to the local density of its nearest neighbor in training dataset (or target class) [160]. The distance D1 between the new instance I and its nearest neighbor N_1 in the training set D_{tr} is compared with the distance between this nearest neighbor N_1 and its nearest neighbor N_2 in the training set D_{tr}. When the first distance D_1 is much larger (larger than a defined threshold) than the second distance D_2, the test instance is identified as an outlier. if D_1/D_2>threshold then C_{te} is an outlier or unknown. We compute an extent of similarity (Euclidean distance) between all instances of training dataset D_{tr}. As a result of this, we get a distance matrix of size n*1. We take a harmonic mean of these distances and come up with the threshold value th.

4.4.2.2 Support Vector Machine Algorithm

Support Vector Machine (SVM) is a supervised learning based classification algorithm that projects the data points in a multi-dimension space. SVM classifier discriminates the data point by separating them with a hyperplane [161]. While several hyperplanes can be created around the data points, SVM identifies the right hyperplane that has the maximum distance from the nearest data point (called as a margin). Despite the low dimensional space of our dataset, we use SVM classification algorithm for our experiments since the size of training data is relatively small [162]. We develop an algorithm that classifies most positive

Algorithm 5: One Class LibSVM Algorithm

Data: Training Dataset D_{tr}, Testing Dataset D_{te}
Result: List of class labels for test dataset C_{te}
Algorithm $OneClassLibSVM(D_{tr},\ D_{te})$
1 $Class\_Label \leftarrow SVM.setTargetClass(D_{tr})$
2 $Model \leftarrow SVM.buildClassifier(D_{tr})$
3 Preprocessed dataset $T_p \leftarrow FilterTweets(D_{te})$
4 **for *each*** *instance* $i \in T_p$ **do**
5 Class $c \leftarrow Model.classifyTweets(i)$
6 $C_{te} \leftarrow c$
7 return C_{te}

class tweets from outliers. In our research, we use LibSVM Java library 3.18[8] for Weka 3.7.10[9], originally proposed by Chang et. al. [163]. LibSVM is a wrapper class that allows one class SVM classifier supported by LibSVM tool. In one class LibSVM, all SVM formulations are supported as a quadratic minimization problem. Equation 4.2 shows the formulation of unconstrained dual form of standard SVM classifier, subject to a Lagrange multiplier α that varies between 0 & a constant value C. Q is a n*n matrix where n is the size of training vectors, and e is a vector of all ones represented as [1,1,....,1]. To constraint in minimization, we optimize margin hyperplane as $y^T\alpha = 0$. In one class LibSVM (Equation 4.3), we solve a scaled version of Equation 4.2 subject to α that varies between 0 & 1 [163]. Given training vectors x_i where $i = 0, 1, ..., n$, $v \in (0, 1)$, where 0 denotes a lower limit of support vectors and 1 denotes an upper limit on errors made in training a model. Equation 4.4 shows the kernel function Q_{ij} of one class LibSVM i.e. a dot product of two training vectors.

Algorithm 5 describes basic modules of LibSVM that we implement in our classifier. We give an input of a training and testing dataset to the algorithm i.e. D_{tr} and D_{te} respectively. Training dataset contains a set of labeled feature vectors of only target class tweets. In steps 1 and 2, we set the target class label and build our model on the training dataset. In step 3, we perform data pre-processing on testing dataset and remove all garbage (non-informative and non-content bearing) data using Algorithm 3. Steps 4 to 6 performs classification and predicts most likely class for a given instance of the testing dataset.

$$min_\alpha\{\tfrac{1}{2}\alpha^T Q\alpha - e^T\alpha\}, \quad 0 \leq \alpha \leq C, \quad y^T\alpha = 0 \tag{4.2}$$

$$min_\alpha\{\tfrac{1}{2}\alpha^T Q\alpha\}, \quad 0 \leq \alpha \leq 1, \quad e^T\alpha = vn \tag{4.3}$$

$$Q_{ij} \equiv K(x_i, x_j) = (x_i.x_j) \tag{4.4}$$

We also implement leave-p-out cross validation strategy in both KNN and LibSVM classifiers. Most of the cross-validation techniques split data into a fixed size of training and testing dataset and use all features dimensions for classification. The leave-p-out cross validation considers all training samples and instead leave p features out at a time. Each possible subset of feature vectors is left-out iteratively, and training sample for the remaining features is used for the cross-validation [164]. Leave-p-out is an exhaustive method to find the relevance and importance of each feature vector in the classification of target classes. We perform a column-wise partition on both training and testing datasets and remove p feature/s at a time. We repeat this process for all features and run our proposed classifiers $2 * {}^mC_p$ times, where m is the size of feature

[8] `http://www.csie.ntu.edu.tw/$\sim$cjlin/libsvm/`
[9] `http://www.cs.waikato.ac.nz/ml/weka/`

Table 4.3: Confusion Matrix and Accuracy Results for KNN and LibSVM Classifiers

(a) KNN Classifier

		Predicted	
		Positive	**Unknown**
Actual	**Positive**	67,798	15,522
	Unknown	74,968	841,712

(b) LibSVM Classifier

		Predicted	
		Positive	**Unknown**
Actual	**Positive**	73,555	9,765
	Unknown	20,420	896,260

(c) Accuracy Results of Classifiers

Classifier	Precision	Recall	TNR	NPV	F-Score	Accuracy
KNN	0.48	0.81	0.92	0.98	0.60	0.90
LibSVM	0.78	0.88	0.98	0.99	0.83	0.97

space, and p is the number of features we remove per iteration, $p = 1$ in our case. As a result of this, we get a $1*m$ matrix for one classifier, where each instance shows the overall accuracy of the respective classifier.

4.4.3 Empirical Analysis & Performance Evaluation

4.4.3.1 Experimental Dataset

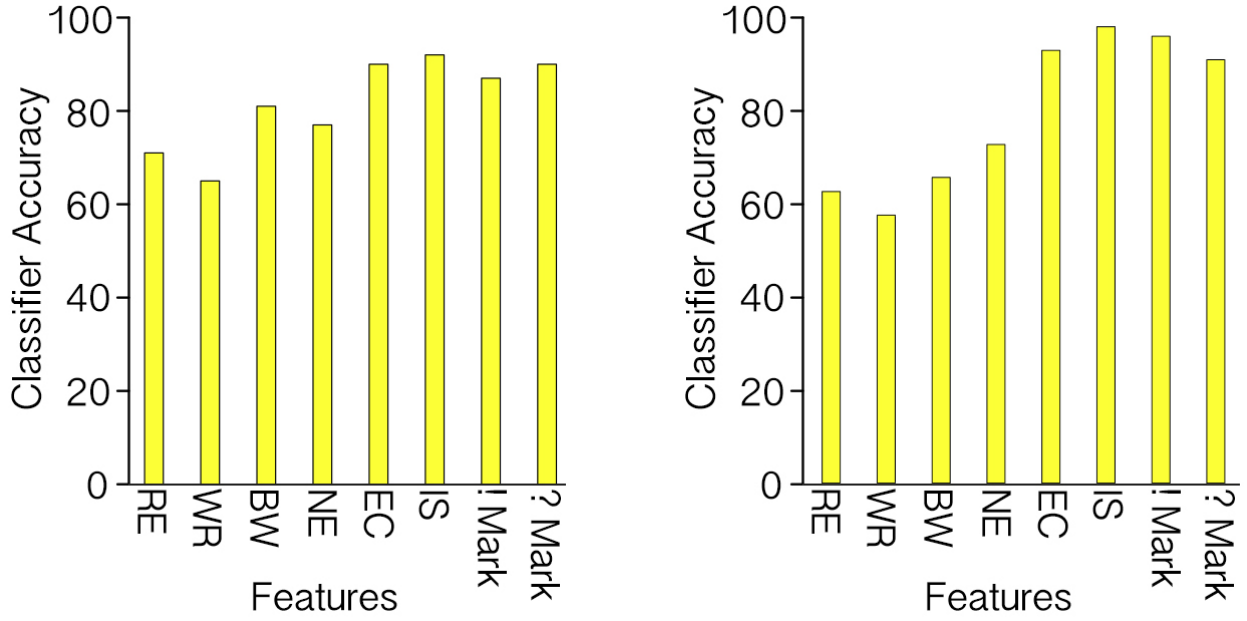
We conduct experiments on publicly available dataset so that our results can be replicated or used for benchmarking or comparison purposes. We download two datasets: UDI-TwitterCrawl-Aug2012[10] and ATM-TwitterCrawl-Aug2013[11]. UDI-TwitterCrawl-Aug2012 consisted of 50 million tweets approximately and was collected in May 2011 [156]. ATM-TwitterCrawl-Aug2013 consisted of 5 million English tweets and was collected in June 2011 [155]. We use language detection library6 for Java for language identification (supports 53 languages) of tweets and find 29 different languages in the dataset. Using Language Detection API, we were able to identify only 85% tweets as English language tweets. Initially, we have $53,234,567$ tweets in our dataset. In this work, we focus only on the English language tweets. Therefore we discard all non-English $(7,889,609)$ tweets and remain with a total of $45,344,958$ tweets. We perform a semi-supervised learning approach on the experimental dataset and collect only hate & extremism promoting tweets. To avoid overfitting in classification, we collect only $10,486$ labeled tweets for training dataset, which is a very small fraction of experimental dataset. We perform a random sampling on all 45.3 million tweets of the experimental dataset and collect a random sample of 1 million tweets as testing (or validation) dataset. This dataset includes both hate promoting and unknown tweets.

4.4.3.2 Experimental Results

To evaluate the performance of our proposed solution approach, we use basic measures of relevance used in information retrieval and machine learning. We asked four graduate students to manually annotate each tweet in the dataset and based upon their decisions we validate our results. We compute accuracy of our classifier using precision, recall, and f-score metrics. Table 4.3a and 4.3b shows the standard confusion matrix for KNN and LibSVM classifiers. We execute our classifiers on a testing dataset of size 1 million records containing tweets from both target class (positive) and outliers. One class KNN algorithm classifies

[10] `https://wiki.engr.illinois.edu/display/forward/Dataset-UDI-TwitterCrawl-Aug2012`

[11] `https://wiki.engr.illinois.edu/display/forward/Dataset-ATM-TwitterCrawl-Aug2013`

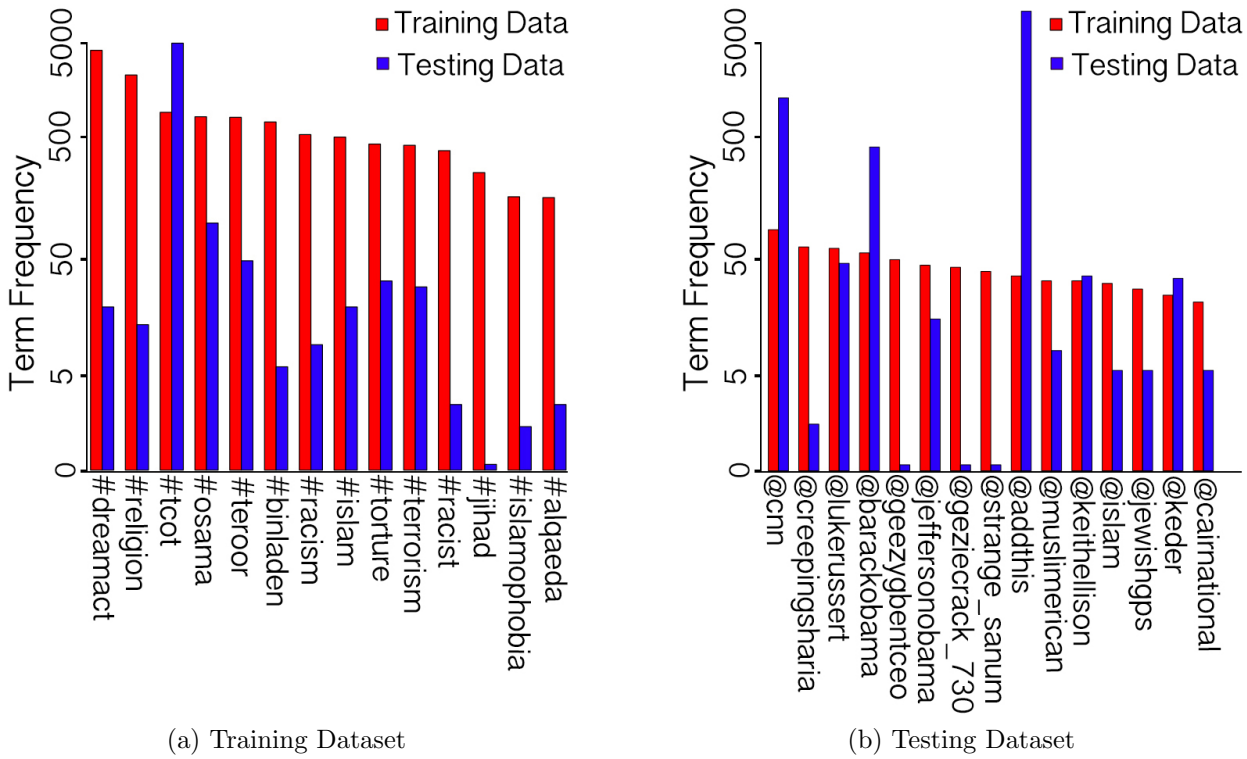


(a) KNN (b) LibSVM

Figure 4.8: Impact of Individual Feature on Overall Accuracy of A Classifier. RE= Religious, WR= War Related, BW= Bad Words, NE= Negative Emotions, EC= Emoticons, IS= Internet Slangs.

142,766 (67,798 + 74,968) tweets as positive and 857,054 (15,522 + 841,712) tweets as unknown. Table 4.3a reveals that there is a misclassification of 18.6% and 8.2% in predicting target (positive) class and outlier (unknown) instances. Similarly, given an input of 1 million tweets, one class LibSVM algorithm predicts 103,975 (73,555 + 20,420) tweets as positive and 906,025 (9,765 + 896,260) tweets as unknown. Table 4.3b shows that 11.7% and 2.2% of tweets are wrongly classified as positive and unknown respectively. Table 4.3c shows accuracy results (precision, recall, f-score) for both KNN and LibSVM classifiers. Table 4.3c reveals that overall LibSVM classifier (accuracy of 97%) outperforms KNN classifier (accuracy of 90%). Results shows that precision, f-score, and accuracy of LibSVM classifier are much higher in comparison to KNN classifier and similarly recall is reasonably high for LibSVM classifier.

We apply leave-p-out strategy for both KNN and LibSVM classifiers ($p = 1$) and compute their accuracy. As discussed in Section 4.4.1, we use 8 discriminatory features to classify a tweet as hate promoting or unknown. Figure 4.8 shows variance in overall accuracy of one-class classifiers (KNN and LibSVM) after removing one feature vector at a time. Figure 4.8a reveals that if we remove religious or war-related terms, then the accuracy of KNN classifier decreases by 20% to 25%. Whereas, by removing bad words or negative emotions from feature vectors, the accuracy falls by 11% to 13%. Figure 4.8a reveals that internet slangs, emoticons and punctuations (! and ? marks) are less important features and doesn't affect the accuracy by a significant difference but we cannot neglect them entirely because they influence the overall accuracy by 2% to 3%. Figure 4.8b reveals that in one class LibSVM classifier, the presence of religious, war-related terms, bad words, and negative emotions plays an important role. While, by removing any of these features, the overall accuracy of classifier decreases by 20% to 45%. Ignoring presence of internet slangs and exclamation marks doesn't affect accuracy. Unlike KNN classifier, removing emoticons and question marks decreases



(a) Training Dataset

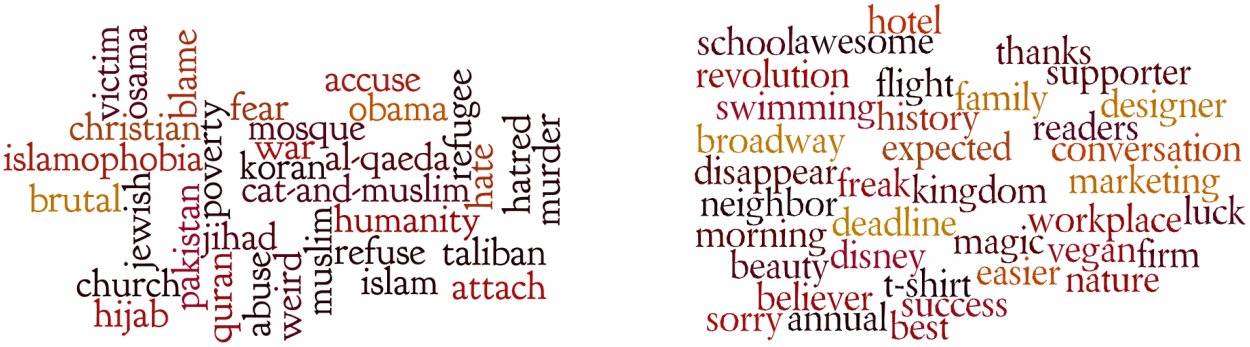
(b) Testing Dataset

Figure 4.9: Frequency of Top K Hashtags and @Usernames in the Tweets Present in our Experimental Dataset

the performance at a reasonable rate. The reason of this misclassification is the presence of noisy content and sparsity in datasets. Feature space of testing dataset is a matrix of size 1M*8, where 70% of entries are 0.

4.4.4 Technical Challenges and Manual Analysis of Tweets

Automatic identification of hate and extremism promoting tweets is a technically challenging problem. Tweets are user generated short text messages containing low-quality content. The presence of noisy data and unknown words (not present in training data or lexicons) increases the number of false alarms. We perform a content-based characterization on tweets present in the training and testing datasets. Table 4.4 shows a sample of tweets containing various key terms that are a clear evidence of a tweet to be hate promoting. For example, islamophobia, muslim, racism, black, Sharia, attack, terrorism, etc. We perform a content-based analysis on the tweets and observe that many non-hate promoting tweets present in the testing dataset contain these key terms. These tweets are posted as news or for general awareness (shown in Table 4.4). The presence of such tweets in testing dataset increases the number of false positives and degrades the classifier's performance. Figure 4.10 shows a word cloud of common terms that are present in training and testing datasets. Table 4.4 shows a sample of hate promoting tweets containing internet slangs and misspells words. These tweets contain many words that vary from user to user and therefore are unknown to the classifier. Excessive presence of abbreviations and noisy data makes it a challenging task to classify a short text message.



(a) Training Dataset (b) Testing Dataset

Figure 4.10: Word Cloud Presentation of Common Terms Present in the Training and Testing Datasets.

In Twitter, hashtags are a very important feature to identify the topic of a tweet. We perform a manual analysis on hashtags present in training and testing dataset. We find several non-hate promoting tweets present in our dataset containing hashtags that are frequently being used in positive class tweets and vice versa. Table 4.4 shows a sample of such less informative tweets. We perform an analysis on the training dataset and find top K hashtags present in the dataset (K=15). We further compute frequency of these hashtags in the testing dataset and plot them together in a graph (refer to Figure 4.9a). All values are plotted in logarithmic scale. Figure 4.9a reveals that frequency of these hashtags is very high in training dataset in comparison to the testing dataset. Therefore, the presence of such highly frequent hashtags from training dataset can be used a discriminatory feature to classify a new tweet as hate promoting or unknown. But similar to hashtag "#tcot", there are many hashtags present in training dataset that are frequently being used in testing dataset as well. Based upon the annotator's decisions, we label our testing dataset and observe that a large number of tweets containing these hashtags are an outlier. Therefore, if we use the presence of hate promoting hashtags as a discriminatory feature, it increases the number of false alarms and affects the accuracy of the classifier. Similar to hashtags, we perform an analysis on usernames mentioned in training and testing datasets. We find top K usernames mentioned in training dataset and compute their frequency in testing dataset. Figure 4.9b reveals that several usernames are mentioned in hate promoting tweets more frequently than the testing dataset. For example, @muslimerican, @islam, @creepingSharia, @jewishgps etc. Therefore tracking these users can be helpful in locating more active users and hidden groups that post hate and extremism promoting tweets and share a common agenda.

4.5 Case Study 2: Intent Classification of Racist Posts on Tumblr

As discussed in Section 4.1 that many like-minded people use popular microblogging websites for posting hateful speech against various religions and race. Automatic identification of racist and hate promoting posts is required for building social media intelligence and security informatics based solutions. However, just keyword spotting based techniques cannot be used to identify the intent of a post accurately. The work presented in this case study is motivated by the need to develop a system for automatically identifying the purpose of a racist and radicalized post. The specific research aim of the present study is the following:

Table 4.4: Content Based Characterization of Tweets- tweets consisting of hate promoting content, tweets posted as news, presence of slang and noisy words and the extremist tweets with less information.

Hate and Extremism Promoting Tweets		
Herman Cains Islamophobia: Why does he treat Muslims the way racists treat blacks? - @Slate http://slate.me/j25oNr		
The Islamophobia Machine - @thinkprogress on the organized right-wing coalition behind the various state sharia bans http://bit.ly/iU12Gq		
My heart goes out to people of #Oslo in aftermath of tragic attack; but shame on those who jumped to conclusions #Islamophobia		
Tweets Posted as News		
#Terrorism News Father Would Be Suicide Bomber Convicted Obstructing Terrorism Investigation: Mo... #HiTechCJ		
#Terrorism News ISAF Condemns Insurgent Attacks Afghanistan: Compiled from International Security A...		
BBC News Bid rescue #Danish hostages from #Somali #pirates fails south asia 12708168 #denmark #terrorism		
Tweets Containing Slang and Misspelld Words		
U can quote da bible tell other ppl bout it...but u dnt live like wut u telln?? Can u say #Hypocrite ???		
omg what atrocious TT #blamethemuslims lmao f****d up		
Lady Gaga txtd Anna Win tour when she received txt abt winning #CFDA "yes bitch we did it" thinking another Anna.		
Oooooh apolgies muslim friends but omg those piggie ribbies kicked asssss		
Less Informative Tweets		
#Tag	**Tweet**	**Class**
✓	What does mean #Muslim America ? #islam #religion **#islamophobia**	X
✓	Does Poverty Cause Terrorism ? #gop **#terrorist** #obama #dems #holder #taliban **#bin-laden** #pakistan	X
✓	Thanskgiving come passed but thankful would downtown #portland **#terrorist** stopped.	X
X	The birthers going through complex grieving process: Shock Denial Amnesia Blame Muslims	✓
X	The British considered Amer Revolutionaries terrorists because they refused fire lines.	✓
X	But I do think Al Qaeda other organizations force us change rules traditional war.	✓
X	I suspect Native Americans would disagree. Im not condoning torture state sponsored assassination.	✓

1. To investigate the efficacy of natural language processing techniques on microblogging dataset for topic and intent classification.
2. To investigate the application of linguistic features such as taxonomy, emotions, language cues, personality traits and text semantics for classifying the intent of Tumblr posts.
3. To conduct empirical analysis on real word dataset and examine the effectiveness of proposed one-class text classification approach. To compute the relative influence of each linguistic feature for identifying the posts having racist intent.

4.5.1 Problem Statement

Given a dataset D of Tumblr Posts P_i, $D = \{P_i \mid 1 \leq i \leq n\}$, a set of topics $N = \{N_j \mid 1 \leq j \leq m\}$ and a target class C; identify the intent of $P_i \in D$ when $P_i \in N$.

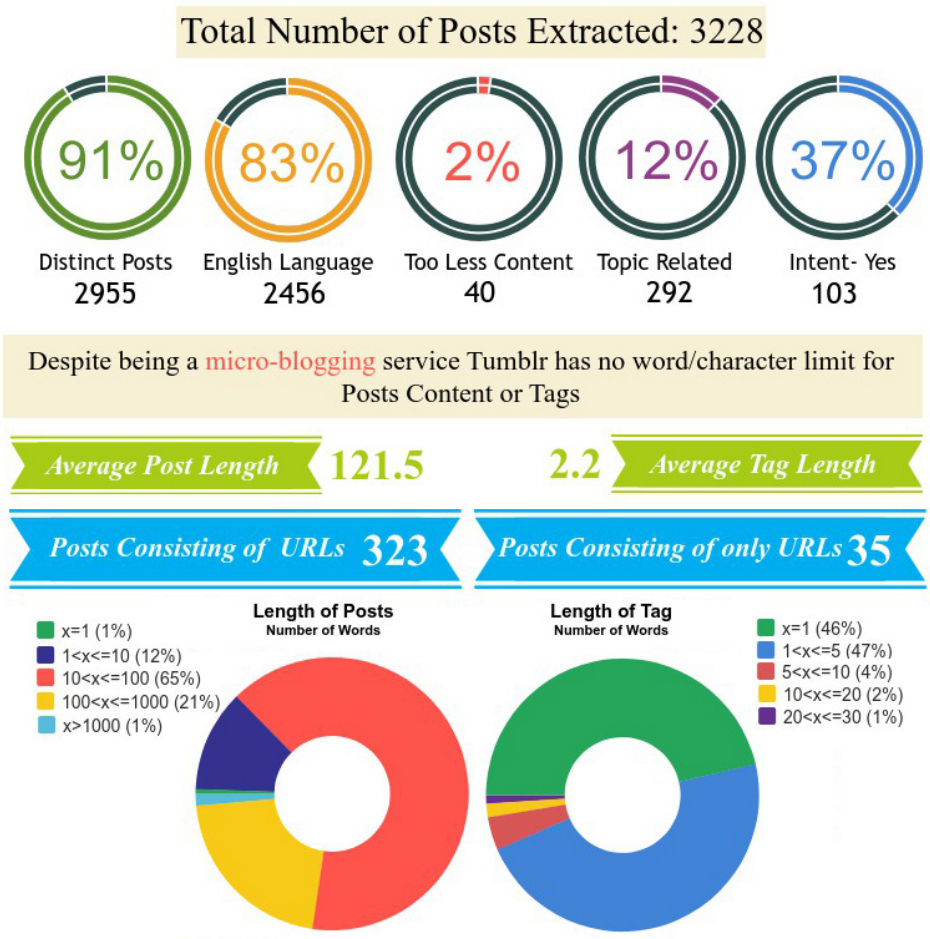


Figure 4.11: An Infographic Presentation of Tumblr Posts Collected, Filtered and Sampled after the Pre-processing and Annotation. Further Illustrating the Statistics of Multimedia and Other Contextual Metadata Present in English Language Posts.

Based on the definition of freedom of expression by Joshua Cohen [118], we define a Tumblr post P_i as a racist intent post if 1) the topic of the content belongs to a race or a religion and 2) the post targets a community in an offensive or persuasive manner (in a recognizable way). To identify a racist or radicalized intent post, we propose following two hypotheses:

1. In the absence of topic related key terms, natural language processing can be an efficient approach to identify the hidden taxonomy of a Tumblr post.
2. Sentiment and semantic enrichment of text can be two discriminatory features for identifying the language of narrative and classifying the intent posts.

Table 4.5: Detailed Schema of Tumblr Database Consisting of Posts and Bloggers Metadata Extracted using Tumblr Search API.

Posts
Post_ID | Timestamp | GMT | Blogger | URL | Type | Tags | Num_Tags | Notes | Re-Blogged_From | Title | Description
Blogger
Blogger_ID | Ask | Ask_Anon | #Likes | #Posts | Title | Description

4.5.2 Experimental Setup

4.5.2.1 Data Collection

We conduct our experiments on an open source and real-time dataset extracted from Tumblr microblogging website. We perform a manual inspection and find most popular Tumblr posts having racist and radicalized intent. We extract the list of unique tags associated with these posts and create a lexicon of top K tags that are the most commonly used by racist or radicalized groups. For example, #islamophobia, #islam is evil, #supremacy, #blacklivesmatter, #white racism, #jihad, #isis and #white genocide. We implement a bootstrapping method to create our dataset and use this lexicon as seed tags for the Tumblr Search API[28]. For each tag, we extract only textual posts (text and quote) and extend our lexicon by acquiring other (unique) related tags associated with these posts. We execute our model until we get the desired number of posts or the model converges (it starts extracting duplicate posts). Using Tumblr Search API, we were able to extract a total of $3,228$ text posts made by $2,224$ unique bloggers consisting of $10,217$ unique tags. Table 4.5 shows a complete schema of additional metadata extracted for each post and unique blogger. The aim of the study presented in this case study is to build a one-class text classifier for identifying racist and hate promoting intent posts. Therefore, we conduct our experiments on post content (referred as description in Tumblr). Since, Tumblr generates a new identification number for each post (re-blogged or posted), despite having the unique Post IDs, we discard 9% (273) of the posts having similar or duplicate content and remove the bias from our data.

The study presented in this case study focuses on the intent mining on English-language posts. We identify the language of each record by applying Alchemy language detection API[114] on the description of each post. Figure 4.11 reveals that only 83% ($2,456$ out of $2,955$) of the posts have English language content and 459 posts were identified as non-English. The language of remaining 40 posts (2% of the data) was identified as 'unknown' due to the insufficient content in post description, for example, the posts containing only URLs. Figure 4.11 reveals that 35 out of $2,955$ posts contain only URLs. We conduct our experiments on $2,456$ English language posts and discard the other non-English or unknown language records. We apply various natural language processing techniques for semantic and sentiment enrichment of our data (discussed in Subsection 4.5.3.1). We enhance our data and make it publicly available so that our experiments can be used for benchmarking and comparison [63]. Our dataset is the first ever published data of Tumblr posts and bloggers labeled with various sentiments and semantic features and can be downloaded from Mendeley Data[12]. Figure 4.11 summarizes the statistics of our experimental dataset. Despite being a microblogging website, Tumblr has no word or character limit and allows users to make long posts and tag with any number or length of keywords. We remove all noisy text from the post descriptions and tags including special characters, emoticons, extra white spaces and compute their length. Data statistics reveals that 21% of the posts have word length between 100 and $1,000$ while 25 posts have a length greater than $1,000$ words. Similarly, 4% (408 out of $10,217$) of unique tags have a word length between 5 to 10 while 10 unique tags have a length between 20 to 30 words.

[12] `https://data.mendeley.com/datasets/hd3b6v659v/2`

Table 4.6: Inter-Annotator Agreement Results for Topic and Intent Labelling of Experimental Dataset

(a) Topic Annotation

		A2	
		Topic	**NA**
A1	**Topic**	292	24
	NA	13	2127

(b) Intent Annotation

		A2	
		Intent	**NA**
A1	**Intent**	103	2
	NA	12	175

(c) Cohen's Kappa Coefficient

	Topic	**Intent**
Observed Agreement Po	0.98	0.95
Random Agreement Pr(e)	0.77	0.51
Kappa Coefficient κ	0.91	0.95

4.5.2.2 Data Annotation

We use 2456 English language posts for annotation which spans only 83% of the extracted data. Since we are using bootstrapping method to collect our data, it retrieves a large number of noisy posts that do not belong to the defined topic (race and religion). Therefore, we first identify the topic related posts and later label them as intent (racist/radicalized) or unknown (we don't know the intent of the author). To annotate these posts, we employ two annotators with 2 to 3 years of experience of using Tumblr website. Each annotator first labels a post as topic or unknown (NA) based on the content description and the tags associated with the post. If a post is annotated as a topic-related post, then these annotators further label it as intent or unknown (NA). To create ground truth for our data, we measure the inter-annotator agreement and compute Cohen's Kappa coefficient between both annotations. Table 4.6 shows the results of topic and intent annotation performed on $2,456$ posts. Table 4.6a reveals that we get $2,419$ (292 topic and $2,127$ unknown) posts as same label from both the annotators. We discard the remaining 37 posts with inconsistent annotation. Both the annotators further label these 292 topic posts as intent or unknown. Table 4.6b reveals that the annotators agree on 278 posts (103 intent, 175 unknown) while there is an inconsistency in remaining 14 posts. Table 4.6c shows the value of Cohen's Kappa coefficient between annotators for both topic and intent annotation. Results reveal that the annotators agree more than 90% of the time. Figure 4.11 shows that the intent posts are only 37% of topic posts and only 4% of the complete experimental dataset, revealing that the labeled data is highly imbalanced. Since we use a tag search based bootstrapping method, we analyze all the tags extracted during the process and find that it happens due to the various limitations of user generated tags. For example, the presence of noisy content (spell errors), long text, multi-lingual tags, use of featured tags and tags that redirects to a non-topic based post such as 'vote', 'lol', 'media', 'news', 'life', 'travel'.

4.5.3 Proposed Approach

Figure 4.12 shows the high-level architecture of proposed approach primarily consisting of three phases: Data Extraction, Feature Identification, and Classification. subsection 4.5.2 describes the bootstrapping method used for data collection and inter-annotator agreement used for creating ground truth. We describe the remaining two phases in the following subsections:

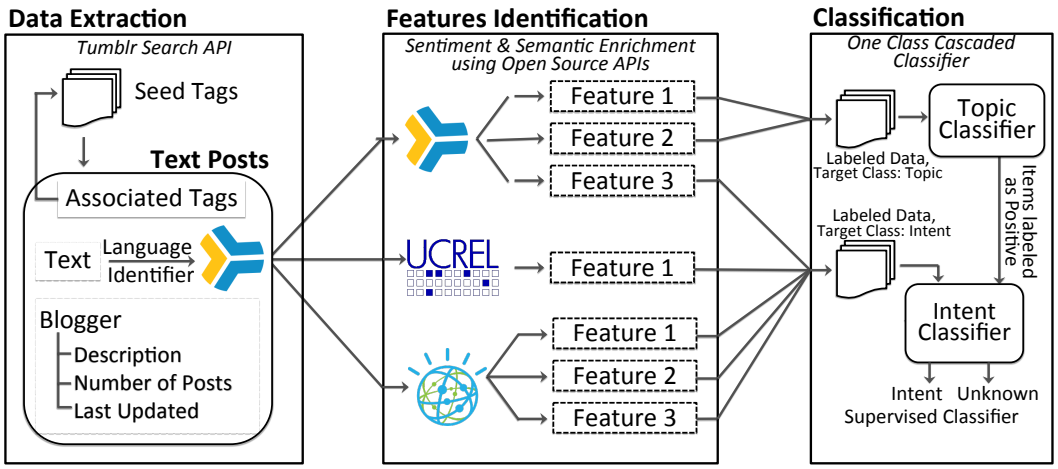


Figure 4.12: A General Research Framework for the Experimental Setup and Proposed Methodology for the Identification of Intent-Based Racist and Radicalized Posts on Tumblr Website.

4.5.3.1 Features Identification

Based on the prior literature and our hypothesis design, we create our feature space by analyzing the linguistic features (semantic and sentiment tone) of Tumblr posts. We divide our features set into three categories: Topic Modeling, Tone Analysis, and Semantic Tagging. We also discuss other contextual metadata features that can be extracted from Tumblr posts but are not applicable in intent classification.

4.5.3.1.1 Topic Modeling

The existing literature shows that there has been a lot of work in the area of mining user-generated content on social media related to offensive speech [132], racism and radicalization [165] [166]. However, our analysis and annotation reveal that despite not having certain topic specific key terms, a post can be an intent post for which keyword-based classification method do not work accurately and generates a large number of false alarms [30]. Therefore, we use statistical and natural language processing techniques to perform topic modeling on Tumblr posts. We use Alchemy Taxonomy API[13] to classify the post into the most likely topic and sub-topic categories. Alchemy API supports over 1000 categories broadly divided into 23 topics. Sub-topic categories allow us to identify the more focused and targeted topic of the post (up to 5 levels of hierarchy). For example, society/crime/personal offense/hate crime. We also use Alchemy Concept Tagging API[14] to identify the hidden concepts in the text that are similar to human annotation. Alchemy API learns about a post from 9 linked data resources[15] such as Freebase, DBpedia, yago, and tags the concepts that are high likely to be related to the given text. For example, for a Tumblr post "*If the Arabs put down their weapons today, there would be no more violence. If the Jews put down their weapons today, there would be no more Israel.*", Alchemy tags "Ashkenazi Jews", "Palestinian people" and "Jewish ethnic divisions" with a confidence score of 0.74, 0.78 and 0.70 respectively. We use these concepts to perform the topic modeling of a text along with the taxonomy. Statistically, the API returns a confidence score of

[13] `http://www.alchemyapi.com/products/alchemylanguage/taxonomy`
[14] `http://www.alchemyapi.com/api/concept-tagging`
[15] `http://www.alchemyapi.com/api/concept/ldata.html`

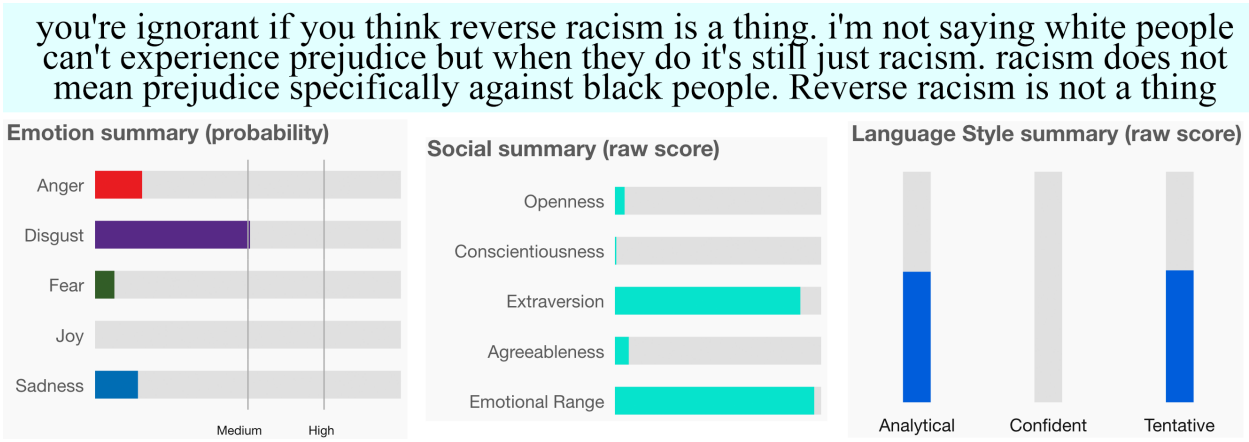


Figure 4.13: Example of Emotion, Social and Writing Tone Features Computed for a Tumblr Post, Topic: Race, Intent: No

each taxonomy conveying how likely the post belongs to a derived category. We discard a category from taxonomy and concept lists if the confidence score is below 40%.

4.5.3.1.2 Sentiment and Tone Analysis

Inspired by the existing literature [49], we investigate the language of the narrative by analyzing various types of sentiments and personality traits in a post such as a document sentiment, social tone, writing tone and emotions. We use Alchemy Document Sentiment API[16] to identify the document-level polarity of overall sentiment of a post. We define five categories of sentiment polarity: strongly negative, negative, neutral, positive and strongly positive and categorize each post based on its sentiment score. The sentiment of a document or post differs from the tone analysis of the content. Sentiment analysis can only identify the positive and negative polarity of a post while tone analysis measures the level of three categories including emotion, social and writing tones. We conduct a linguistic analysis on Tumblr posts using IBM Watson Tone Analyzer API[17]. Emotions tone analyzes the text of a post and gives a distribution of 5 emotions namely joy, fear, sadness, anger and disgust. Social tendencies analyze the personality traits from the text that includes openness, conscientiousness, extraversion, agreeableness and emotional range of a narrative. Writing tone identifies the language cues of the author in context to the content written in a Tumblr post. It includes the analytical, confident and tentative style of writing. The Tone Analyzer API analyzes the content of a post and computes two scores (document level and sentence level) for all three categories of tones. Since the text length of posts in our experimental dataset varies from 1 to 1200 words, we select only document level measures of these tones. Similar to sentiment score, we create a feature vector of each tone and categorize each post based on the confidence score: very low, low, medium, high, and very high. Figure 4.13 shows a concrete example of Tumblr post related to Race topic and shows the level of emotion tone, language and personality traits of author.

4.5.3.1.3 Semantic Tagging

Semantic tagging of a post identifies the semantic role of each term present in the content. It also identifies the hidden phrases playing a major role in the post. We use UCREL Semantic Analysis System

[16] http://www.alchemyapi.com/api/sentiment-analysis
[17] https://www.ibm.com/watson/developercloud/

(USAS)[18] to semantically tag each post in our dataset. USAS contains a hierarchy based lexicon of 232 categories with 21 major labels at the top of the hierarchy. All the semantic tags in a post are composed of a general or high-level label and a numeric value showing the division of each label in the lexicon. A numeric value after the decimal shows a further sub-division of categories in the hierarchy. For example, term "*refugee*" is tagged as "*M1/S2mf*" where *M1* denotes the tag 'moving from one location to another', *S2* denotes 'people' and *mf* denotes the 'gender'. We use USAS for semantic tagging because it not only tags each word of the document but also tags multi-words unit in the post if any. For example, term "New York Times" is tagged as "New_Z3c[i4.3.1 York_Z3c[i4.3.2 Times_Z3c[i4.3.3" where *Z3* denotes the name of a company, *c* denotes an anaphora, *i* denotes a multi-words unit and following numeric terms present the number of words present in a unit (3). We remove all punctuations and special characters (tagged as PUNC) from semantically tagged content and decode all remaining terms with their respective labels in tags' hierarchy[19]. USAS tags a term as *Z99*, if the term is not identified and not present in USAS database. We, however, do not remove them from the tagged content. Because USAS labels various topic-specific terms as *Z99* that are important for the intent identification. For example, 'Jihadist', 'racial', 'anti-white', 'pro-black', join words such as 'BlackLivesStillMatter'. It also includes the terms with hashtags, URLs, misspelled words, acronyms and abbreviations.

4.5.3.1.4 Contextual Metadata

Tumblr API allows us to extract the following contextual metadata associated with each Tumblr post: number of tags, terms used in the tags, the number of notes (reblog + like count) and link to multimedia content such as image, video or audio attached with the post. By further mining the content of a post, we can extract the following contextual information: hashtags, URLs, emoticons and Internet slang. However, due to various limitations, we exclude this contextual metadata from our feature space.

1. As discussed in subsection 4.5.2, the length of unique tags present in our dataset varies from 1 to 35 and contains a large amount of noisy text (multilingual terms, misspell words). Tags are user generated content, and a Tumblr post can have any number of tags (up to 30 in our dataset) or no tags at all. Further, the presence of a comma in a long sentence splits a tag into two separate terms. Given the length of tags in our dataset, the number of tags cannot be a discriminatory feature.
2. For a given tag, Tumblr API allows us to extract only most recently published posts. These posts automatically have relatively less number of reblogging or like count (referred as notes) in comparison to the posts containing featured or popular tags or uploaded before the current timestamp. Hence, the number of notes is not a valid feature for our experimental data.
3. Since, we extract only textual posts for our analysis, our dataset does not contain any multimedia content such as image, video or audio attached in the post description.
4. We conduct an exploratory data analysis on all topic related posts. Our data reveals that for both intent and unknown posts, there are very few (up to maximum 10) posts that contain either of hashtags (hashtags in Tumblr posts are not clickable and searchable), emoticons, Internet slangs (usually present in tags than the post content), @user mention or external URLs. We exclude contextual metadata from our feature space as those are not discriminatory for intent or topic classification.

4.5.3.2 Classification

The third phase of our proposed framework is a cascaded ensemble learning based classifier primarily consisting of two stages: topic classification and intent classification. We train our model from feature vectors created in Phase 2 and perform one-class classification on Tumblr posts.

[18] `http://ucrel.lancs.ac.uk/usas/`

[19] USAS published a list of all semantic tags is available at `http://ucrel.lancs.ac.uk/usas/semtags.txt`

Table 4.7: Illustrates the Simplified Codes Grouped Feature Vectors. Further, Shows the List of Individual Features (Personality Traits) Grouped Together in one Vector.

Code	Grouped Features
F1	Document Sentiment
F2	Semantic Tagging
F3	Emotion {Anger, Fear, Joy, Disgust, Sadness}
F4	Writing {Analytical, Confident, Tentative}
F5	Social {Openness, Conscientiousness, Extraversion, Agreeableness and Emotional Range}

4.5.3.2.1 Topic Classification

To identify the posts that belong to a defined topic (Race or Religion), we use topic modeling linguistic features extracted using natural language processing. We take a random sample of 50 posts out of 292, annotated as topic posts and extract their taxonomy and concepts from the feature space. We create two independent lexicons of these concepts and labeled topics that have a confidence score above 0.40. We manually filter the list of taxonomy and finalize the following 6 labels that strictly belong to the topic of this study:

1. religion and spirituality
2. society/unrest and war
3. society/racism
4. society/personal offense/hate crime
5. law, govt & politics/espionage and intelligence/ terrorism
6. law, govt & politics/legal issues/human rights

We use a look-up based method and check if the post belongs to any of these taxonomies and has a confidence score above 0.40. If yes, then we classify it as a topic post. However, if a post contains a wide range of taxonomies (>5), then we identify the top K concepts in the text and check if they exist in the concept lexicon of labeled topic posts. This stage of cascaded classifier is a one-class classifier that takes complete experimental dataset as an input and classifies topic related posts from unknown posts.

4.5.3.2.2 Intent Classification

An intent of a post (consisting of free-form text) cannot be fully determined only by mining the keywords in the content. But it also requires to understand and predict the psychological tendency, sentiment tones, and language of the narrative. It also requires analyzing the semantic role of topic-related keywords used in the post. We perform classification on Tumblr posts by training our model on sentiment, semantic and language cues based features of a text. On a high level, we create a vector space of 5 features set (F1 to F5) which is further categorized into 15 unique vectors. Table 4.7 shows the list of all features extracted and grouped into 5 feature vectors. We define intent classification as a one-class classification problem. Therefore, our training data contains only positive class (intent) posts. We implement three different one-class classifiers (Random Forest (RF), Naive Bayes (NB) and Decision Tree (DT)) and compare their accuracy for the posts classified as a topic in Stage 1. We train our model for each classifier and perform 5 fold cross validation. As discussed in subsection 4.5.2 and shown in Figure 4.11, only 12% of the posts are labeled as intent posts making our experimental dataset highly imbalanced. Further, intent classifier takes only the topic posts as an input classified by a topic classifier which is again a small subset of the whole dataset. Therefore, we select classification algorithms that work for small training data.

Table 4.8: Confusion Matrix for Topic Classification

		Predicted	
		Topic	**Unknown**
Actual	**Topic**	TP=253	FN=39
	Unknown	FP=93	TN=2034

Table 4.9: Performance Evaluation Metrics for Intent Classification.

	Test-Data1			Test-Data2		
	DT	**RF**	**NB**	**DT**	**RF**	**NB**
Recall	0.79	0.82	0.79	0.82	0.84	0.83
Precision	0.72	0.78	0.74	0.75	0.81	0.78

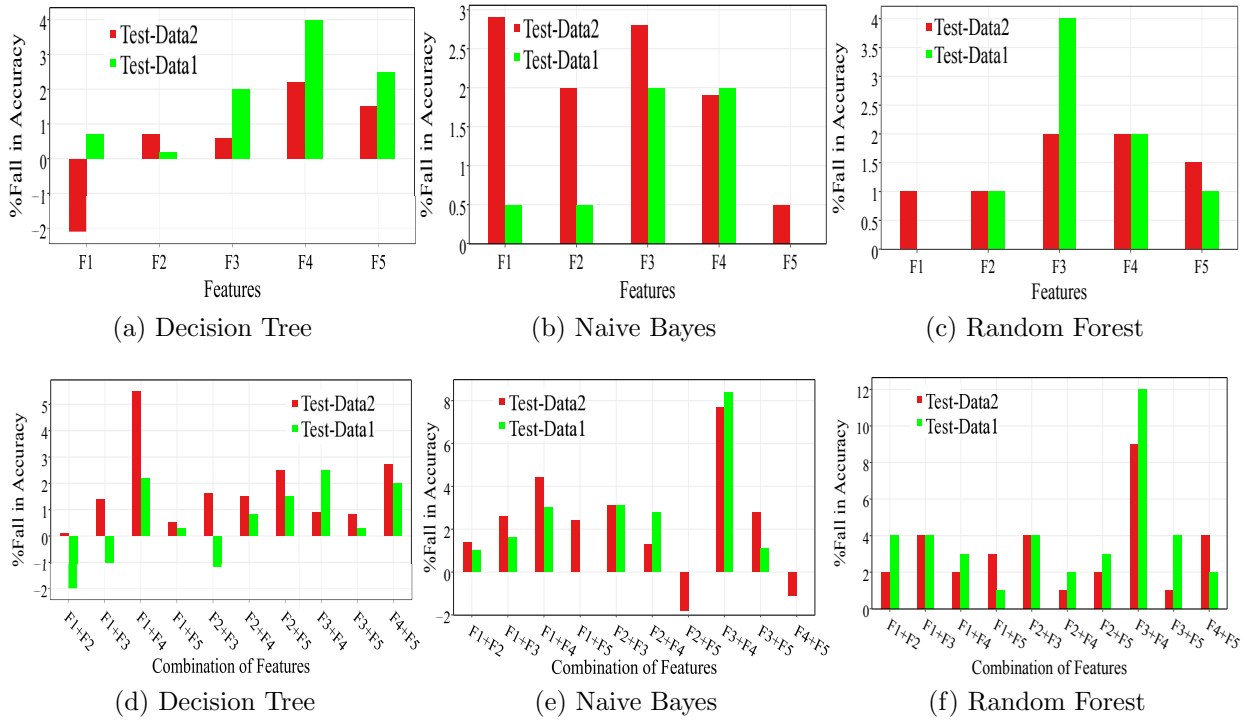
4.5.4 Performance Evaluation

As described in subsection 4.5.3.2, proposed method is a cascaded ensemble learning classifier in which topic classifier uses the complete experimental dataset as an input while intent classifier takes input from Stage 1. In this subsection, we present the accuracy results of each classifier and also discuss the influence of topic classification's accuracy on the classification of the intent post. Based on the inter-annotator agreement results, we evaluate the accuracy of our classifier by comparing the observed results against labeled class. We conduct our experiments on $2,419$ posts, consistently labeled by both annotators. Proposed topic classifier classifies 346 posts as target (topic) class and $2,073$ posts as unknown. Table 4.8 reveals that there is a misclassification of 3.8% and 1.6% in identifying target and outliers (unknown) posts. Since the focus of our study is to identify all such posts that have racist or radicalized intent, our aim is to achieve high precision as well as high recall. Our results reveal that for topic classification, we are able to achieve a precision of 73% (253/(253+93)) and a recall of 86% (253/(253+39)).

Given that our data is highly imbalanced and only 12% of the posts are labeled as target (intent) class, we execute each of our classifiers (RF, NB, and DT) using a 5 fold cross-validation over the experimental dataset. Since the accuracy measures are biased towards the majority class, we evaluate the performance of intent classifier using two standard information retrieval metrics i.e. precision and Area Under Operator Receiver Curve (AUC). Due to the misclassification in topic modeling, we evaluate the performance of intent classification in two steps. We first execute our model on all 346 posts (Test-Data1) classified as a topic in the previous stage. In the second iteration, we evaluate the performance of the intent classifier on 253 Tumblr posts (Test-Data2) correctly classified as a topic. Table 4.9 shows the accuracy metrics for Random Forest (RF), Decision Tree (DT) and Naive Bayes (NB) algorithms.

Our results reveal that one-class intent classifier gives higher precision rate for Test-Data1 (refer to Table 4.9). However, filtering non-topic based posts from the dataset further improve the accuracy of intent classification. It is probably associated with the fact that the unknown posts represent a broad range of sentiments and language cues. Table 4.9 reveals that Random Forest outperforms Naive Bayes and Decision Tree algorithms and gives the maximum precision (0.78, 0.81) and recall (0.82, 0.84) for Test-Data1 and Test-Data2. In fact, both Naive Bayes and Random Forest generate almost similar classification results for topic posts with a difference of 1% to 2%. Our results reveal that wrongly classified posts at Stage 1 provoke a decrement in the accuracy of intent classification. As shown in Table 4.9, classification accuracy for Test-Data2 is higher than Test-Data1. Figure 4.15 shows the ROC curves generated for each type of classifiers executed for both Test-Data1 and Test-Data2. Graphs in Figure 4.15 shows that given a set of posts $P = \{P_i \mid 1 \leq i \leq n \mid C(P_i) = Topic\}$, Decision Tree based intent classifier has the high probability ($\sim$ 0.7) to classify them as target class. Whereas, Random Forest and Naive Bayes have almost equal probability (0.55) to classify a post as intent or unknown. Figure 4.15 reveals that if the taxonomy of a post is unknown (Test-Data1), then each algorithm has a probability of approximately 0.60 to classify it as intent post.

In order to evaluate the impact of each feature on intent classification performance, we test the leave-p-out cross-validation for both Test-Data1 and Test-Data2. Figure 4.14 illustrates the percentage of fall in the precision of each classifier and for both the datasets. Negative values rather show the increment in precision. Figure 4.14a shows that in Test-Data1, removing F2 and F3 individually from the feature space does not im-



(a) Decision Tree (b) Naive Bayes (c) Random Forest

(d) Decision Tree (e) Naive Bayes (f) Random Forest

Figure 4.14: Percentage Fall in Accuracy of One-Class Classifiers During Leave-P-Out Compilation (P=1- One Feature (Top), P=2- Two Features (Bottom))

pact the overall performance of Decision Tree (<1%). While removing writing tone feature i.e., F4 decreases the precision by 4%. In fact, for Test-Data2, removing document sentiment vector from the feature space, it increases the performance of Decision Tree by 2%. It is possibly due to the reason is that emotion tone gives a detailed classification of emotions (anger, fear, joy, sadness, and disgust) while document sentiment feature gives overall sentiment of a post that can be biased in longer posts (word length >100). Figure 4.14b reveals that in Naive Bayes algorithm, removal of any feature from Test-Data2 impacts the performance of classifier with a reasonably high percentage of fall in precision. If we remove feature F1 or F4, it decreases the overall precision up to 3%. Similarly, if the taxonomy of a post is unknown (Test-Data1) then removing emotion tone (F3) or language tone (F4) decreases the precision by 2%. Similar to Naive Bayes, for Random Forest algorithm (Figure 4.14c), removal of any feature declines the classifier's performance up to 2%. While, for any unknown post, emotion tone ($F3$) and writing cues of the narrative ($F4$) are the most discriminatory features as removal of these features can decrease the performance of algorithm up to 4%.

We also report the variation in performance of classifiers if a combination of two features is removed from the training model. Leaving out two features at once also reveals the relative influence of each vector in feature space. Figure 4.14d reveals that feature F3 and F4 are the most discriminatory features as removal of any of these vectors does not influence the performance of other features, but we observe a fall in the overall precision rate. For example, removing feature F1 (that increases the precision of Decision Tree algorithm upon leaving out individually) with F4 decreases the precision by 6% for Test-Data1 and 2.25% for Test-Data2. Similarly, leaving features F2 or F3 along with most of other features (F2 and F4) decreases the performances by 1% to 2% for both datasets. However, for Test-Data1, leaving these features out along with F1 rather increases the performance. It reveals that in Decision Tree intent classification, Feature F1

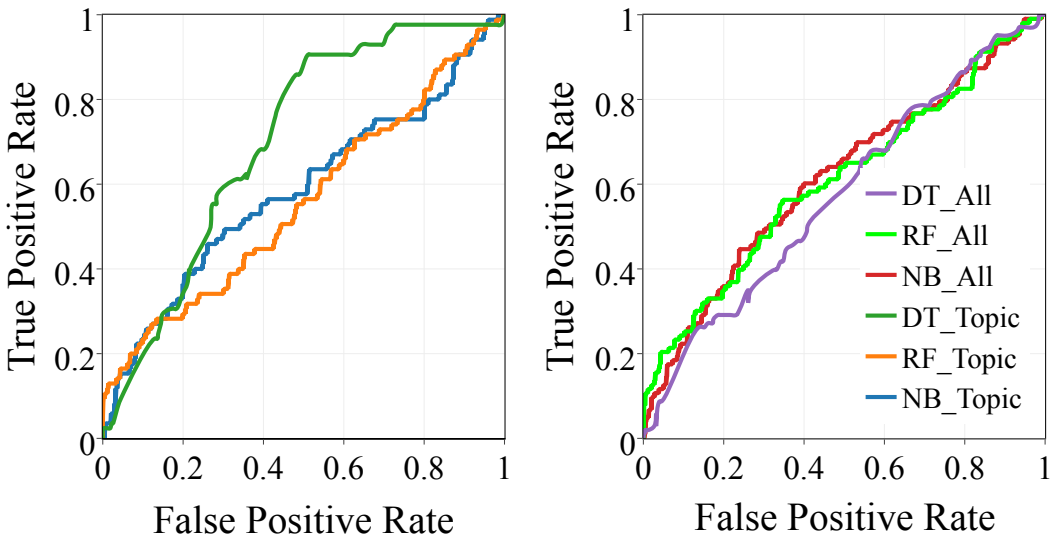


Figure 4.15: ROC Curve for Test-Data1 (Right) and Test-Data2 (Left) Showing the Relationship Between TPR and FPR Values for Decision Tree, Random Forest and Naive Bayes Classifiers.

negatively impacts the performance of other features. In Naive Bayes intent classification, we find that for Test-Data1, F2 is an important feature for identifying intent posts (Figure 4.14e). This is possible because if the taxonomy of a post is unknown, then semantic tagging of text can be an important feature for identifying the topic related posts. Figure 4.14d also reveals that in Naive Bayes classifier (Test-Data1), the social tone of a text (F5) declines the performance of other features. For example, removing F1 individually decreases the precision by 3% while combining it with F5 does not make any change in the accuracy. Similarly, leaving out F3 and F4 features from training model individually makes a fall of 2% in overall performance while combining any of them with F5, the accuracy rather improves by 1% to 2%. It happens because if the posts are not topic related, then they might have a wide range of taxonomy which impacts the social tone of a narrative. Due to the sparsity in social tendency attributes, it increases the number of false alarms. Unlike Decision Tree or Naive Bayes algorithms, in Random Forest, removing a combination of any two feature vectors decreases the performance rate of the intent classifier for each dataset. Figure 4.14f reveals that removing any feature along with F3 declines the precision by at least 4%. While removing them with F4 can lower the performance by 2% to 4%. Our results reveal that emotion tone (F3) and writing cues (F4) are the two most discriminatory features for identifying intent post while using any of three classifiers and datasets. Semantic tagging (F2) and the social tendency of narratives (F5) are two other important features if the post has a wide range of topics or emotional range making a post ambiguous. Classification results support our hypotheses that sentiment and semantic of text can be used to identify the language cues and personality traits of author and classify the intent post on microblogging platforms.

4.5.5 Limitations

In this work, we conduct our analysis only on English-language posts. Our proposed approach has dependencies with the open source APIs used for the feature extraction. If a post contains multi-lingual text (for example, Arabic + English), then the APIs might not be able to extract the taxonomy or semantic features accurately. We make our model generalized, and it can be used to identify racist and radicalized intent for any given text. However, the model might require some pre-processing and large training data for

microposts as the topic modeling, and tone analysis might not be 100% accurate for very short text such as tweets.

4.6 Case Study 3: Mining Hate & Extremism Promoting Users and Communities on YouTube

As discussed in Section 4.1, YouTube has become a convenient platform for many hate and extremist groups to share information and promote their ideologies. The reason because a video is the most usable medium to share views with others [143]. Previous studies show that extremist groups put forth hateful speech, offensive comments and messages focusing their mission [167]. The presence of such extremist users and communities in a large amount is a major concern for YouTube moderators (to uphold the reputation of the website), government and law enforcement agencies (identifying extremist content and user communities to stop such promotion in the country). However, due to the dynamic nature of the website detecting such hate promoting communities on YouTube is a significant and technically challenging problem. Further, the manual inspection for locating such users and groups by the keyword-based search is overwhelmingly impractical. The work presented in this case-study is motivated by the need for a solution to combat and counter online radicalisation. We frame our problem of radicalized community detection as

1. Identification of hate promoting videos leading to the uploaders of such content on the website
2. Investigating the links among such users and locating their virtual and hidden communities on YouTube sharing a common agenda or group mission
3. Identifying users with strong connections and playing central role in a community

4.6.1 Research Framework & Methodology

Figure 4.16 presents a general framework for the proposed solution approach. The proposed method is a multi-step process primarily consists of three phases, Training Profile Collection, Statistical Model Building and Focused Crawler labeled as Phase 1, 2 and 3 respectively. We perform a manual analysis, and a visual inspection on activity feeds and contextual metadata of various YouTube channels. We collect 35 positive class channels (promoting hate and extremism) used as training profiles. We build our training dataset by extracting the discriminatory features (user activity feeds- titles of videos uploaded, shared, favorited & commented by the user and profile information) of these 35 channels using YouTube API[20]. In the training dataset, we observe several terms relevant to hate and extremism and divide them into 9 main categories shown in Table 4.10. We build a statistical model from these training profiles by applying character n-gram based language modeling approach. We chose character-level analysis (low-level features) as it is language independent and does not require extensive language specific pre-processing. The other advantage of character n-gram based approach is that it can capture sub-word and super-word features and is suitable for noisy text found in social media. The paper by Peng et al. lists the advantages of character-level n-gram language models for language independent text categorization tasks [168]. In phase 3, we build a focused crawler (best first search and shark search) which is a recursive process. It takes one YouTube channel as a seed (a positive class channel) and extracts its contextual metadata (user activity feeds and profile information) using YouTube API. We find the extent of textual similarity between these metadata and training data by using statistical model (build in phase 2) and LingPipe API [21]. We implement a binary classifier to classify a user channel as relevant or irrelevant. A user channel is said to be relevant (hate

[20] `https://developers.google.com/youtube/getting_started`
[21] `http://alias-i.com/lingpipe/index.html`

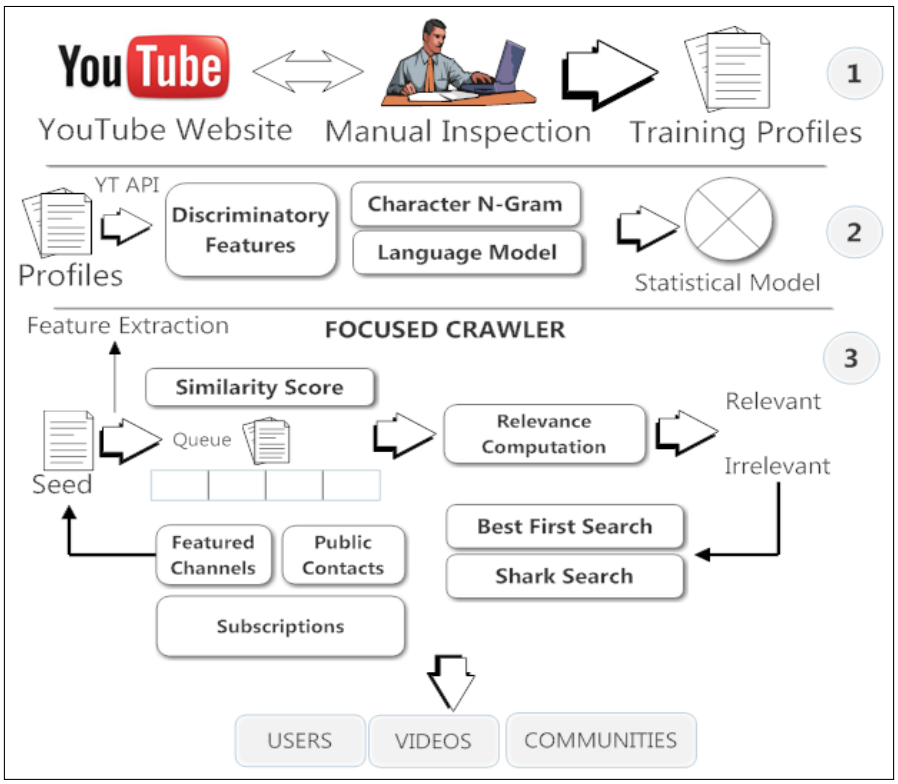


Figure 4.16: A General Research Framework of the Proposed Solution Approach for Identifying Hate and Extremist Promoting Videos and Users on YouTube Website. Further Illustrating the High-Level Demonstration of the Application of Focused Crawler Approach for Locating Radical Groups.

and extremism promoting channel) if the computation score is above a predicted threshold. If a channel is relevant, then we further extend it's frontiers (links to other YouTube channels) i.e. the subscribers of the channel, featured channels suggested by the user and it contacts available publicly. We extract these frontiers by parsing users' YouTube homepage using jsoup HTML parser library[22]. We execute focused crawler phase for each frontier recursively which results in a connected graph, where nodes represent the user channels and edges represent the links between two users. We perform social network analysis on the output graph to locate hidden communities of hate promoting users.

4.6.2 Solution Implementation

In this subsection, we present the methodology and solution implementation details for the design and architecture articulated in the previous subsection. In focused crawler, we first classify a seed input as relevant or irrelevant which further leads to more relevant channels. In proposed method, we use focused crawler for two different graph traversing algorithms i.e. Best First Search (BFS) Algorithm and Shark-Search Algorithm (SSA). Algorithm 6 and Algorithm 7 describe the focused crawlers we develop to locate a group of connected hate and extremist channels on YouTube. The result of both algorithms is turned

[22]`http://jsoup.org/apidocs/`

Table 4.10: Categorization of Sample Terms Occurring in Exemplary Documents for Focused Crawler.

Category	Terms
Important Dates	13th January, 26th January, 23rd March, 5th August, 14th August, 15th August, 21st September, 9th November, 3rd December, 25th December
Region	Hindustan, Pakistan, India, Kashmir, Bhindustan, Lahore, Afganistan, America, China, Turkey, Mumbai, Khalistan, Indo-Pak, US, Jammu & Kashmir, Agartala, Bangladesh, England, Israel, Karbala, Arabia, Argentine, Syria, Egyptian, Goa, Orrisa, Bihar, Canadian, Arab, Sindh, Balochistan, Punjab
Religion	Islam, Muslim, Hindu, Allah, Khuda, Quran, Maulana, Mosque, Kabba, Jihad, Azan, Jewish, Burka, Prophet, Religious, Koum, Islamic, Jews Christians, Apostates, Sikh, Buddhist, Hinduism, Muhajirs, Immigrant Muslims
People Name	Obama, Osama, Laden, Zaid Hamid, Zakir Naik, Parvez Musharraf, Mark Glenn, Jinnah, Saed Singh, Imran Khan, Nawaz Shareef, Quaid, Iqbal, Tahir Ashrafi, Emad Khalid, Yousuf Ali, Shaykh Feiz, Mustafa Kamal, Khalid Yaseen, Asma Jahangir, Chandragupt, Gandhi, Nehru, Pramod Mahajan
Negative Emotions	Horrible, Hate, Hatred, Murder, Cheating, Ice-Blood, Honour, Loathing, Humanity, Violence, Bloody, Blood, Revenge, Torture, Extremism, Humiliation, Abuse, Poverty, Fear, Scoundrel, Lies, Fraud, Friendship, Hesitation, Fake, Filthy, Discrimination
Communities	Paki Punjabi, CIA, ISI, Takmel-E-Pakistan, Brass Tacks, Azad Kashmir, Liberate Kashmir, Taliban, Aman Ki Asha, Flag Attack, Gang, IAF, Air Force, RAW, PMLN, NATO, TTP, Threek-E-Taliban, SWAT, WUP, PPP, Pakistani People Party, Operation Shudhi Karan, Aryavrat
Politics Terms	Conspiracy, Leader, Democracy, Inqalab, Awami, Strike, Khilafat, Against, Rights, Partition, Corruption, Media, Resolution, Objective, Rule, Party, League, Protest, Politician, Slogan, Division, Public, President, Secularism, Domestic, Congress, Election, Witnessed, Tribal, Rallies, persecuted, Youth
War Related Terms	LOC, Bomb, Blast, Attack, Holywar, Warfare, Tribute, Soldier, Jawan, Refugee, Enemies, Fighting, Patriot, Assassination, Expose, Propaganda, Army, Protocol, Security, Anthem, Threat, Nukes, Border, Shaheedi, Military, Zindabad, Hijab, Dirty War, Black Day, Terror, Mission, Operation, Jail, Prison, Open Fire, Destruction
Others	Pig, Monkey, Faith, Ideology, Earthquake, Thunder, Uneducated, Awareness, Debate, Foreign, Leaked, Press, Affair, Economic, Destiny, Flood, Endgame, Rebuttal, Documentary, Respect, Argue, Patrol, Scandal, Survival, Rapist, Rape, Ideological, Geographical, subsections, Sects, Government, Interview

out to be a directed cyclic graph where each node represents a user channel, and an edge represents a link between two users. The goal of BFS and SSA is to classify a channel to be relevant (target class) or irrelevant (unknown class) and then exploring the frontier channels of a relevant user (in case of BFS) and both users (in case of SSA). Inputs to these algorithms are a seed channel (a positive class user) U, width of graph w (i.e. maximum number of children of a node), size of graph s (i.e. maximum number of nodes in graph), threshold th for classification, n-gram value Ng for similarity computation, and a lexicon of 35 positive class channels U_p. Figure 4.17 shows a list of all seed inputs we have used for different iterations. We compare each training profile with all profiles and compute their similarity score for each mode. We take an average of these 35 scores and compute the threshold values. Both algorithms are different in their approach explained in following subsections:

Algorithm 6: Focused Crawler- Best First Search

Data: Seed User U, Width of Graph w, Size of Graph s, Threshold th, N-gram Ng, Positive Class Channels U_p
Result: A connected directed cyclic graph, Nodes=User u

1 **for** ***all*** $u \in U_p$ **do**
2 | $D.add$(ExtractFeatures(u))
Algorithm BFS(U)
3 | **while** $graphsize < s$ **do**
4 | | userfeeds $U_f \leftarrow$ExtractFeatures(U)
5 | | score $score \leftarrow$LanguageModeling(D, U_f, N_g)
6 | | **if** *(score <th)* **then**
7 | | | $U.class \leftarrow$Irrelevant
| | **else**
8 | | | $U.class \leftarrow$Relevant
9 | | | Hashmap $U_{sorted}.InsertionSort(U, score)$
| | **for** $i \leftarrow 1$ ***to*** w **do**
10 | | | Hashmap $U_{graph}.add(U_{sorted}(i))$
11 | | **for** ***all*** $U_g \in U_{graph}$ **do**
12 | | | $fr = Extract\_Frontiers(U_g)$
13 | | | Hashmap $U_{crawler}.add(fr)$
14 | | **for** ***all*** $U_{fr} \in U_{crawler}$ **do**
15 | | | BFS(U_{fr})

4.6.2.1 Focused Crawler- Best First Search

The proposed method (Algorithm 6) follows the standard best first traversing to explore relevant user to seed input. Best-First Search examines a node in the graph and finds the most promising node among its children to be traversed next [169]. This priority of nodes (users) is decided based upon the extent of similarity with the training profiles. A user with the similarity score above a specified threshold is said to be relevant and allowed to be extended further. If a node is relevant and has the highest priority (similarity score) among all relevant nodes, then we extend it first and explore its links and discard irrelevant nodes. We process each node only once, and if a node appears again, then we only include the connecting edge in the graph. Steps 1 and 2 extract all contextual features for 35 training profiles using Algorithm 8 and build a training dataset. Algorithm SSA is a recursive function which takes U as a seed input. Steps 4 and 5 extract all features for seed user U and compute its similarity score with training profiles using character n-gram and language modeling (using LingPipe API). Steps 6 to 8 represent the classification procedure and labeling of users as relevant or irrelevant depending upon the threshold measures. BFS method has non-binary priority values assigned to each node. The priority values are the similarity score, which is computed by comparing the users' contextual metadata (user activity feeds and profile information) with training profiles. Steps 9 and 10 make a list of top w (maximum number of children, a node can have) users among relevant users based upon their similarity score, sorted in decreasing order. Step 16 extracts frontiers of a user channel using Algorithm 9. Steps 18 and 19 repeat steps 3 to 15 for each frontier extracted. We execute this function till we get a graph with the desired number of nodes or there is no more node remaining to extend.

4.6.2.2 Focused Crawler- Shark Search

We propose a focused crawler for Shark-Search Algorithm (Algorithm 7), an adaptive version of the same algorithm introduced in M. Hersovici et. al. [170]. Shark Search algorithm is different from Best First

Algorithm 7: Focused Crawler- Shark Search

Data: Seed User U, Width of Graph w, Size of Graph s, Threshold th, N-gram Ng, Positive Class Channels U_p, Decay Factor d
Result: A connected directed cyclic graph, Nodes=User u

```
1  for all u ∈ U_p do
2  |  D.add(ExtractFeatures(u))
   Algorithm SSA(U)
3  |  while graphsize < s do
4  |  |  userfeeds U_f ←ExtractFeatures(U)
5  |  |  score score ←LanguageModeling(D, U_f, N_g)
6  |  |  if (U is a child of Irrelevant node) then
7  |  |  |  score ← score * d
8  |  |  if (U has appeared before) then
9  |  |  |  score ← max(new_score, old_score)
10 |  |  if (score <th) then
11 |  |  |  U.newclass ←Irrelevant
   |  |  else
12 |  |  |  U.newclass ←Relevant
13 |  |  Hashmap U_sorted.InsertionSort(U, score)
   |  |  for i ← 1 to w do
14 |  |  |  Hashmap U_graph.add(U_sorted(i))
15 |  |  for all U_g ∈ U_graph do
16 |  |  |  fr = Extract_Frontiers(U_g)
17 |  |  |  Hashmap U_crawler.add(fr)
18 |  |  for all U_fr ∈ U_crawler do
19 |  |  |  SSA(U_fr)
```

Search algorithm in a way that it explores frontiers of both relevant and irrelevant nodes. In SSA if the parent of a node is an irrelevant node then the inherited score of the child node is $score_{child} * d$, where d is a decay factor, an extra input for SSA which directly impacts on the priority of user. This inherited score is dynamic because a node can have more than one parent. Steps 1 to 5 are similar to Best First Search (Algorithm 6). Steps 6 to 9 check if the user is a child of an irrelevant node then it computes an inherited score for the user by multiplying the original score by a decay factor d. If a node has appeared before and has not been extended further, then we update its similarity score by the maximum value of old and new inherited score. Steps 10 to 12 represent the classification procedure and labelling of users as relevant or irrelevant similar to Algorithm 6. The SSA method also uses non-binary priority values same as similarity score of users. Steps 13 and 14 make a list of top w (maximum number of children, a node can have) users (could be relevant or irrelevant unlike BFS) based upon their similarity score, sorted in decreasing order. Steps 15 to 19 extract frontiers of a user channel using Algorithm 9 and repeats steps 3 to 19 for each linked user.

4.6.2.3 Features Extraction

In Algorithm 8, we retrieve contextual metadata of a YouTube user channel using YouTube API. Step 1 extracts the profile summary of the user. Steps 2 to 5 extract the titles of videos uploaded, commented, shared and favorited by given user U. The result of the algorithm is a text file containing all the video titles and user profile information.

Algorithm 8: Features Extraction for a YouTube User

Data: User u
Result: User Activity Feeds and Profile Information
Algorithm *ExtractFeatures*(*U*)

1 $u_{Profile}$ ←u.getSummary()
2 $u_{Uploads}$ ←u.getUploadedVideo()
3 $u_{Commented}$ ←u.getCommentedVideo()
4 u_{Shared} ←u.getSharedVideo()
5 $u_{favorited}$ ←u.getFavoritedVideo()

Algorithm 9: Frontier Extraction for a YouTube User

Data: User u
Result: Frontiers of a channel
Algorithm *Extract_Frontiers*(*U*)

1 u_{subs} ←u.getSubscribers()
2 u_{fc} ←u.getFeaturedChannels()
3 u_{con} ←u.getFriends()

Figure 4.17: Name of 10 Seed Inputs Used for BFS and SSA Graph Traversal and Focused Crawler

Figure 4.18: YouTube Channels of Uploaders of Hate and Extremism Promoting Videos Being Used As Exemplary Documents For Training A Text Classifier

4.6.2.4 Frontiers Extraction

In Algorithm 9, we extract all external links of a YouTube channel to other YouTube channels. These links could be the subscribers, featured channels (suggestions by the user) and public contacts (friends). YouTube API does not allow users to retrieve the contacts of other users which is why we use jsoup HTML parser library to fetch all frontiers and public contacts list. This algorithm returns a vector of all channels user U is linked with, and we make sure that there is no redundant channel in the list.

4.6.3 Experimental Setup and Dataset

A focused crawler needs to classify if a given webpage is relevant to the given topic or not. The crawler requires exemplary documents or training examples to learn the specific characteristics and properties of documents in the training dataset. A statistical model (text classifier) needs to be built from a collection of documents pertaining to a predefined topic. Figure 4.18 shows a list of 35 user ids used as training profiles. The 35 user ids consist of 612 videos, and hence the training is performed on 612 videos. We obtain the training dataset by manually searching (keyword based) for anti-India hate and extremism promoting channels using YouTube search and traversing related video links (using the heuristic that videos on a similar topic will be connected as relevant on YouTube). The training dataset profile consists of profile information of users and the title of videos uploaded, favorited, shared and commented by the user. We believe the title of such videos reflects user interests and can be used for building a predictive model. We select 10 random positive class (hate and extremist) channels for creating test dataset. Each user works as a seed input to the focused crawler. Figure 4.17 shows the list of all 10 seeds we select for our experiments. To evaluate the effectiveness of our solution approach, we execute our focused crawler sixty times for both Shark Search and Best First Search. Here we use 10 different seeds, 3 different threshold values and 2 different n-gram values for similarity computation. We make 6 pairs of threshold and n-gram values calling them as six different "Modes". For both approaches (BFS and SSA), we run our focused crawler 60 times for 10 seeds and each seed for all 6 modes.

4.6.4 Experimental Results

In this subsection, we present the characterization of hate and extremist videos. We demonstrate the empirical analysis and performance results. To measure the effectiveness of our focused crawler, we used standard confusion matrix [171]. Each user can only be classified in one of the classes: Positive (Relevant) and Negative (Irrelevant). We evaluate the accuracy of our classifier by comparing the predicted class (each column in the matrix) against the actual class (each row in the matrix) of each user channel. We evaluate the performance of our classifier using TPR (or recall), TNR (or specificity), PPV (or precision), NPV, F1-score, and accuracy. Precision is the number of positives predicted correctly. TNR is the proportion of actual negative, which is predicted negative. The recall is the number of actual positives, which are predicted correctly. NPV is the number of negatives predicted correctly. F1-Score is the weighted harmonic mean between Precision and Recall. Accuracy is the performance of the binary classifier to predict true results (both true positives and true negatives).

4.6.4.1 Focused Crawler Results

As mentioned above, we execute our focused crawler 60 times for both BFS and SSA. Table 4.11 (a) and (b) present the complete picture of users statistics based upon their similarity scores in each iteration. Table 4.11 (a) and (b) show the number of unique relevant, irrelevant users, the total number of users present in the output graph and the total number of users processed during execution of BFS and SSA focused crawlers respectively. In Table 4.11 we notice that for both BFS and SSA, five-gram performs better than tri-gram. And for five-gram we achieve maximum number of relevant users in mode F (threshold= -3.0, n-gram= 5). These statistics show that the number of relevant and irrelevant nodes vary for different seeds. For example, for seed 3 and 8, we have only one relevant node. Despite being positive class channels, these users have no links to other hate and extremist users on YouTube. Table 4.11 (a) and (b) reveal the difference in BFS and SSA performance for same seed. For seed 7, 9 and 10, we have an empty graph for BFS while in SSA we have 25 connected users for mode A. Similarly, for other modes, SSA has higher number of relevant users in comparison to BFS.

Figure 4.19a and 4.19b illustrate the variance in number of nodes (shown on Y-axis) for different modes

Table 4.11: Results of Focused Crawler for 6 Different Seeds. Modes Represent 6 Different Thresholds (Th) & N-gram (Ng) Pairs. (A: Th=-2.0, Ng=3, B: Th=-2.5, Ng=3, C: Th=-3.0, Ng=3, D: Th=-2.0, Ng=5, E: Th=-2.5, Ng=5, F: Th=-3.0, Ng=5), (C1: Mode, C2: Relevant, C3: Irrelevant, C4: Processed Nodes, C5: Nodes in the Graph.)

(a) Focused Crawler- Best First Search

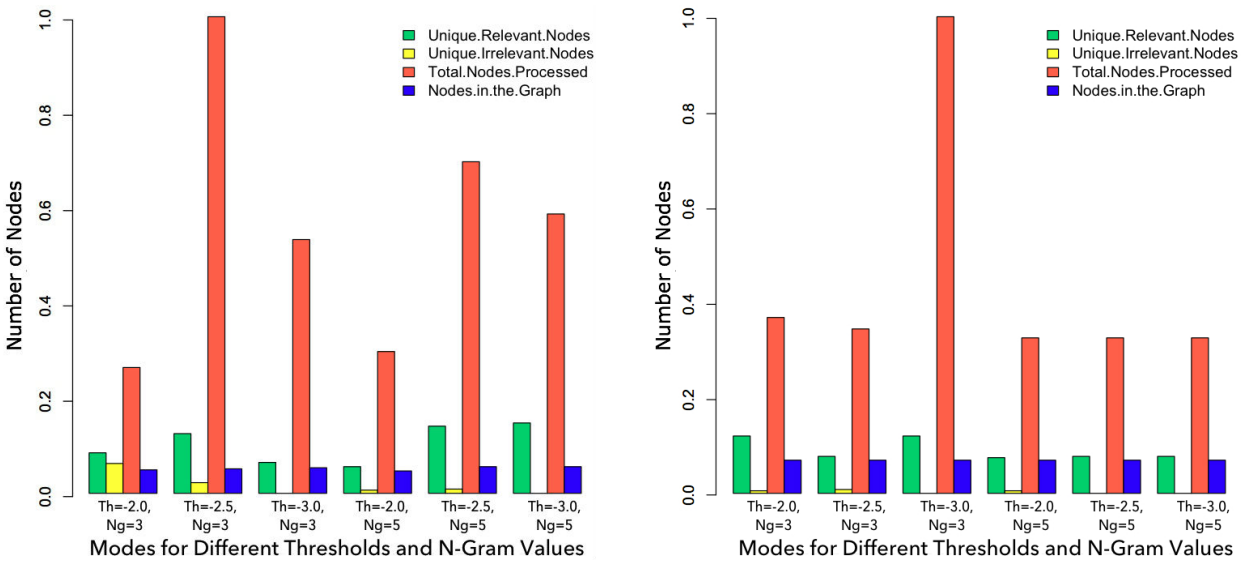
	Seed 1						Seed 2						Seed 3						Seed 4						Seed 5					
C1	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
C2	26	19	58	21	56	60	39	57	30	26	64	67	1	1	1	1	1	1	40	54	31	27	40	74	1	1	1	1	1	1
C3	23	2	9	3	11	7	11	1	4	5	1	1	0	0	0	0	0	0	29	11	1	4	1	2	1	1	1	1	1	1
C4	119	448	239	134	159	145	119	448	239	134	312	263	1	1	1	1	1	1	129	325	212	203	309	403	1	1	1	1	1	1
C5	23	19	25	21	26	26	23	24	25	22	26	26	1	1	1	1	1	1	23	23	25	22	25	25	1	1	1	1	1	1
	Seed 6						**Seed 7**						**Seed 8**						**Seed 9**						**Seed 10**					
C1	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
C2	5	34	27	23	32	32	0	28	25	21	31	31	1	1	1	1	1	1	0	30	19	27	33	33	0	0	18	17	22	22
C3	2	4	10	11	6	6	1	5	1	4	2	2	0	0	0	0	0	0	1	4	1	2	2	2	1	1	1	2	3	3
C4	20	313	290	258	332	332	0	212	318	256	252	274	1	1	1	1	1	1	0	0	217	321	385	385	0	237	218	189	254	254
C5	5	22	25	22	25	25	0	22	25	21	26	26	1	1	1	1	1	1	0	22	19	24	23	23	0	0	18	17	22	22

(b) Focused Crawler- Shark Search

	Seed 1						Seed 2						Seed 3						Seed 4						Seed 5					
C1	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
C2	37	34	27	56	29	27	45	29	45	28	29	29	1	1	1	1	1	1	45	32	32	29	76	35	1	1	1	27	40	74
C3	6	2	2	3	3	2	2	3	0	2	0	0	0	0	0	0	0	0	2	3	0	2	0	0	1	1	1	1	1	1
C4	198	167	123	177	110	110	138	129	374	122	122	122	1	1	1	1	1	1	147	198	199	157	221	221	1	1	1	1	1	1
C5	26	26	26	26	26	26	25	26	26	26	26	26	1	1	1	1	1	1	25	25	25	25	25	25	1	1	1	1	1	1
	Seed 6						**Seed 7**						**Seed 8**						**Seed 9**						**Seed 10**					
C1	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
C2	23	35	27	37	23	23	22	36	26	36	23	23	1	1	1	1	1	1	17	33	23	27	40	40	14	18	19	23	30	30
C3	4	3	0	0	2	2	2	3	0	3	2	2	0	0	0	0	0	0	3	3	0	1	2	2	1	1	0	1	2	3
C4	158	169	88	131	80	80	242	213	65	107	54	54	1	1	1	1	1	1	47	195	50	127	336	359	41	41	41	49	124	124
C5	25	25	25	25	25	25	25	25	25	25	21	21	1	1	1	1	1	1	25	25	25	25	25	25	25	25	25	25	25	25

(shown on X-axis) for one seed, where each node represents a YouTube user. Figure 4.19b depicts that for each mode number of irrelevant nodes for SSA are negligible in comparison to BFS. We also notice that for Seed 2, the graph size is almost similar in both BFS and SSA approach. In BFS we extract frontiers of only the relevant nodes, unlike SSA. Therefore, for BFS, we see a radical change in the number of processed nodes for each mode. For SSA, the number of relevant as well as the processed nodes are similar for all modes except mode C. Figure 4.20 and 4.21 show the variance in the statistics of similarity or relevance score (shown on Y-axis) for different modes (shown on the x-axis). These statistics are measured for one seed used for both BFS and SSA approaches and same configuration of threshold and n-gram values. In Figure 4.20 we see that the first quartile for mode A is below the threshold value and it is smaller than third quartile unlike in Figure 4.21. It is evidence that for BFS the number of relevant nodes is lesser in comparison to the SSA. In SSA approach, we are able to find channels which are more relevant (shown as outliers) to training profiles. Figure 4.20 and 4.21 show that for modes E and F (Th=-2.5, Ng=5 and Th=3, Ng=5 respectively) all users are classified at relevant.

We asked 3 graduate students of our department to validate our results, and they manually annotated each YouTube channel. Based upon the validation, we evaluate the accuracy of our classifier by comparing the predicted class against the actual class of each user channel. Table 4.12a shows the confusion matrix for binary classification performed during Best First Search approach. Given the input of 10 seed users and six modes (pairs of threshold and n-gram values), we get a different number of connected users in each iteration.



(a) Best First Search (b) Shark Search

Figure 4.19: Illustrating The Variance Between Number of Unique Relevant Nodes, Unique Irrelevant Nodes, Nodes Present in The Graph and Total Number of Nodes Processed for Six Different Modes of Seed 2

Table 4.12: Confusion Matrix for Focused Crawlers- Illustrating the Predicted Results Against Actual Classes of Records

(a) Best First Search

		Predicted	
		Relevant	**Irrelevant**
Actual	**Relevant**	991	295
	Irrelevant	55	29

(b) Shark Search Algorithm

		Predicted	
		Relevant	**Irrelevant**
Actual	**Relevant**	921	314
	Irrelevant	125	67

To measure the accuracy of our proposed approach we collect results of all 60 iterations and classify 1046 (921 + 125) users as relevant and 381 (314 + 67) as irrelevant users. There is a misclassification of 25.42% and 65.10% in predicting the relevant and irrelevant users respectively. Table 4.12b shows the confusion matrix for binary classification during Shark Search approach. Given the input of 10 seed users and 6 n-gram & threshold pairs, it classifies 1046 (991 + 55) users as relevant and 324 (295 + 29) as irrelevant users. There is a misclassification of 22.93% and 65.47% in predicting the relevant and irrelevant users respectively. This misclassification occurs because of the noisy data such as lack of information, non-English text, and misleading information. Table 4.13 shows the accuracy results (precision i.e. PPV, recall i.e. TPR, NPV, TNR, f1-score and accuracy) of the focused crawler for both Best First Search and Shark Search approaches. Table 4.13 reveals that overall SSA approach (accuracy of 74%) performs better than BFS approach (accuracy 69%). Precision and accuracy of SSA are much higher than BFS and similarly recall, and f1- score are reasonably higher for SSA.

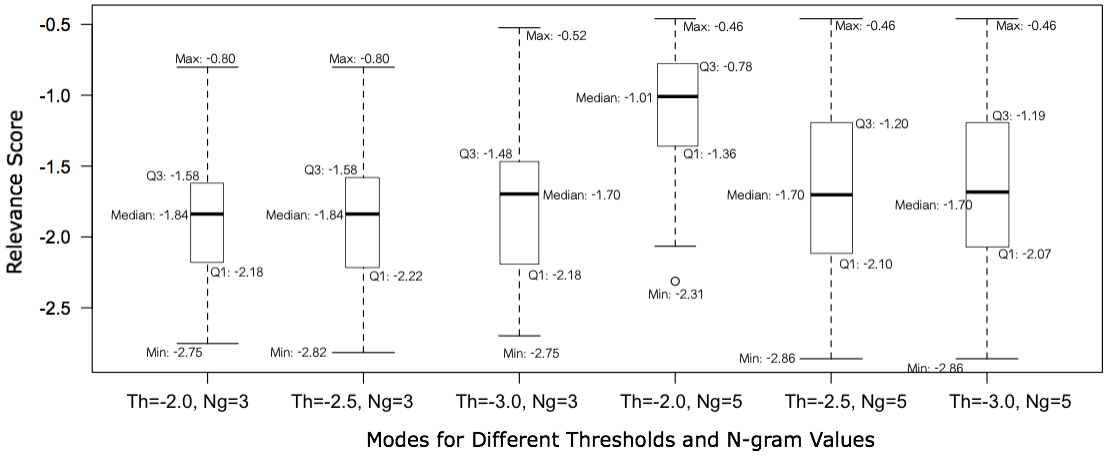


Figure 4.20: Box-Plot and Descriptive Statistics for Six Different Configurations of Best First Search Crawler

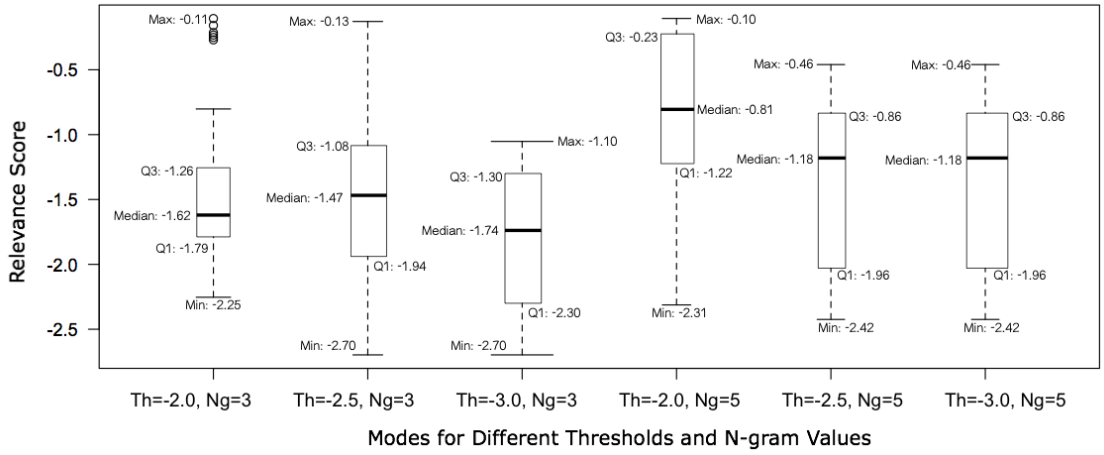


Figure 4.21: Box-Plot and Descriptive Statistics for Six Different Configurations of Shark Search Crawler

4.6.4.2 Social Network Analysis

We perform social network analysis on the output graph of the focused crawler, where each node represents a YouTube user channel and each edge accounts for a relation (friend, subscriber and featured channel) between two users. Table 4.14 illustrate the network level measurements we perform on the output graphs of BFS and SSA focused crawlers. These values have been computed for seed 2 in mode B (configuration of threshold=-2.5 and n-gram=3). In Table 4.14 we notice that in SSA approach users are strongly connected in comparison to BFS approach because the average density of network graph is more in SSA approach. Network diameter shows that in SSA each user is reachable in maximum 3 hops while in BFS it takes 4 hops. In SSA, we have a higher number of connected components than BFS. Therefore we are able to locate a larger number of hate promoting communities in SSA. The average clustering coefficient of SSA is reasonably higher than BFS. Hence the clusters formed in SSA include only highly relevant users.

Table 4.13: Accuracy Results for Focused Crawler- Best First Search and Shark Search. TPR= True Posoitive Rate, FPR= False Positive Rate, PPV= Positive Predictive Value, NPV= Negative Predicted Value.

	TPR	TNR	PPV	NPV	F1-Score	Accuracy
BFS	0.75	0.35	0.88	0.18	0.81	0.69
SSA	0.77	0.35	0.95	0.09	0.85	0.74

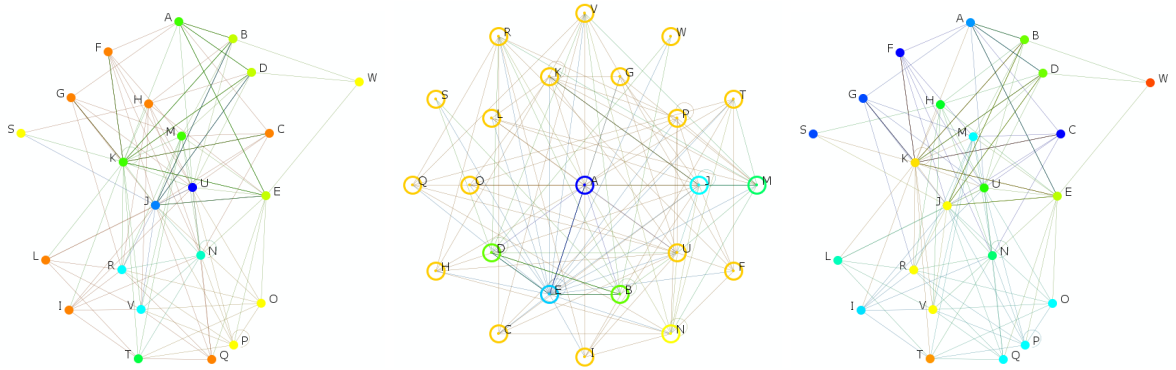
Table 4.14: Illustrating The Network Level Measurements for Focused Crawlers- Best First Search (Left) and Shark Search Algorithm (SSA). NN= Number of Nodes, NE= Number of Edges, SL= Number of Self Loops, Dia= Network Diameter, AD= Average Density, ACC= Average Clustering Coefficient, IBC= In- Betweenness Centrality, CC= In- Closeness Centrality, #W/SCC= Number of Weak/Strong Connected Components.

	NN	NE	SL	Dia	AD	ACC	IBC	ICC	#WCC	#SCC
BFS	23	119	3	4	0.225	0.388	0.046	0.356	1	7
SSA	24	137	8	3	0.238	0.788	0.009	0.320	1	16

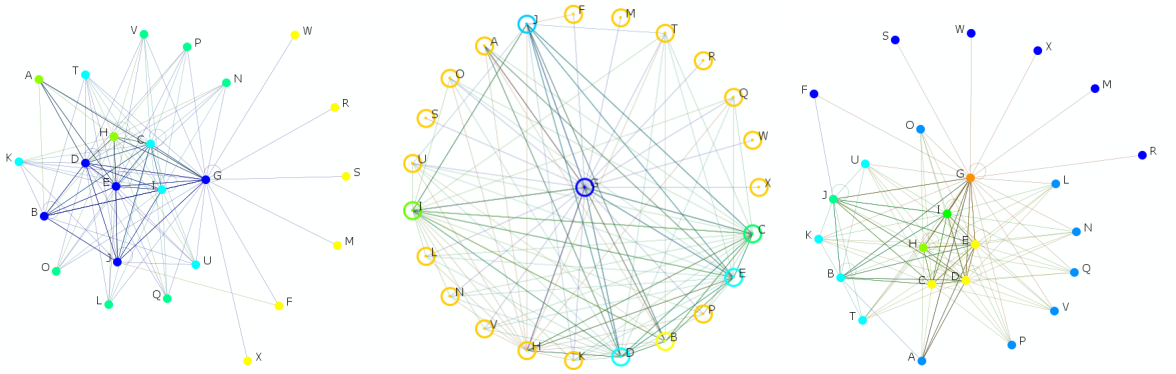
Figures 4.22a and 4.22b shows three different representations of network graphs, outputs for BFS and SSA focused crawler respectively (seed 2 and mode B- threshold=-2.5, n-gram=3). We use ORA[23], a social network visualization tool to create the network of these nodes (user channels). ORA is a dynamic meta-network analysis and evaluation tool developed by CASOS (Center for Computational Analysis of Social and Organizational Systems) group at Carnegie Mellon University. ORA contains a large number of social network analysis, graph and network metrics that compute the relations between input nodes and identifies the patterns in nodes, groups, and network [172].

The network graph on the left shows a directed connected cyclic graph. Colors of nodes represent the different 'in-degree' of users, and the width of an edge is scaled based upon the number of links between two users. In Figures 4.22a and 4.22b node A is the seed user. In community graphs, we see that for BFS, all nodes are connected to each other while in SSA a few nodes are connected to only one user. Despite the existence of these nodes, we find many strongly connected components in SSA which is very less in BFS because all nodes are equally connected. Graphs in the middle in Figures 4.22a and 4.22b are a different representation of the output graph (betweenness centrality). A node in the center has the highest centrality among all users and connected to all users of outer shells. In Figures 4.22a and 4.22b, graphs in the right are the cluster representation of network. As we see in Table 4.14, the average clustering coefficient of the network in SSA approach is very large in comparison to BFS. Similarly in the Figure 4.22a we see that in BFS approach network has 13 clusters, where the total number of nodes is 23. Among these 13 clusters, 6 clusters have only one user node which shows the lack of similarity among users. In Figure 4.22b the cluster representation of network graph (rightmost graph) has 7 clusters for 24 nodes where only 2 clusters are formed with one user channel. In Figure 4.22b network graph, each cluster shows the level of connectivity to other users. We see the existence of three strong communities made by nodes C, D, E, H, I and B, C, D, E, G, H, I, J and A, D, E, G where G is the center of all communities and connected to all users.

[23]http://www.casos.cs.cmu.edu/projects/ora/



(a) Best First Search Approach



(b) Shark Search Approach

Figure 4.22: Community (Left), Betweenness Centrality (Middle) and Cluster (Right) Graph Representation for Best First Search (Top) Shark Search Crawler (Bottom) With Configuration: Th=-2.5, Ng=3 and Seed 2.

4.6.4.3 Manual Analysis of Videos

We perform a manual analysis on YouTube and collect 274 hate and extremist videos uploaded by 35 unique users. We perform a characterization on these 274 and divide them into 5 different sets, shown in Table 4.15. We categorize these videos based upon three main parameters: 1) focus of the content shown in the videos, 2) targeted audience of the users uploading these videos and 3) the keywords presented in the title & description and used or spoken in the video. We also perform a characterization of these videos based upon the content shown in the video. Table 4.15 reveals that the average duration of these videos is from 3 minutes to 45 minutes. We observe that the 43 of total videos were small clips showing women and children harassment in India and Pakistan. For example, child labor, prostitution, slave. We find that majority of videos focus on Islam promotion. These videos are relatively longer in the duration and defined as education videos on YouTube. Table 4.15 also shows that majority of videos fall under news and politics category and very few of them are uploaded for entertainment purpose. These videos target the audience who are affected by the incidents shown in the videos.

For example, 1947 partition, liberate Kashmir and hate speech videos against Pakistan and India targeting the haters of respective nations. The keywords shown in the Table 4.15 are the clear evidence of these videos to be hate promoting. We notice that all these videos are not just public recording, but users have used more creative ways to present their messages in front of their audiences. We divide these 274 videos

Figure 4.23: Categorization of Hate & Extremism Videos Based Upon the Content Shown in the Video.

Table 4.15: Charcaterization of Videos Based Upon Keywords in Video Content & Title and Target Domain of the Uploader, #VD= Number of Videos, YT Category= Youtube Category, Avg Len= Average Duration of Video (in seconds).

#VD	YT Category	Avg Len	Content Focus	Target Audience
43	News & Non-Profit	151.68	Honour Killing, Harassment	Women, Refugee People, Child
93	News, Auto, Vehicle, Politics & Education	2526.16	Islam Promotion	Jewish And Muslim People
25	News, Politics & Education	1225.56	Liberate Kashmir	Kashmiri People
83	News & Politics	349.28	Anti-Muslims	Pakistan Haters
30	Entertainment, Travel, News & Politics	319.61	Anti-India	India Haters

into 12 categories based upon the type of content shown in the video. Figure 4.23 shows that now users have used animation, cartoon, drawings, group discussions and textual messages in their videos to promote hate and extremism. These videos leave a negative impact on the audience and provoke them to write hateful comments.

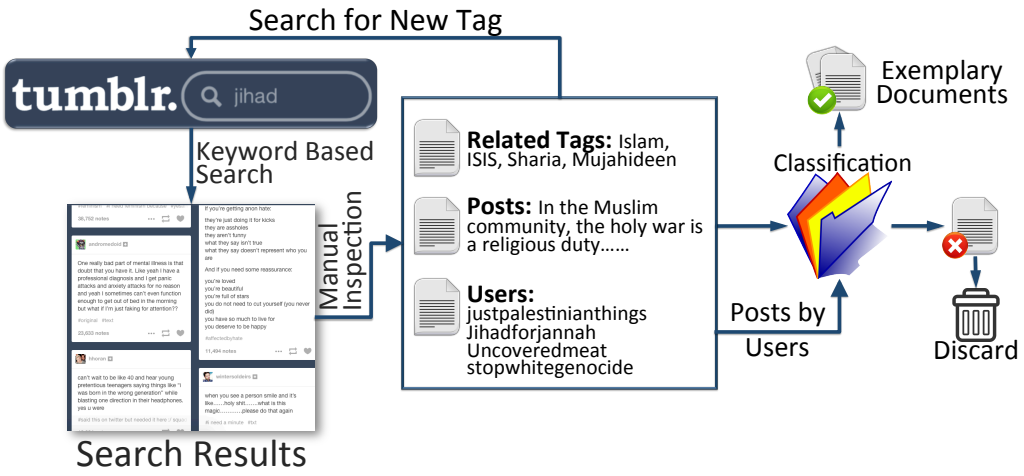


Figure 4.24: Shows High-Level Diagram of Bootstrapping Method used for the Collection of Experimental Dataset and Exemplary Documents

4.7 Case Study 3: Uncovering Hidden Communities of Extremist Micro-Bloggers on Tumblr

As discussed in Chapter 1 and Section 4.1 of this chapter, Tumblr allows users to posts 8 different types of content including images, text, video and audio. The simplicity of navigation, high reachability across a wide range of viewers, low publication barriers, social networking, and anonymity has led Tumblr to be a convenient platform for extremist groups and communities. Due to the dynamic nature of the website and presence of huge amount of text, pictures, and other multimedia content, it is impractical to find every hate promoting post using keyword based flagging search. Further due to the low-quality of user-generated data, the automatic identification of such users and communities is a technically challenging problem. The work presented in this study is motivated by the need of investigating solutions to counter and combat the online extremism on Tumblr. The specific research aim of the work presented in this case-study is the following:

1. To investigate the application of topical crawling based algorithm for retrieving hate promoting bloggers on Tumblr. Our aim is to examine the effectiveness of random walk in social network graph traversal and measuring its performance.
2. To investigate the effectiveness of contextual metadata such as the content of the body, tags and caption or title of a post for computing the similarity between nodes in graph traversal. To examine the effectiveness of *re-blogging* and *like* on a post as the links between two bloggers.
3. To conduct experiments on a large real-world dataset and demonstrate the effectiveness of proposed approach for locating virtual and hidden communities of hate and extremism promoting bloggers and apply Social Network Analysis based techniques to locate central and influential users.

4.7.1 Experimental Setup

In a graph traversal, a topical crawler returns relevant nodes to a specific topic. To define the relevance of a node, it learns the characteristics and features of given topic and computes the extent of similarity

Figure 4.25: A Word Cloud of Key Terms Commonly Used by Extremist Bloggers

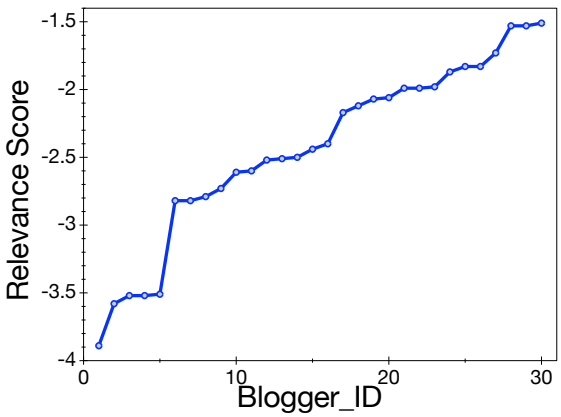


Figure 4.26: Shows Relevance Score Statistics of Positive Class Bloggers

against a bunch of exemplary documents. Figure 4.24 illustrates the general framework to obtain these documents. As shown in Figure 4.24, to collect these training examples, we perform an iterative search on Tumblr using keyword based flagging, where the keyword is a search tag; for example, jihad, anti-islam and hate. We perform a case study on Jihad and by a manual search on Tumblr posts, we collect several relevant tags that are commonly used by extremist bloggers. We use these tags to initiate our process and collect all textual posts (avoiding picture, audio, video, and URLs), tags (associated with resultant posts) and linked bloggers (post reblogged by and liked by) with no redundancy. We perform a manual inspection on resultant posts and posts made by linked bloggers to filter relevant (hate promoting) and unknown results. We further extract more posts and linked bloggers from related tags and run this framework recursively to collect our exemplary documents. These training examples contain the body and caption of only positive class (hate and extremism promoting content) posts which are used to train the model. During the manual inspection of exemplary documents and resultant posts, we observe many keywords that are frequently used by extremist bloggers. Figure 4.25 shows a word cloud of such terms. We use these linked bloggers and posts to compute the threshold value for language modeling. We take a sample of 30 bloggers and compare their posts with the exemplary documents. For each blogger, we get a relevance score. To compute the threshold value for similarity computation, we take an average of these scores. Figure 4.26 illustrates the relevance score statistics of each blogger (Sorted in increasing order). We notice that 80% of the bloggers have relevance score between -2.7 and -1.5. We take the average (turns out to be -2.58) of these scores to avoid the under-fitting and over-fitting of bloggers during classification.

4.7.2 Research Methodology

In this Subsection, we present the general research framework and methodology of proposed approach for classifying extremist bloggers on Tumblr. Figure 4.27 illustrates the design and architecture of topical crawler to locate radical communities. As shown in Figure 4.27, solution framework is an iterative multi-step process primarily consists of five phases: features (posts) extraction, data pre-processing, classification, frontier extraction, and graph traversal. In phase 1, we initiate our process using a positive class (hate promoting) blogger U_i called as 'seed'. We use Tumblr API[24] to fetch the URLs of n number of textual posts and by using Jsoup Java library[25] we extract the content and caption of these posts (used as contextual

[24] `https://www.tumblr.com/docs/en/api/v2`
[25] `http://jsoup.org/apidocs/`

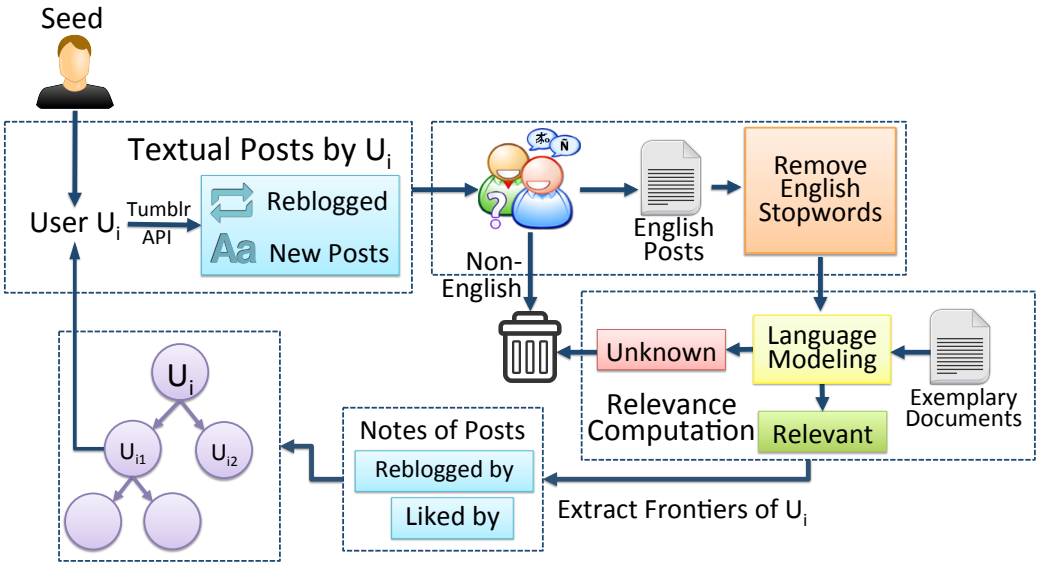


Figure 4.27: A High-Level Block Diagram Presentation of Proposed Methodology for Locating Extremist Communities on Tumblr Primarily Divided in 5 Phases: 1) activity feeds extraction, filtering non-English language posts, classification, external links extraction, and graph traversal.

metadata). These posts can be either re-blogged from other users or originally posted by the user U_i. These posts consist of multiple languages. Therefore, in phase 2, we perform data pre-processing and filter English and non-English posts using language detection library[26]. We perform data pre-processing on these posts and remove English stopwords. In phase 3, we build a statistical model from the exemplary documents collected separately by using a semi-automatic process (refer to Figure 4.24). To compute the relevance of each blogger, we use the character-level n-gram language modeling approach. We find the extent of similarity between metadata and exemplary documents using LingPipe API[27]. We implement a one-class classifier and filter extremism promoting bloggers from unknown bloggers.

In phase 4, we extract the notes associated with the posts (collected in phase 1) of relevant bloggers. These notes contain the list of bloggers who liked and re-blogged a particular post. The number of notes represent the popularity of a post and indicate the similar interest between original poster and other bloggers on the list who may or may not be the direct followers of each other. We use notes to extract frontier nodes of a blogger because of two reasons:

1. Due to the privacy policies, the Tumblr API does not allow developers to extract the followers and followings blogs of Tumblr users.
2. Tumblr facilitates bloggers to track any number of tags so that whenever there is a new post published publicly on Tumblr containing any of those tags, it automatically appears in a menu on user's dashboard. Bloggers can spread that post among their followers by reblogging it. Tracked tags allow bloggers to form a virtual community without following each other.

For each frontier extracted in phase 4, we compute the relevance score against exemplary documents and discard unknown bloggers. In phase 5, we manage a queue of relevant bloggers and perform directed graph

[26] `https://code.google.com/p/language-detection/`
[27] `http://alias-i.com/lingpipe/index.html`

Algorithm 10: Extracting Textual Posts on Tumblr

```
Data: User U, Consumer Key C_k, Consumer Secret C_s, Search Tag tag_name
Result: Text based posts made by User U or associated with tag tag_name
Algorithm ExtractPost()
1   | SetParameters()
2   | select a method to extract posts TaggedPost() OR BloggerPost()
3   | Generate URL of post to fetch post content and caption
4   | for all postP ∈ Posts do
5   |   | Slug=P.getSlug()
6   |   | id=P.getID()
7   |   | URL="http://blog_name.tumblr.com/post/id/slug"
8   |   | Document=Jsoup.connect(URL).get()
9   |   | post_content=Document.getDescription()
10  |   | post_caption=Document.getTitle()
Algorithm SetParameters()
11  | Authenticate the client via API Keys C_k and C_s
12  | Set the parameters
13  | params.put("type", "text")
14  | params.put("filter", "text")
15  | params.put("reblog info", true)
16  | params.put("notes info", true)
Algorithm TaggedPost(tag_name)
17  | Posts = client.tagged(tag_name, params)
Algorithm BloggerPost(blog_name)
18  | Posts = client.tagged(blog_name, params)
```

traversal using random walk algorithm. To expand our graph, we select the next blogger in uniform distribution and extract its frontiers. We execute our focused crawler for each frontier without revisiting a blogger. This traversal results in a connected graph, where nodes represent a blogger (hate promoting) and edges represent the links (re-blog and like) between two bloggers. We perform social network analysis on the resultant graph and locate extreme right communities of hate promoting bloggers.

4.7.3 Solution Implementation

A topical crawler starts from a seed node, traverses in a graph navigating through some links and returns all relevant nodes to a given topic. In proposed solution approach we divide our problem into three subproblems. First, we classify the given seed node S as hate promoting or unknown according to the published post (originally posted by blogger or re-blogged from other Tumblr users). Second, if the node is relevant then we extend this node into its frontiers, and it further leads us to more hate promoting bloggers. In third sub-problem, we perform topical crawling on Tumblr network and use random walk algorithm to traverse along the graph.

4.7.3.1 Retrieval of Published Posts

Algorithm 10 describes the method to search Tumblr posts using keyword based flagging and extraction of posts published by a given blogger. The work presented in this study focuses on mining textual metadata on Tumblr. Therefore, we set a few parameters and extract only text-based posts for further analysis. For

Algorithm 11: Extracting Frontiers of a Given Blog

```
   Data: Blogger U
   Result: Frontiers F < name, type > of U
   Algorithm ExtractFrontiers(U)
1  |  SetParameters()
2  |  Posts=BloggerPost(U)
3  |  for all postP ∈ Posts do
4  |  |  Notes=P.getNotes()
5  |  |  for all NoteN ∈ Notes do
6  |  |  |  Linked_Blog_Name=N.getBlogName()
7  |  |  |  Note_Type=N.getType()      Liked or Reblog
8  |  |  |  Frontiers F.add(Linked_Blog_Name, Note_Type)
```

each blogger, we set the limit of 100 posts published recently. Function `SetParameters()` (steps 13 and 14) filters the search results and displays only the textual posts. Function `TaggedPost()` with given parameters searches for text posts that exclusively contain given tag name. `BloggerPost()` fetches the textual posts published by given blogger name. Both the functions makes a Tumblr API request to fetch these data. Function `ExtractPost()` filters the response and extract body content & caption of each post. Tumblr API allows us only to extract the summary of large posts. Therefore we use HTML parsing for obtaining the whole message in a blog post. In steps 4 to 7, we generate the URL from post summary and id to fetch the remaining post details. ID is a unique identifier of Tumblr posts, and slug is a short text summary of that post which is appended at the end of every URL. We invoke this URL using Jsoup library and parse the HTML document to get the post content.

4.7.3.2 Retrieval of External Links to Bloggers

Algorithm 11 describes the steps to extract frontiers of a given node U. Due to the privacy policies, Tumblr API does not allow developers to extract subscriptions and followers of a Tumblr user. The link between two bloggers indicates the similar interest between them, so that number of frontiers vary for every post published by a blogger. For each user, we extract 25 bloggers for each relation i.e. users who have liked and reblogged that post recently. If a blogger B_1 re-blog and as well as likes a post published by another blogger B_2 then in the graph G, we create an edge with both labels i.e. (B_1, B_2, <like, re-blog>). To avoid the redundancy, we extract one more frontier who have either liked or re-blogged the post recently. To extract the linked bloggers of a Tumblr user we first need to extract the posts made by U. We can extract notes information only when notes and re-blog information parameters are set to be true (refer to steps 15 and 16 of Algorithm 10). As described in Algorithm 11, in step 4, we extract notes for each textual post (hate promoting) made by User U. In steps 5 to 8, we extract the name of unique bloggers who liked and re-blogged the post P. In step 8, F represents the list of frontiers and relation of U with each frontier. We maintain a list of all processed bloggers and the number of hit counts on their recent 100 posts. These number of notes varied from 0 to 25K, therefore, we perform smoothing on data points and plot median of these values. Figure 4.28 shows the statistics of the number of notes collected on 100 posts of each blogger extracted during topical crawler. Figure 4.28 reveals that the overall number of hit counts (number of reblogging and likes) for extremism promoting users is very high. These hit counts reveal the popularity of extremist content and the number of viewers connected to such bloggers.

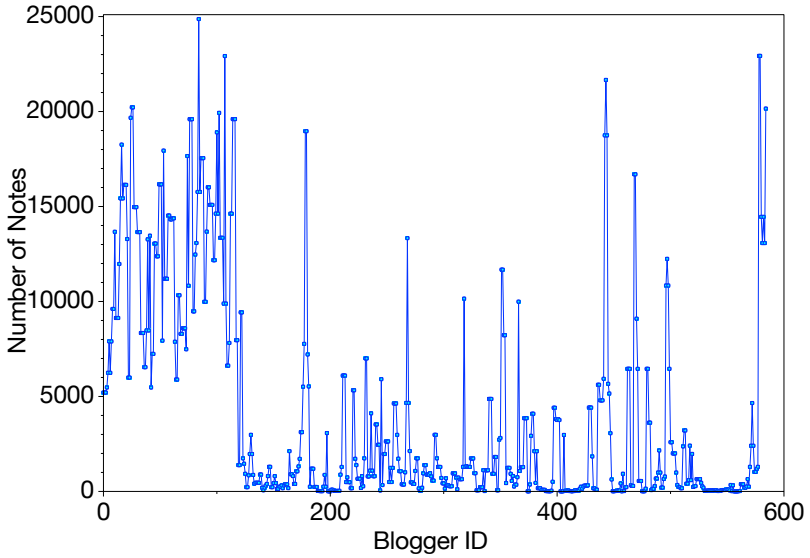


Figure 4.28: Illustrating the Number of Notes For Each Blogger Processed in Random Walk Graph Traversal Based Topical Crawler

4.7.3.3 Topical Crawler Using Random Walk

Algorithm 12 describes the proposed crawler for locating a group of hidden extremist bloggers on Tumblr. The goal of this algorithm is to compare each blogger against training examples and then to connect all positive class bloggers (hate promoting or relevant). Algorithm 12 takes several inputs: seed blogger (positive class user) S, size of the graph S_g i.e. maximum number of nodes in a graph, width of the graph W_g i.e. the maximum number of frontiers or adjacency nodes for each blogger, a set of exemplary documents D_e, threshold th and n-gram value N_g for relevance computation. We create a list of 30 positive class bloggers extracted during experimental setup (refer to Subsection 4.7.1) and compute their relevance score against the exemplary documents. We take an average of these scores and compute the threshold value for language modeling. We use n-gram language modeling (N_g=3) to build our statistical model. Algorithm 12 is a recursive process that results in a cyclic directed graph. We run this algorithm until we get a graph of size S (1000 bloggers) or there is no node left in the queue for further extension. We perform a self-avoiding random walk that means we make sure a node is never being re-visited. If a node re-appears in the frontiers list then there are two possibilities: 1) the frontier has already been processed (extended or discarded based upon the relevance score- Steps 4 to 7). If it exists in the processed nodes list, then we create a directed edge between the node and its parent and avoid further extension. 2) If the re-appearing node is in frontiers list and is not yet processed, we created a directed edge in the graph and continued the traversal.

The topical crawler is a recursive process that adds and removes nodes after each iteration. The resultant graph is dynamic and not irreducible that means given a graph $G(V, E)$, if there is a directed edge between two nodes u and v, it is not necessary that there exists a directed path from v to u. Consider that object (topical crawler) processed node i at time $t-1$. In the next iteration object moves to an adjacency node of i. The probability that object moves to node j at time t is $\frac{1}{d_i^+}$ when there exists a direct edge from i to j. M_{ij}=$\frac{1}{d_i^+}$ denotes the probability to reach from i to j in one step where d_i^+ is the out-degree of node i. Therefore we can define:

Algorithm 12: Graph Traversal Using Random Walk Algorithm

```
Data: S, th, N_g, S_g, W_g, D_e
Result: Directed Graph G
SetParameters()
U_i = S, F.add(S)
Algorithm TopicalCrawler(S)
    while (graphsize < S_g OR F.size > 0) do
        Posts=ExtractPost(U_i)
        Relevance_Score = LanguageModeling(D_e, Posts, N_g)
        if (score >th) then
            Linked_Users=ExtractFrontiers(U_i)
            ProcessedNodes PN.add(U_i)
            for all LU ∈ Linked_Users do
                if (!(F.contains(LU) AND (PN.contains(LU))) then
                    F.add(LU)
                else
                    Discard the node LU
        else
            Discard the node U_i
        Compute the Markov Chain over graph G
        New_Blogger= node with maximum probability in Markov chain array
        F.remove(New_Blogger)
        TopicalCrawler (New_Blogger)
```

$$M_{i,j} = \begin{cases} \frac{1}{d_i^+}, & if\ (i,j)\ is\ an\ edge\ in\ digraph\ G \\ 0, & otherwise \end{cases} \tag{4.5}$$

Therefore for each vertex i, the sum of the probability to traverse an adjacency node of i is 1.

$$\forall i \sum_{j \in A(i)} M_{ij} = 1 \tag{4.6}$$

A(i) denotes the list of adjacency nodes i. In random walk on graph G topical crawler traverse along the nodes according to the probability of M_{ij}. Graph G is a dynamic social networking graph. Therefore we compute a Markov chain M after each iteration and compute the probability matrix over graph G. Markov chain is a random process where the probability distribution of node j depends on the current state of the matrix [173]. The probability matrix M^k gives us a picture of graph G after k iterations of a topical crawler. Using this matrix, we compute the probability distribution P that object moves to a particular vertex. P^k is the probability distribution of a node j after k iteration then probability of i to be traversed in $k + 1^{th}$ iteration is the following:

$$P^{k+1} = P^k.Mwhere, P^k = P^0 * M^k \tag{4.7}$$

Where P^0 is the initial distribution fixed for the seed node.

Table 4.16: Confusion Matrix and Accuracy Results for One Class Classifier

(a) Confusion Matrix

		Predicted	
		Positive	**Unknown**
Actual	**Positive**	290 (Tp)	45 (Fn)
	Unknown	92 (Fp)	173 (Tn)

(b) Accuracy Results

Precision	**Recall**	**F-score**
Tp/Tp+Fp	Tp/Tp+Fn	2PR/P+R
0.75	0.86	0.80

Table 4.17: Illustrating The Network Level Measurements for Topical Crawler. Dia= Network Diameter, Mod= Modularity, ACC= Average Clustering Coefficient, IBC= In- Betweenness Centrality, CC= In- Closeness Centrality, #SCC= Number of Strongly Connected Components.

Graph	**#Nodes**	**#Edges**	**Dia**	**#SCC**	**#ACC**	**#Mod**	**IBC**	**ICC**
TC	382	275	4	137	0.026	12.00	11.36	0.20
LB	27	60	1	21	0.0231	1.307	0	0.38
RB	355	215	6	185	0.021	7.01	6.284	0.40

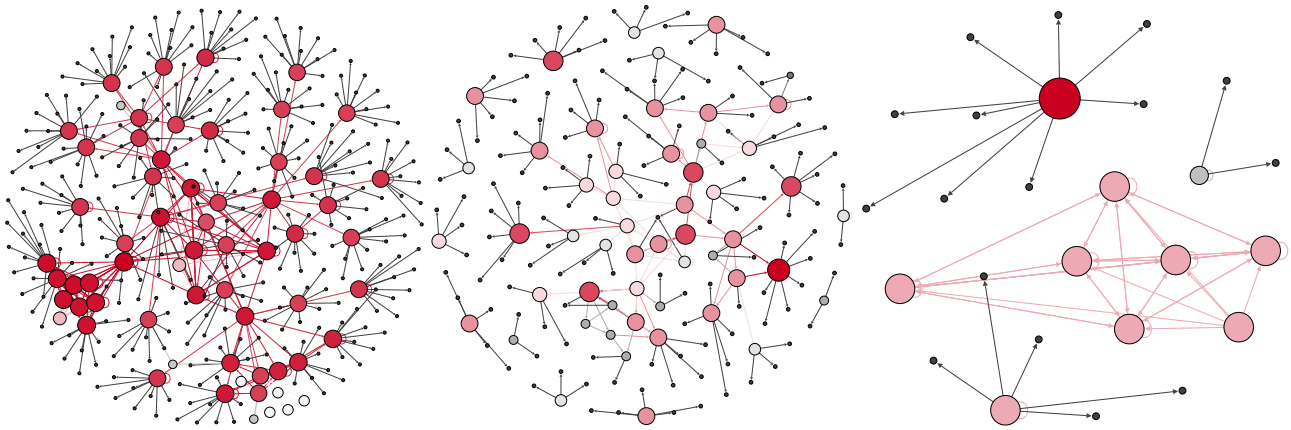
4.7.4 Experimental Results

4.7.4.1 Topical Crawler Results

We execute our topical crawler for a given seed blogger and traverse through Tumblr network using random walk algorithm. For every new blogger, we compute its relevance and classify it as hate promoting or unknown using one-class classifier. To examine the effectiveness of our classifier, we measure the accuracy using standard information retrieval techniques. In one execution of our topical crawler, we were able to collect 600 bloggers. We hired 30 graduate students as volunteers from the different departments to label these bloggers as hate promoting or unknown according to their published posts and given guidelines for annotation. To avoid the biases and to collect correct annotated results we perform a horizontal and vertical partition on nodes and arrange these 600 bloggers into a 2D matrix where rows are the numbers of annotators grouped in 10 sets, 3 members each. Columns of the matrix are the number of bloggers assigned to each member for annotation i.e. 60. We use majority voting approach for final annotation, the class of a blogger is the one which is voted by at least two annotators. Based on the validation results, we evaluate the accuracy of our model. Table 4.16a shows the confusion matrix for one class classification. Table 4.16a reveals that our model predicts 382 (290+92) bloggers as hate promoting and 218 (173+45) bloggers as unknown. Table 4.16a shows that there is a misclassification of 13% and 34% in predicting hate promoting and unknown bloggers. Table 4.16b shows the accuracy results of our classifier. Results show that the precision, recall, and f-score are reasonably high and we are able to predict hate promoting bloggers with an accuracy of 77%.

4.7.4.2 Social Network Analysis

We perform social network analysis on topical crawler's network resulted into a directed graph $G(V, E)$, where V represent a set of Tumblr bloggers accounts and E represent a directed edge between two bloggers. We define this edge as a relation having two labels 'posts liked by' and 'post re-blogged by'. To examine the effectiveness of these links, we generate two independent networks exclusively for 'like' and 'reblog' links



(a) (b) (c)

Figure 4.29: Cluster Representation of Social Network Graphs- Topical Crawler using (a)Random Walk, (b) 'Posts Liked by' and (c) 'Posts Re-blogged By'

among bloggers. Figure 4.29 illustrates the representations of these networks. In each graph, the size of the node is directionally proportional to its out-degree. A node with the maximum number of adjacency vertices is biggest in size. Colors in the graph represent the clusters of nodes having similar properties. Here, we define the similarity measure as the ratio of out-degree and in-degree.

We also perform several network level measurements on these graphs. Table 4.17 reveals that re-blogging is a good indicator of connection between two bloggers. Here, we observe that the graphs generated for topical crawler and re-blogging link have the same pattern in network measurements. Both graphs are dense (also evident from the Figure 4.29) and have higher modularity in comparison to the network created for 'liked' link. Table 4.17 and Figure 4.29 also reveal that by navigating through re-blogging links we can locate a large number of connected components in an extreme right community. While following 'like' as a link, we are able to detect a small number of connected blogs. Though, as illustrated in Figure 4.29b, we cannot entirely avoid this feature since a set of blogs extracted using this link are irreducible. Table 4.17 also shows that the graph created for 'like' relation has slightly larger value for average clustering coefficient. The reason is that the number of nodes in the network is very less and a major set of these nodes is strongly connected. A Higher value of 'In-between' centrality shows the presence of bloggers who are being watched by a large number of users. As the Figure 4.29c shows there are many users who are not directly connected to each other (shown in red color) but has a huge network of common bloggers. These disjoint bloggers are two or three hops away and are connected via other bloggers (with the second largest number of adjacency nodes). These bloggers are connected with the maximum number of other bloggers present in the graph and have a widespread network in extreme right communities. These nodes have the maximum closeness centrality and play the central role in the community. Nodes represented as black dots have a minimum number of out-degree nodes. They don't have a directed path to the central users or source of extremist posts. Based on our study, we find that these bloggers to be the target audiences who share their posts on their own network. These users are very crucial for such communities though they don't actively participate in the network.

4.8 Conclusions and Future Work

The immense popularity and wide reachability of social media platforms also gained the attention of several groups of people and individual who use social media as a platform to convey extremist thoughts among their followers. It is seen that due to low barriers to publication and anonymity, social media platforms such as YouTube, Twitter and Tumblr are being misused by extremist groups to spread hate, promote their ideology, recruiting young people in their groups and forming virtual communities sharing a common agenda. The presence of such content on social media is a major concern for the government and law enforcement agencies. In the work presented in this chapter, we formulate the problem of online radicalization detection as a classification and graph traversal problem and divide it into three sub-problems: 1) identification of hate promoting content on social media, 2) identification of radicalized posts based on the intent of the author and 3) identification of users and communities disseminating such content on social networking websites. To solve each of these sub-problems, we conduct our studies on Twitter, Tumblr, and YouTube platforms and evaluate the performance of our proposed approach.

In first case study, we mine Twitter data for identifying hate promoting content. We conduct a manual analysis of tweets and identify linguistic features which can be used as discriminators for the task of identifying hate and extremism promoting tweets. We demonstrate a correlation between such tweets and features like presence of war, religious, negative emotions and offensive terms. We train a one-class SVM and KNN on $10,486$ positive class tweets and observe an F-Score of 0.83 and 0.60 respectively. We implement a leave one out strategy and examine the influence of each discriminatory feature on overall accuracy of classifiers. Based upon the accuracy results, we conclude that presence of religious, war related terms, offensive words and negative emotions are strong indicators of a tweet to be hate promoting. Unlike KNN classifier, presence of internet slang and question mark plays an important role in LibSVM classifier. The reason why SVM outperforms KNN classifier is that our testing dataset is scattered in a high dimensional space and is trained on a relatively smaller size of dataset. We perform a content based characterization on training and testing datasets. Presence of low quality content (misspells words, abbreviations and ambiguity) in tweets increases the data sparsity that eventually degrades the performance of our classifiers. We also conclude that classifying tweets based upon the presence of certain hashtags is an in-efficient approach. Overall accuracy of classifier falls down as it increases the number of false positives.

In second case study, we study the problem of identifying racist and radicalized Tumblr posts based on the intent of narrative. We formulate our problem as a cascaded ensemble learning problem and propose a two-stage one-class classification approach to solve the problem. Our result shows that the proposed approach is effective for identifying intent posts unlike previous keyword based techniques. Our experimental results shows that emotion tone, writing cues and social personality traits of an author are discriminatory features for identifying the intent of the post. Further, topic classification of posts and filtering non-topic based (or noisy) posts improves the performance of the proposed intent classification. Random Forest is an ensemble learning based model trained from multiple Decision Tree based models and used as a boosting approach. Therefore, the performance of Random Forest is higher than the Decision Tree based algorithm. Naive Bayes algorithm computes the probability of a new test object to be in the target class or an outlier. Due to the small size of training dataset, we record a lower precision rate of Naive Bayes classifier in comparison to the Random Forest algorithm.

In third case study, we present a focused-crawler based approach for identification of hate and extremism promoting users and community YouTube. We propose to use best first search and shark search- two graph traversal algorithm for navigating through the website and identify extremist users. Experimental results reveal higher precision, recall and accuracy (0.74) for shark-search approach in comparison to best-first search (0.69). We conduct a series of experiments by varying various algorithmic parameters such as the similarity threshold for the language modelling based text classifier and n-grams. We conclude that by performing

social network analysis on network graphs, we are able to locate hidden communities. We identify the users who play major roles in the communities and have highest centrality among all. We reveal the communities by dividing the network graph into clusters formed by similar users. In SSA we find more strongly connected components (16) and communities in comparison to BFS (7). We perform a characterization on the content and contextual information of several hate promoting videos. The analysis reveals that hate promoting users upload videos targeting some specific audiences. Majority of videos are very large in the duration (3 to 45 minutes). Keywords present in the contextual information and video content are the evidence of these videos doing hate promotion among their viewers.

In order to propose solution for the 4^{th} sub-problem, we perform a case study on Jihadist groups and locate their existing extreme right communities on Tumblr. We conduct experiments on real world dataset and use topical crawler based approach to collect textual data (published posts) from Tumblr users. We perform one class classification and identify hate promoting bloggers according to the content present in their posts. We use random walk algorithm for graph traversal and extract exclusive links to these bloggers. We conclude that by performing social networking analysis on a graph (vertices are the Tumblr bloggers and edges are the links among these bloggers: re-blog and like) we are able to uncover hidden virtual communities of extremist bloggers with an accuracy of 77%. We compute various centrality measures to locate the influential bloggers playing major roles in extremist groups. We also investigate the effectiveness of link features (likes and re-blogs) in order to find the communities. Our results reveal re-blogging is a strong indicator and a discriminatory feature to mine strongly connected communities on Tumblr. We perform a manual inspection on Tumblr and perform a characterization on several hate promoting posts. Our study reveals that these posts are very much popular among extremist bloggers and get large number of hits. These posts are published targeting some specific audiences. Keywords present in the blog content, tags associated with post and comments by other bloggers are clear evidence of hate promotion among their viewers.

Future work includes addressing the limitations of present study and improving the accuracy of linguistic features. Identification of multilingual posts by doing a sentence level language detection and enhancing the translated content for identifying hate promoting posts. As mentioned in the previous sections, social media allows users to embed multimedia content in their posts. Therefore, our future work involves mining users' reactions from attached external images and enrichment of linguistic features of a post.

Chapter 5

Using Commonsense Knowledge for Detecting Word Obfuscation in Adversarial Behavior

Research shows that terrorists use social media platforms such as Facebook [4] and Twitter [5] for communication[1]. While several posts by terrorists are aimed at spreading their propaganda, a large number of posts are also aimed at communication between them by staying hidden. It is important for the law enforcement and intelligence agencies to not only identify social media content disseminating hatred but also detect communication between terrorists in which the terrorists have carefully concealed the content to avoid any attention from monitoring agencies [174]. Intelligence and security agencies intercept and scan billions of messages and communications every day to identify the dangerous communications between terrorists and criminals. Surveillance by Intelligence agencies consists of intercepting mail, mobile phone and satellite communications[2]. Message interception to detect harmful communication is not only done by Intelligence agencies to counter terrorism but also by law enforcement agencies to combat criminal and illicit acts [174] [175]. Law enforcement and Intelligence agencies have a watch-list or lexicon of red-flagged terms such as attack, bomb and heroin [176]. The watch-list of suspicious terms are used for keyword-spotting in intercepted messages which are filtered for further analysis [177]. The article by BBC[3] titled "How do terrorists communicate" defines two categories of communication: secret and public messages. Secret messages on social media between terrorists are not aimed at online radicalization and rather activities such as planning and organization. The secret messages between terrorists are written in such a way that they are hard to intercept and filter by monitoring agencies. Terrorists make use of open source social media for secret communication as a preferred strategy than using closed source communication like person email accounts and encrypting messages. This is because techniques like encryption can draw suspicion and emails can be intercepted well. Term obfuscation is a common strategy followed by terrorists to communicate secretly (and yet openly on an open-source publicly available social media channel) which does not draw suspicion. Textual or word substitution consists of replacing a red-flagged term (which is likely to be present in the watch-list[4]) with an "ordinary" or an "innocuous" term. Innocuous terms are those terms which are less likely to attract the attention of security agencies. For example, the word attack being replaced by the phrase birthday function and bomb being replaced by the term milk. Research shows that terrorist use low-tech word substitution

[1] http://fortune.com/2016/05/03/terrorists-email-social-media/

[2] https://www.wired.com/2007/08/wiretap/

[3] http://www.bbc.com/news/world-24784756

[4] https://www.state.gov/strategictrade/redflags/

Table 5.1: List of Previous Work in the Area of Detecting Word Obfuscation in Adversarial Communication. ED: Evaluation Dataset, RS: Resources Used in Solution Approach, SA: Solution Approach

	Deshmukh et al. 2008 [178]	Fong et al. 2008 [176]
ED	Google News	Enron e-mail dataset, Brown corpus
RS	Google search engine	British National Corpus (BNC), WordNet, Yahoo, Google and MSN search engine
SA	Measuring sentence oddity (MSO), enhance sentence oddity and K-grams frequencies (KGF)	MSO, KGF, Hypernym Oddity (HO) and Pointwise Mutual Information (PMI)

	Jabbari et al. 2008 [177]	Fong et al. 2006 [174]
ED	British National Corpus (BNC)	Enron e-mail dataset
RS	1.4 billion words of English Gigaword v.1 (newswire corpus)	BNC, WordNet, Google search engine
SA	Probabilistic or distributional model of context	MSO, semantic measure using WordNet, and KGF (bigrams)

than encryption as encrypting messages itself attracts attention [178] [176]. Al-Qaeda used the term wedding for attack ("The wedding will be in two weeks" and "The grooms are ready for the big wedding") [179] and architecture for World Trade Center in their email communication. Automatic word obfuscation detection is natural language processing problem that has attracted several researchers' attention. The task consists of detecting if a given sentence has been obfuscated and which term(s) in the sentence has been substituted. The research problem is intellectually challenging and non-trivial as natural language can be vast and ambiguous (due to polysemy and synonymy) [177].

In Chapter 4, we present our work on detecting content on social media aimed at online radicalization. In this chapter, we present our work on identifying content on social media aimed at detecting and decoding secret messages between terrorists. There is no publicly available social media data for the research community for analyzing secret messages between terrorists, and it is not possible to create and annotate such data from social media because the messages are hidden, secret and not explicit. Hence we conduct our experiments on an alternate dataset validate the effectiveness of our proposed technique on term obfuscation detection and secret message identification.

5.1 Background

In this Section, we discuss closely related work to the study presented in this chapter. Term obfuscation in adversarial communication is an area that has attracted several researchers' attention. Table 5.1 displays a list of traditional techniques consisting of the evaluation dataset (ED), lexical resource and the high-level solution approach applied in each of the four methods. Table 5.1 reveals that majority of the studies are conducted on news corpora and email communications. Whereas, web search engines and English language word dictionaries and lexical databases are used as a back-end database resources. Prior literature shows that keyword-based-flagging and probabilistic methods are the commonly used techniques for identifying substituted term in a sentence. Existing solution approaches consist of measuring sentence oddity [174] [176]

[178], probability distribution [177], semantic similarity [174] and Pointwise Mutual Information [176] to compute out-of-context terms in a given sentence. During our literature survey, we observe that all proposed approaches in existing studies are focused towards the substitution of the first noun in a sentence.

5.1.1 ConceptNet- A Commonsense Knowledgebase

ConceptNet[5] is a semantic network consisting of nodes (representing concepts) and edges (representing relations between the concepts). ConceptNet is a freely available commonsense knowledgebase which contains everyday basic knowledge [107] [180]. It has been used as a lexical resource and natural language processing toolkit for solving many NLP and textual reasoning tasks [181]. We hypothesize that ConceptNet can be used as a semantic knowledgebase to solve the problem of textual or word obfuscation. We believe that the relations between concepts in ConceptNet can be exploited to find conceptual similarity between given concepts and use to detect out-of-context terms or terms which typically do not co-occur together in everyday communication. We briefly discuss some of the recent and related work using ConceptNet as a lexical resource. Wu et al. [182] use relation selection to improve value propagation in a ConceptNet-based sentiment dictionary (sentiment polarity classification task). Bouchoucha et al. [183] use ConceptNet as an external resource for query expansion. Revathi et al. [184] present an approach for similarity based video annotation utilizing commonsense knowledgebase. They apply a Local Binary Pattern (LBP) and commonsense knowledgebase to reduce the semantic gap for non-domain specific videos automatically. Poria et al. [105] propose a ConceptNet-based semantic parser that deconstructs natural language text into concepts based on the dependency relation between clauses. Their approach is domain-independent and is able to extract concepts from the heterogeneous text.

5.2 Research Contributions

In context to existing work, the study presented in this chapter makes the several unique and novel research contributions:

1. The study presented in this chapter is the first focused research investigation on the application of ConceptNet commonsense knowledgebase for solving the problem of textual or term obfuscation. While there has been work done in the area of using a corpus as a lexical resource for the task of word obfuscation detection, the application of an ontology like ConceptNet for determining the conceptual similarity between given terms and identifying out-of-context or odd terms in a given sentence is novel in context to previous work.
2. We conduct an in-depth empirical analysis to examine the effectiveness of the proposed approach. The test dataset consists of examples extracted from research papers on term obfuscation, Enron email dataset (having over 600000 emails generated by 158 employees of Enron Corporation) and Brown corpus (totaling about a million words drawn from a wide variety of sources).

5.3 Proposed Solution Approach

Figures 5.1 and 5.2 illustrates the general research framework for the proposed solution approach. The proposed solution approach primarily consists of two phases labeled as A and B (refer to Figure 5.1). In Phase A, we tokenize a given sentence S into a sequence of terms and tag each term with their part-of-speech.

[5] `http://conceptnet5.media.mit.edu/`

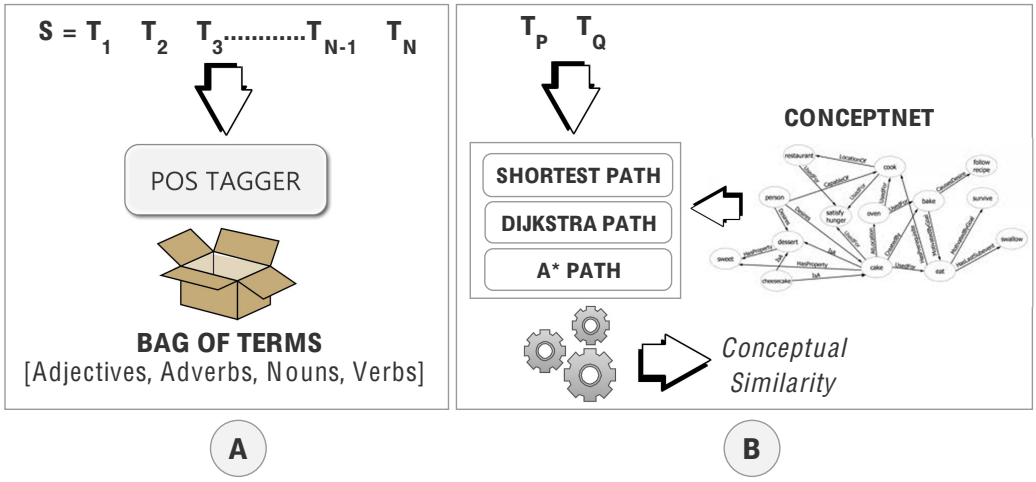


Figure 5.1: Solution framework demonstrating two phases in the processing pipeline. Phase *A* shows tokenizing given sentence and applying the part-of-speech-tagger. Phase *B* shows computing conceptual similarity between any two given term using ConceptNet as a lexical resource and applying graph distance measures.

We use Natural Language Toolkit[6] (NLTK) part-of-speech tagger for tagging each term. We exclude non-content bearing terms using an exclusion list. For example, we exclude conjunctions (and, but, because), determiners (the, an, a), prepositions (on, in, at), modals (may, could, should), particles (along, away, up) and the base form of verbs. We create a bag-of-terms (a set) with the remaining terms in the given sentence. As shown in Figures 5.1 and 5.2, Phase B consists of computing the Mean Average Conceptual Similarity (MACS) score for a bag-of-terms and identify obfuscated term in a sentence using the MACS score. The conceptual similarity between any two given terms T_p and T_q is computed by taking the average of number of edges in the shortest path between T_p & T_q and T_q & T_p (and hence the term *average* in MACS). We use three different algorithms (Dijkstra's, A* and Shortest path) to compute the number of edges between any two given terms. The different algorithms are experimental parameters, and we experiment with three different algorithms to identify the most effective algorithm for the given task.

Let us say that the size of the bag-of-terms after Phase A is N. As shown in Figure 5.2, we compute the MACS score N times. The number of comparisons (computing the number of edges in the shortest path) required for computing a single MACS score is twice of ${}^{(N-1)}P_2$ times. Consider the scenario in Figure 5.2, the MACS score is computed 4 times for the four terms: A, B, C and D. The comparison required for computing the MACS score for A are: $B-C$, $C-B$, $B-D$, $D-B$, $C-D$ and $D-C$. Similarly, the comparisons required for computing the MACS score for B are: $A-C$, $C-A$, $A-D$, $D-A$, $C-D$ and $D-C$. The obfuscated term is the term for which the MACS score is the lowest. Lower number of edges between two terms indicate higher conceptual similarity. The intuition behind the proposed approach is that a term will be out of-context in a given bag-of-terms if the MACS score of terms minus the given term is low. The out-of-context term will increase the average conceptual similarity and hence the MACS score.

[6] `www.nltk.org`

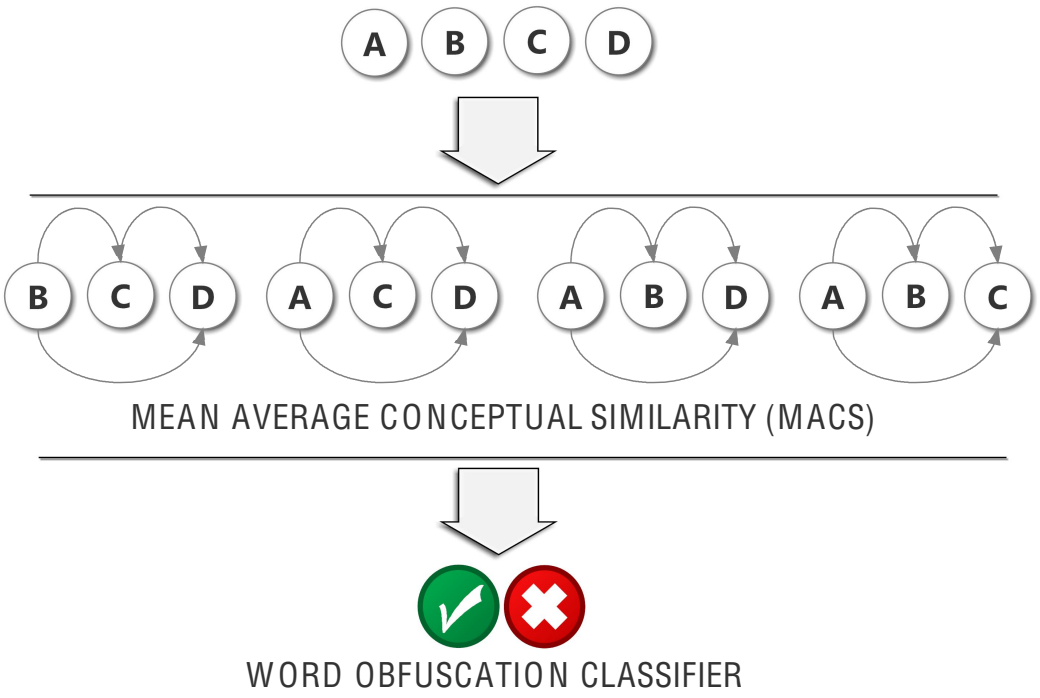


Figure 5.2: High-Level framework demonstrating the procedure of computing Mean Average Conceptual Similarity (MACS) score for a bag-of-terms and for determining the term which is out-of-context. The given example consisting of four terms *A*, *B*, *C* and *D* requires computing conceptual similarity between two terms 12 times.

5.3.1 Worked-Out Example

We take two concrete worked-out examples in-order to explain our approach. Consider a case in which the original sentence is: "We will attack the airport with bomb". The red-flagged term in the given sentence is bomb. Let us say that the term bomb is replaced with an innocuous term flower and hence the obfuscated textual content is: "We will attack the airport with flower". The bag-of-terms (nouns, adjectives, adverbs and verbs and not including terms in an exclusion list) in the substituted text is [attack, airport, flower]. The conceptual similarity between airport and flower is 3 as the number of edges between airport and flower is 3 (airport, city, person, flower) and similarly, the number of edges between flower and airport is 3 (flower, be, time, airport). The conceptual similarity between attack and flower is also 3. The number of edges between attack and flower is 3 (attack, punch, hand, flower) and the number of edges between flower and attack is 3 (flower, be, human, attack). The conceptual similarity between attack and airport is 2.5. The number of edges between attack and airport is 2 (attack, terrorist, airport) and the number of edges between airport and attack is 3 (airport, airplane, human, attack). The Mean Average Conceptual Similarity (MACS) score is $(3 + 3 + 2.5)/3 = 2.83$. In the given example consisting of 3 terms in the bag-of-terms, we computed the conceptual similarity between two terms six times.

Consider another example in which the original sentence is: "Pistol will be delivered to you to shoot the president". Pistol is clearly the red-flagged term in the given sentence. Let us say that the term Pistol is replaced with an ordinary term Pen as a result of which the substituted sentence becomes: "Pen will be delivered to you to shoot the president". After applying part-of-speech tagging, we tag pen and president

as noun and shoot as a verb. The bag-of-terms for the obfuscated sentence is: [pen, shoot, president]. The conceptual similarity between shoot and president is 2.5 as the number of edges between president and shoot is 2 (president, person, shoot) and similarly, the number of edges between shoot and president is 3 (shoot, fire, orange, president). The conceptual similarity between pen and president is 3. The number of edges between president and pen is 3 (president, ruler, line, pen) and the number of edges between pen and president is 3 (pen, dog, person, president). The conceptual similarity between pen and shoot is 3.0. The number of edges between shoot and pen is 3 (shoot, bar, chair, pen) and the number of edges between pen and shoot is 3 (pen, dog, person, shoot). The Mean Average Conceptual Similarity (MACS) score is $(2.5 + 3 + 3)/3 = 2.83$.

Algorithm 13: Obfuscated Term Detection

Data: Substituted Sentence S', ConceptNet Corpus C
Result: Obfuscated Term O_T

```
1  for all record r ∈ C do
2  |  Edge E.add(r.node_1, r.node_2, r.relation)
3  |  Graph G.add(E)
4  tokens = S'.tokenize()
5  pos.add(pos_tag(tokens)
6  for all tag ∈ pos and token ∈ tokens do
7  |  if tag is in (verb, noun, adjective, adverb) then
8  |  |  BoW.add(token.lemma)
9  for iter = 0 to BoW.length do
10 |  concepts = BoW.pop(iter)
11 |  for i = 0 to concepts.length-1 do
12 |  |  for j = i to concepts.length do
13 |  |  |  if (i! =j) then
14 |  |  |  |  path c_{i,j} = Dijkstra_path_len(G, i, j)
15 |  |  |  |  path c_{j,i} = Dijkstra_path_len(G, j, i)
16 |  |  |  |  avg.add(Average(c_{i,j}, c_{j,i})
17 |  mean.add(Mean(avg))
18 O_T = BoW.valueAt(min(mean))
```

5.3.2 Solution Pseudo-code and Algorithm

Algorithm 13 describes the proposed method to identify an obfuscated term in a given sentence. Inputs to our algorithm is a substituted sentence S' and the ConceptNet 4.0 corpus C (a commonsense knowledgebase). In Steps 1 to 3, we create a directed network graph from ConceptNet corpus where nodes represent concepts and edge represents a relation between two concepts (for example, HasA, IsA, UsedFor). As described in the research framework (refer to Figures 5.1 and 5.2), in Steps 4 to 5, we tokenize S' and apply part-of-speech tagger to classify terms according to their lexical categories (such as noun, verbs, adjectives and adverbs). In Steps 6 to 8, we create a bag-of-terms of the lemma of verbs, nouns, adjectives and adverbs that are present in S'. In Steps 9 to 17, we compute the mean average conceptual similarity (MACS) score for bag-of-terms. In Step 18, we compute the minimum of all MACS scores to identify the obfuscated term. In proposed method, we use three different algorithms to compute the shortest path length between the concepts.

Figure 5.3 shows an example of shortest path between two terms Pen and Blood using Dijkstra's, A*

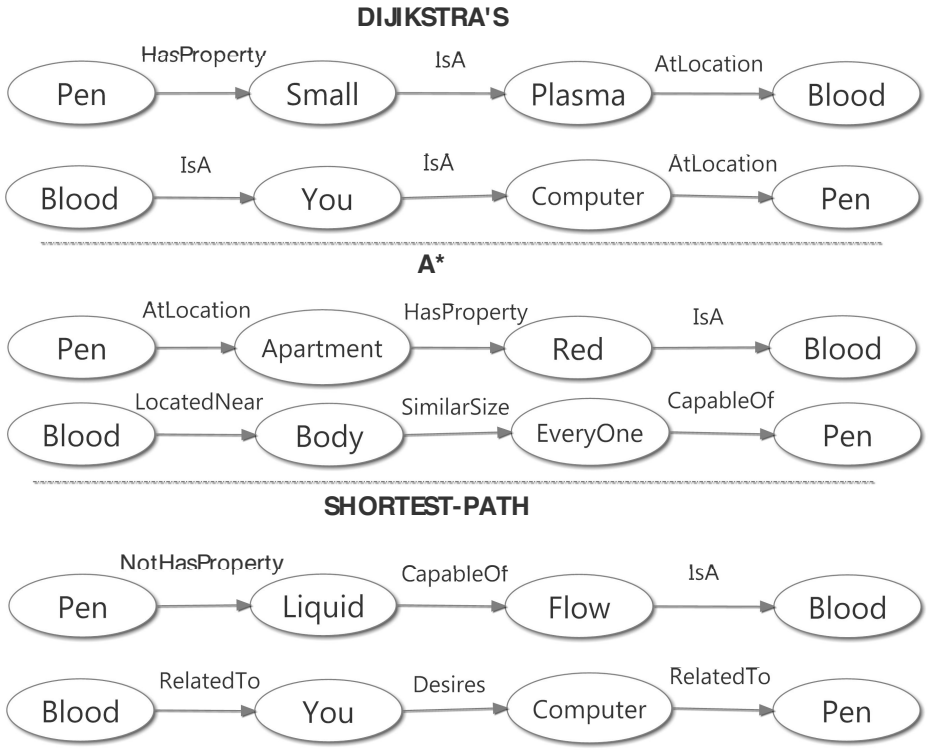


Figure 5.3: ConceptNet paths (nodes and edges) between two concepts 'Pen' and 'Blood' using three different distance metrics

Table 5.2: Concrete Examples of Computing Conceptual Similarity between Two Given Terms Using Three Different Distance Metrics or Algorithms (NP: Denotes No-Path between the Two Terms and is Given a Default Value of 4).

		Dijkstra's Algo			A-Star Algo			BFS Algo		
Term 1	Term2	T1-T2	T2-T1	Mean	T1-T2	T2-T1	Mean	T1-T2	T2-T1	Mean
Tree	Branch	1	1	1	1	1	1	1	1	1
Pen	Blood	3	3	3	3	3	3	3	3	3
Paper	Tree	1	1	1	1	1	1	1	1	1
Airline	Pen	4(NP)	4	4	4(NP)	4	4	4(NP)	4	4
Bomb	Blast	2	4(NP)	3	2	4(NP)	3	2	4(NP)	3

and Shortest path algorithms. As shown in Figure 5.3, the path between the two terms Pen and Blood can be different than the path between the terms Blood and Pen (terms are same but the order is different). For example, the path between Pen and Blood using A* consists of Apartment and Red as intermediate nodes whereas the path between Blood and Pen consists of nodes Body and EveryOne using the A* algorithm. Also, the Figure 5.3 demonstrates that the path between the same two terms is different for different algorithms.

Two terms are related to each other in various contexts. In ConceptNet, the path length describes the extent of semantic similarity between concepts. If two terms are conceptually similar, then the path length will be smaller in comparison to the terms that are highly dissimilar. Therefore if we remove an obfuscated word from the bag-of-terms the MACS score of remaining terms will be minimum. Table 5.2 shows some

Table 5.3: Concrete Examples of Conceptually and Semantically Un-related Terms and their Path Length (PL) to Compute the Default Value for No-Path

T1	T2	PL	T1	T2	PL	T1	T2	PL
Bowl	Mobile	3	Office	Festival	3	Feather	Study	3
Wire	Dress	3	Coffee	Research	3	Driver	Sun	3

concrete examples of semantic similarity between two concepts. Table 5.2 illustrates that the terms Tree & Branch and Paper & Tree are conceptually similar and has a path length of 1 which means that both the concepts are directly connected in the ConceptNet knowledgebase. NP denotes no path between the two concepts. For example, in Table 5.2 we have a path length of 2 from source node Bomb to target node Blast while there is no path from Blast to Bomb. We use a default value of 4 in the case of no path between two concepts. We conduct an experiment on ConceptNet 4.0 and compute the distance between highly dissimilar terms. Table 5.3 shows that in the majority of cases the path length between semantically unrelated terms is 3. Therefore we use 4 (distance between unrelated terms + 1 for upper bound) as a default value for no path between two concepts.

5.4 Experimental Evaluation and Validation

As an academic researcher, we believe and encourage academic code or software sharing in the interest of improving openness and research reproducibility. We release our term obfuscation detection tool *Parikshan* in public domain so that other researchers can validate our scientific claims and use our tool for comparison or benchmarking purposes. *Parikshan* is a proof-of-concept hosted on GitHub which is a popular web-based hosting service for software development projects. We provide installation instructions and a facility for users to download the software as a single zip-file[7]. Another reason of hosting on GitHub is due to an integrated issue tracker which makes reporting issues easier by our users (and also GitHub facilitates easier collaboration and extension through pull-requests and forking). We believe our tool has utility and value in the domain of intelligence and security informatics and the spirit of scientific advancement, select GPL license (restrictive license) so that our tool can never be closed-sourced.

5.4.1 Experimental Dataset

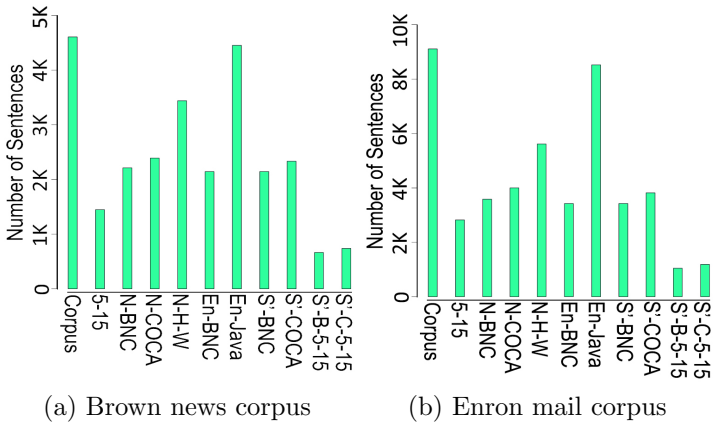
We conduct experiments on publicly available dataset so that our results can be used for comparison and benchmarking. We download two datasets: Enron e-mail corpus[8] and Brown news corpus[9]. We also use the examples extracted from 4 research papers on word substitution. Hence we have a total of three experimental datasets to evaluate our proposed approach. We believe conducting experiments on three diverse evaluation dataset will prove the generalizability of our approach and thus strengthen the conclusions. Enron e-mail corpus consists of about half a million e-mail messages sent or received by about 158 employees of Enron Corporation. This dataset was collected and prepared by the CALO Project[10]. We perform a random sampling on the dataset and select 9000 unique sentences for substitution. Brown news corpus consists of about a million words from various categories of formal text and news (for example, political, sports, society and cultural). This dataset was created in 1961 at Brown University. Since the writing style in Brown

[7] https://github.com/ashishsureka/Parikshan

[8] http://verbs.colorado.edu/enronsent/

[9] http://www.nltk.org/data.html

[10] https://www.cs.cmu.edu/~./enron/



(a) Brown news corpus (b) Enron mail corpus

Figure 5.4: Bar Chart Presentation of the Experimental Dataset Statistics (refer to Table 5.4 for exact values).

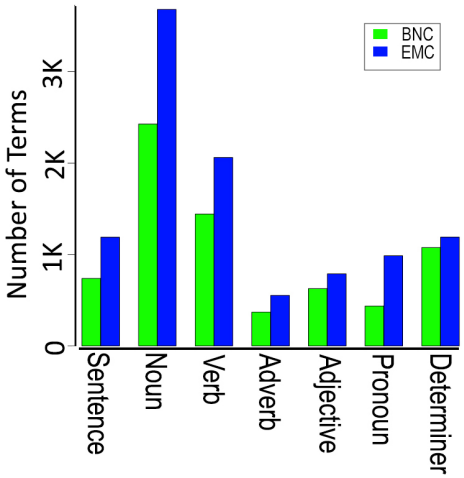


Figure 5.5: Bar Chart for the Number of Part-of-Speech Tags in Experimental Dataset.

Table 5.4: Experimental Dataset Statistics for the Brown News Corpus (BNC) and Enron Mail Corpus (EMC) (Refer to Figure 5.4 for the Graphical Plot of the Statistics), #=Number of

Abbr	Description	BNC	EMC
Corpus	Total sentences in brown news corpus	4607	9112
5-15	Sentences that has length between 5 to 15	1449	2825
N-BNC	Sentences that has their first noun in BNC (british national corpus)	2214	3587
N-COCA	Sentences that has their first noun in 100 K list (COCA)	2393	4006
N-H-W	If first noun has an hypernym in WordNet	3441	5620
En-BNC	English sentences according to BNC	2146	3430
En- Java	English sentences according to Java language detection library	4453	8527
S'-BNC	#Substituted sentences using BNC list	2146	3430
S'-COCA	#Substituted sentences using COCA (100K) list	2335	3823
S'-B-5-15	#Substituted sentences (between length of 5 to 15) using BNC list	666	1051
S'-C-5-15	#Substituted sentences (between length of 5 to 15) using COCA list	740	1191

news corpus is much more formal than Enron e-mail corpus, we use these two different datasets to examine the effectiveness of our approach. We perform a word substitution technique (refer to Section 5.4.2) on a sample of 9000 sentences from Enron e-mail corpus and all 4600 of Brown news corpus. Figure 5.4 shows the statistics of both the datasets before and after the word substitution. Figure 5.4a and 5.4b also illustrates the variation in number of sentences substituted using traditional approach (proposed in Fong et al. [176]) and our approach. Figure 5.4a and 5.4b reveals that COCA is a huge corpus and has more nouns in the frequency list in comparison to BNC frequency list. Table 5.4 displays the exact values for the points plotted in the two bar charts of Figure 5.4.

Table 5.4 reveals that for Brown news corpus, using BNC (British National Corpus) frequency list, we

Table 5.5: Concrete Examples of Sentences Presented in EMC and BNC Discarded During Word Substitution

Corpus	Sentence	Reason
EMC	Since we're ending 2000 and going into a new sales year I want to make sure I'm not holding resource open on any accounts which may not or should not be on the list of focus accounts which you and your team have requested our involvement with.	Sentence length is not between 5 to 15
EMC	next Thursday at 7:00 pm Yes yes yes.	First noun is not in BNC/COCA list
BNC	The City Purchasing Department the jury said is lacking in experienced clerical personnel as a result of city personnel policies	Sentence length is not between 5 to 15
BNC	Dr Clark holds an earned Doctor of Education degree from the University of Oklahoma	First noun does not have a hypernym in WordNet

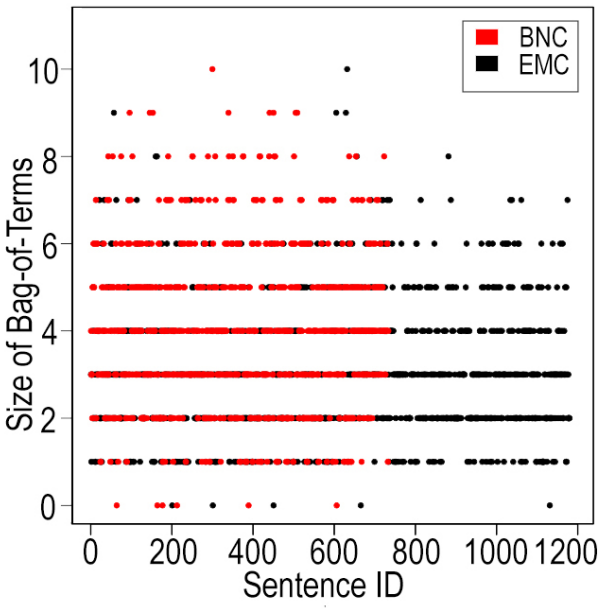


Figure 5.6: Scatter Plot Diagram for the Size of Bag-of-Terms in Experimental Dataset

are able to detect only 2146 English sentences. Whereas, using Java language detection library, we are able to detect 4453 English sentences. Similarly, in Enron e-mail corpus, BNC frequency list recognizes only 3430 English sentences while Java language detection library identifies 8527 English sentences. Therefore using COCA frequency list and Java language detection library, we are able to substitute more sentences (740 and 1191) in comparison to previous approach (666 and 1051). Table 5.4 reveals that initially we have a dataset of 4607 and 9112 sentences for BNC and EMC respectively. After word substitution, we are remaining with only 740 and 1191 sentences. Some sentences are discarded because they do not satisfy several conditions of word obfuscation. Table 5.5 shows some concrete examples of such sentences from BNC and EMC datasets.

Algorithm 14: Text Substitution Technique

Data: Sentence S, Frequency List $COCA$, WordNet DataBase W_{DB}
Result: Substituted Sentence S'

```
1  if (5 < S.length < 15) then
2      tokens ←S.tokenize()
3      POS ←S.pos_tag()
4      NF ←token[POS.indexOf("NN")]
5      if (COCA.has(NF) AND W_DB.has(NF.hypernym)) then
6          lang ← S.Language Detection
7          if (lang == "en") then
8              F_NF ←COCA.freq(NF)
9              F_NF' ← COCA.nextHigherFreq(F_NF)
10             NF' ← COCA.hasFrequency(F_NF')
11             S' ← S.replaceFirst(NF, NF')
12             return S'
```

We use 740 substituted sentences from Brown news corpus, 1191 sentences from Enron e-mail corpus and 22 examples from previous research papers as our testing dataset. As shown in research framework (refer to Figure 5.1) we apply a part-of-speech tagger on each sentence to remove non-content bearing terms. Figure 5.5 illustrates the frequency of common part-of-speech tags present in Brown News Corpus (BNC) and Enron e-mail corpus (EMC). As shown in Figure 5.5, the most frequent part-of-speech in the dataset is nouns followed by verbs. Figure 5.6 shows the length of bag-of-terms for every sentence present in BNC and EMC datasets. Figure 5.6 reveals that 5 sentences in Enron e-mail corpus and 6 sentences in Brown news corpus have an empty bag-of-terms which makes the system difficult to identify an obfuscated term. Figure 5.6 reveals that for majority of sentences size of bag-of-terms varies between 2 to 6. It also illustrates the presence of sentences that have an insufficient number of concepts (size <2) or the sentences that have a large number of concepts (size >7).

5.4.2 Term Substitution Technique

We substitute a term in a sentence using an adaptive version of a substitution technique originally proposed by Fong et al. [176]. Algorithm 14 describes the steps to obfuscate a term in a given sentence. We use WordNet database[11] as a language resource and the Corpus of Contemporary American English (COCA) as a word frequency data. COCA is a corpus of American English that contains more than 450 million words collected from 1990-2012[12]. In Step 1, we check the length of a given sentence S. If the length is between 5 to 15 then we proceed further otherwise we discard that sentence. In Steps 2 and 3, we tokenize the sentence S and apply part-of-speech tagger to annotate each word. In Step 4, we identify the first noun NF from this word sequence POS. In Steps 5, we check if NF is present in COCA frequent list and has an hypernym in WordNet. If the condition satisfies then we detect the language of the sentence using Java language detection library[13]. If the sentence language is not English, then we ignore it and if it is English, then we further process it. In Steps 8 to 11, we check the frequency of NF in COCA corpus and replace the term in the sentence by a new term NF' with the next higher frequency in COCA frequency list. This new

[11] `http://wordnet.princeton.edu/wordnet/download/`
[12] `http://www.wordfrequency.info/`
[13] `https://code.google.com/p/language-detection/`

Table 5.6: Example of Term Substitution using COCA Frequency List. NF= First Noun/ Original Term, ST= Substituted Term

Sentence	NF	Freq	ST	Freq	Sentence
Any opinions expressed herein are solely those of the author.	Author	53195	Television	53263	Any opinions expressed herein are solely those of the television.
What do you think that should help you score women.	Score	17415	Struggle	17429	What do you think that should help you struggle women.
This was the coolest calmest election I ever saw Colquitt Policeman Tom Williams said	Election	40513	Republicans	40515	This was the coolest calmest republicans I ever saw Colquitt Policeman Tom Williams said
The inadequacy of our library system will become critical unless we act vigorously to correct this condition	Inadequacy	831	Inevitability	831	The inevitability of our library system will become critical unless we act vigorously to correct this condition

term NF' is the obfuscated term. If NF has the highest frequency in COCA corpus, then we substitute it with the term which appears immediate before NF in frequency list. If two terms have the same frequency, then we sort those terms in alphabetical order and select immediate next term to NF for substitution. Table 5.6 shows some concrete examples of substituted sentences taken from Brown news corpus and Enron e-mail corpus. In Table 5.6, Freq denotes the frequency of first noun and it's substituted term in COCA frequency list. Table 5.6 also shows an example where two terms have the same frequency. We replace the first noun with the term that has equal frequency and is next immediate to NF in alphabetical order.

In Fong et al.; they use British National Corpus (BNC) as word frequency list. We replace BNC list by COCA frequency list because it is the largest and most accurate frequency data of English language and is five times bigger than the BNC list. The words in COCA are divided among a variety of texts (for example, spoken, newspapers, fiction, and academic texts) which are best suitable for working with commonsense knowledgebase. In Fong et. al; they identify the sentence to be in the English language if NF is present in BNC frequency list. Since the size of BNC list is comparatively small, we use Java language detection library for identifying the language of the sentence [185]. Java language detection library supports 53 languages and is much more flexible in comparison to BNC frequency list.

5.4.3 Experimental Results

5.4.3.1 Examples from Research Papers (ERP)

As described in section 5.4.1, we run our experiments on examples used in previous papers. Table 5.7 shows 22 examples extracted from 4 research papers on term obfuscation (called as ERP dataset). Table 5.7 shows the original sentence, substituted sentence, research paper and the result produced by our tool. Experimental results reveal 72.72% accuracy of our solution approach (16 out of 22 correct output).

Table 5.7: List of Original and Substituted Sentences used as Examples in Papers on Word Obfuscation in Adversarial Communication

	Original Sentence	Substituted Sentence	Result
1	the bomb is in position [174]	the alcohol is in position	alcohol
2	copyright 2001 south-west airlines co all rights reserved [174]	toast 2001 southwest airlines co all rights reserved	southwest
3	please try to maintain the same seat each class [174]	please try to maintain the same play each class	try
4	we expect that the attack will happen tonight [176]	we expect that the campaign will happen tonight	campaign
5	an agent will assist you with checked baggage [176]	an vote will assist you with checked baggage	vote
6	my lunch contained white tuna she ordered a parfait [176]	my package contained white tuna she ordered a parfait	package
7	please let me know if you have this information [176]	please let me know if you have this men	know
8	It was one of a series of recommendations by the Texas Research League [176]	It was one of a bank of recommendations by the Texas Research League	recomm.
9	The remainder of the college requirement would be in general subjects [176]	The attendance of the college requirement would be in general subjects	attendance
10	A copy was released to the press [176]	An object was released to the press	released
11	works need to be done in Hydrabad [178]	works need to be done in H	H
12	you should arrange for a preparation of blast [178]	you should arrange for a preparation of daawati	daawati
13	my friend will come to deliver you a pistol [178]	my friend will come to deliver you a CD	CD
14	collect some people for work from Gujarat [178]	collect some people for work from Musa	Musa
15	you will find some bullets in the bag [178]	you will find some pen drives in the bag	pen drives
16	come at Delhi for meeting [178]	come at Sham for meeting	Sham
17	send one person to Bangalore [178]	send one person to Bagu	Bagu
18	Arrange some riffles for next operation [178]	Arrange some DVDs for next operation	DVDs
19	preparation of blast will start in next month [178]	preparation of Daawati work will start in next month	Daawati
20	find one place at Hydrabad for operation [178]	find one place at H for operation	H
21	He remembered sitting on the wall with a cousin, watching the German bomber fly over [177]	He remembered sitting on the wall with a cousin, watching the German dancers fly over	German
22	Perhaps no ballet has ever made the same impact on dancers and audience as Stravinsky's "Rite of Spring [177]	Perhaps no ballet has ever made the same impact on bomber and audience as Stravinsky's "Rite of Spring	bomber

5.4.3.2 Brown News Corpus (BNC) and Enron Email Corpus (EMC)

To evaluate the performance of our solution approach we collect results for all 740 and 1191 sentences from BNC and EMC datasets respectively. Table 5.8 reveals an accuracy of 77.4% (573 out of 740 sentences) for BNC and an accuracy of 62.9% (629 out of 1191 sentences) for EMC. "NA" denotes the number of sentences where the concepts present in bag-of-terms are not good enough to identify an obfuscated term (bag-of-terms length <2). Table 5.9 shows some concrete examples of these sentences from BNC and EMC

Table 5.8: Accuracy Results for Brown News Corpus (BNC) and Enron Mail Corpus (EMC)

	Total Sentences	Correctly Identified	Accuracy Results	NA
BNC	740	573	77.4%	46
EMC	1191	629	62.9%	125

Table 5.9: Concrete Examples of Sentences Present in our Experimental Dataset, Size of Bag-of-terms (BoT) < 2

Corpus	Sentence	BoT	Size
BNC	That was before I studied both	[]	0
BNC	The jews had been expected	[jews]	1
BNC	if we are not discriminating in our cars	[car]	1
EMC	What is the benefits?	[benefits]	1
EMC	Who coined the adolescents?	[adolescents]	1
EMC	Can you help? his days is 011 44 207 397 0840 john	[day]	1

datasets. Table 5.8 also reveals that for BNC dataset our tool outperforms the EMC dataset with a difference of 14.5% in overall accuracy. The reason behind this major fall in the accuracy is that the sentence written in Enron e-mail corpus are written in a formal manner containing several technical terms and abbreviations. These abbreviations are annotated as nouns in part-of-speech tagging and do not exist in commonsense knowledgebase. Table 5.10 shows some concrete examples of such sentences. Table 5.10 also reveals that there are some sentences that contain both abbreviations and technical terms. Furthermore, the length of bag-of-terms for those sentences is either too small (<2) or too large (>6). Therefore, either there is no term in the bag-of-words to compare and compute the conceptual similarity in commonsense knowledgebase or there are too many terms normalizing the similarity scores making it difficult to identify the out-of-context word. Whereas, the sentences written in Brown News Corpus are a combination of both formal and informal sentences. As discussed above in Section 5.1.1, the ConceptNet is a knowledgebase consisting of the sentence written by random users and consists of informal sentences and writing. Therefore, a majority of terms present in Brown News Corpus are found in commonsense knowledgebase. Due to the presence of these terms, we are able to compute the similarity score and identify out-of-context term in Brown News Corpus with a better accuracy than Enrom email corpus. Experimental results reveals that our approach is effective and able to detect obfuscated term correctly in long sentences containing more than 5 concepts in bag-of-terms. Table 5.11 shows some examples of such sentences present in BNC and EMC datasets.

We believe that our approach is more generalized in comparison to existing approaches. Word obfuscation detection techniques proposed by Deshmukh et al. [178] Fong et al. [176] and Jabbari et al. [177] are focused towards the substitution of first noun in a sentence. The bag-of-term approach is not limited to the first noun of a sentence. We use a bag-of-terms approach that is able to identify any term that has been obfuscated.

Table 5.10: Concrete Examples of Sentences with the Presence of Technical Terms and Abbreviations.

Sentence	Tech Terms	Abbr
#4. artifacts 2004-2008 maybe 1 trade a day.	Artifacts	-
We have put the interview on IPTV for your viewing pleasure.	Interview, IPTV	IPTV
Will talk with KGW off name.	-	KGW
We are having males backtesting Larry May's VaR.	backtesting	VAR
Internetworking and today American Express has surfaced.	Internetworking	-
I do not know their particles yet due to the Enron PRC meeting conflicts.	Enron	PRC
The others may have contracts with LNG consistency owners.	-	LNG

Table 5.11: Concrete Examples of Long Sentences (Length of Bag-of-terms >= 5) Where Substituted Term is Identified Correctly

Corpus	Sentence	Original	Bag-of-Terms
BNC	He further proposed grants of an unspecified input for experimental hospitals	Sum	[grants, unspecified, input, experimental, hospitals]
BNC	When the gubernatorial action starts Caldwell is expected to become a campaign coordinator for Byrd	Campaign	[gubernatorial, action, Caldwell, campaign, coordinator, Byrd]
BNC	The entire arguments collection is available to patrons of all members on interlibrary loans	Headquarters	[entire, argument, collection, available, patron, member, interlibrary, loan]
EMC	Methodologies for accurate skill-matching and pilgrims efficiencies=20 Key Benefits ?	Fulfillment	[methodologies, accurate, skill, pilgrims, efficiencies, benefits]
EMC	PERFORMANCE REVIEW The measurement to provide feedback is Friday November 17.	Deadline	[performance, review, measurement, feedback, friday, november]

5.4.3.2.1 Minimum Average Conceptual Similarity (MACS) Score

Figures 5.7 and 5.8 shows the minimum average conceptual similarity (MACS) score for Brown news corpus and Enron e-mail corpus respectively. Figure 5.7 also reveals that using Dijkstra's algorithm, majority of the sentences have mean average path length between 2 and 3.5. For Shortest path algorithm one-third of sentences have mean average path length between 1 and 2. That means in shortest path metrics, we find many directly connected edges. In Figure 5.7, we also observe that for half of the sentences, Dijkstra's and shortest path algorithms have similar MACS score. If two concepts are not reachable, then we use 4 as a default value for no-path. MACS score between 4.5 to 6 shows the absence of path among concepts or a relatively much longer path in the knowledgebase. Figure 5.7 reveals that for some sentences A* algorithm has a mean average path length between 4 and 6. Figure 5.8 illustrates that using A* and Dijkstra's algorithm, majority

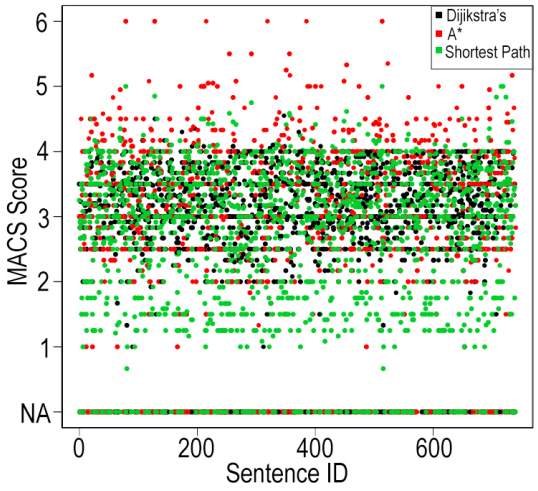


Figure 5.7: MACS Score of Concepts for Each Sentence for Brown News Corpus

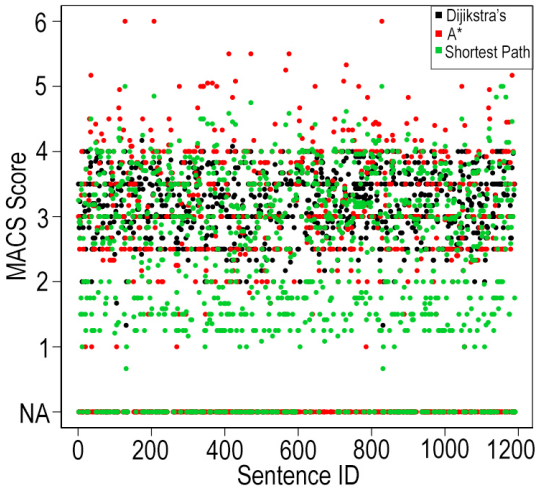


Figure 5.8: MACS Score of Concepts for Each Sentence for Enron E-mail Corpus

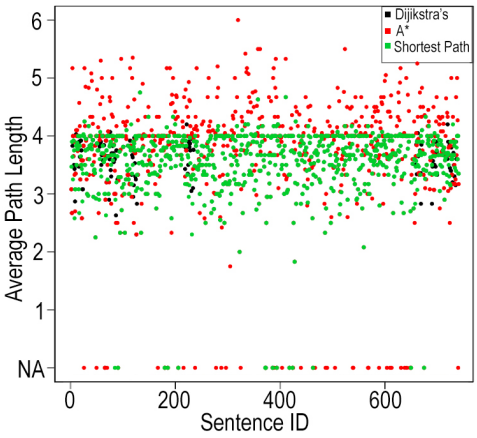


Figure 5.9: Average Path Length of Concepts for Each Sentence for Brown News Corpus

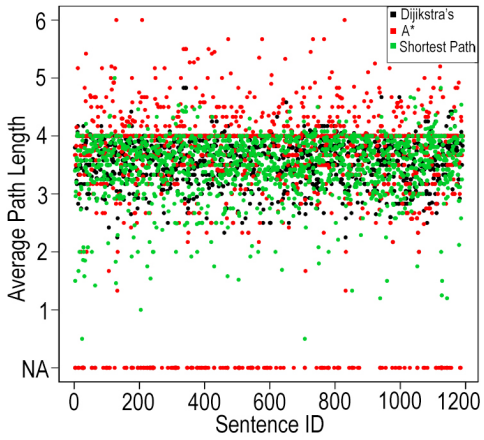


Figure 5.10: Average Path Length of Concepts for Each Sentence for Enron E-mail Corpus

of sentences have a mean average path length between 3 to 4. It shows that for many sentences bag-of-terms have concepts that are conceptually unrelated. It happens because Enron e-mail corpus has many technical terms that are not semantically related to each other in the commonsense knowledgebase. Similar to Brown news corpus, we observe that for half of the sentences, shortest path algorithm has a mean average path length between 1 and 2. Whereas, for Enron e-mail corpus, A* algorithm has higher MACS score for very few sentences. It reveals that either the concepts are connected by one or two nodes in between, or they are not connected at all (no-path).

5.4.3.2.2 Average Path Length Score

Figures 5.9 and 5.10 shows the average path length between concepts for each sentence present in the BNC and EMC datasets respectively. Figure 5.10 reveals that for Dijkstra's and shortest path algorithms, 80% sentences of brown news corpus have same average path length. Also, the majority of sentences have an average path length between 2.5 and 3.5. Similar to Figures 5.7 and 5.8 "NA" denotes the sentences with insufficient number of concepts. Figure 5.9 also reveals the presence of obfuscated term in the sentence. Since no sentence has an average length of 1 and similarly, only 1 sentence has an average length of 2. It implies the presence of terms that are not conceptually related to each other. Figure 5.10 shows that majority of sentences have average path length between 2.5 and 4 for all three distance metrics. Figure 5.10 also reveals that for some sentences shortest path algorithm has average path length between 0.5 and 2. Figure 5.10 shows that for some sentences all three algorithms have average path length between 4 and 6. It happens because of the presence of a few technical terms and abbreviations. These terms have no path in ConceptNet 4.0 and therefore are assigned a default value of 4.0 which increases the average path length for the bag-of-terms.

5.5 Threats to Validity and Limitations

The proposed solution approach for textual or term obfuscation detection uses ConceptNet knowledgebase for computing the conceptual and semantic similarity between any two given words. We use version 4.0 of ConceptNet, and the solution result is dependent on the nodes and the relationships between the nodes in the particular version of the ConceptNet knowledgebase. Hence a change in the version of the ConceptNet may have some effect on the outcome. For example, the number of paths between any two given concepts or the number of edges in the shortest path between any two given concepts may vary from one version to another. One of the limitations of our approach is that as the size of the bag-of-terms increases, the number of times the function to compute the shortest path between two nodes (and hence the overall computation time) increases substantially.

5.6 Conclusions

We present an approach to detect term obfuscation in adversarial communication using ConceptNet commonsense knowledgebase. The proposed solution approach consists of identifying the out-of-context term in a given sentence by computing the conceptual similarity between the terms in the given sentence. We compute the accuracy of the proposed solution approach on three test datasets: example sentences from research papers, Brown news corpus, and email news corpus. Experimental results reveal an accuracy of 72.72%, 77.4% and 62.0% respectively on the three dataset. Empirical evaluation and validation shows that the proposed approach is effective (an average accuracy of more than 70%) for the task of identifying obfuscated term in a given sentence. Experimental results demonstrate that our approach is also able to detect term obfuscation in long sentences containing more than $5-6$ concepts. Furthermore, we demonstrate that the proposed approach is generalizable as we conduct experiments on nearly 2000 sentences belonging two three different datasets and diverse domains.

Chapter 6

Social Media as Human Sensors for Forecasting Civil Protests

6.1 Introduction

Civil unrest or civil disobedience is referred as a social instability and protest movements at the National and International level primarily against the government and policymakers [186]. Civil unrest can be both non-violent demonstrations or strikes as well as violent riots. The reason behind large-scale civil unrest is mainly discontent in the society due to the poor social and economic conditions [187]. The Arab Spring democratic uprising which originated in Tunisia in December 2010[1] and then propagated across various countries in the Arab world in the year 2011 is an example of intense civil unrest and disorder[2]. Similarly, in a recent incident (February 1, 2017) a violent protest took place at Berkeley University, California causing 100,000 USD worth of damage to the university campus[3]. Due to high reachability and popularity of social media websites worldwide, organizations use these websites for planning and mobilizing events for protests and public demonstrations [188]. The study of civil unrest reveals that now most of the protests are planned and mobilized in much advance [187] [189]. Traditionally, newspapers have been used as primary sources for such analysis and prediction. However, the speed and flexibility of publication on social media platforms gained the attention of various organizations for planning and making announcements of various protests, strikes, public demonstrations, and riots. Organizing such large-scale civil protests requires planning and mobilization, and it is seen that the due to its immense popularity and wide reachability, Twitter is being used as a platform for sharing information about civil protest events and mobilize them [187] [190]. Figure 6.1 shows some concrete examples of tweets published for planning and mobilization of several civil protests and unrest related events. Due to the presence of public posts about mobilization, an early prediction of such events can be done by applying OSSMInt on Twitter data. In countries like USA, India, and Australia where protests are legal, early detection or forecasting of such events is valuable for government, tourism and law enforcement agencies. For example, it can help police deployment in those areas to maintain law and order and prevent violence. Similarly, it can help the government to deploy local civil force to retain the route traffic and prevent the interactions between protesters and bystanders [187].

We conduct a literature survey in the area of mining social media for prediction of civil unrest events. Based on our review, we find that while there has been a lot of work done in the area of general-interest event forecast; the prediction of civil unrest and protest events has recently gained the attention of researchers

[1] http://middleeast.about.com/od/humanrightsdemocracy/tp/Arab-Spring-Uprisings.htm

[2] http://www.npr.org/2011/12/17/143897126/the-arab-spring-a-year-of-revolution

[3] http://edition.cnn.com/2017/02/01/us/milo-yiannopoulos-berkeley/

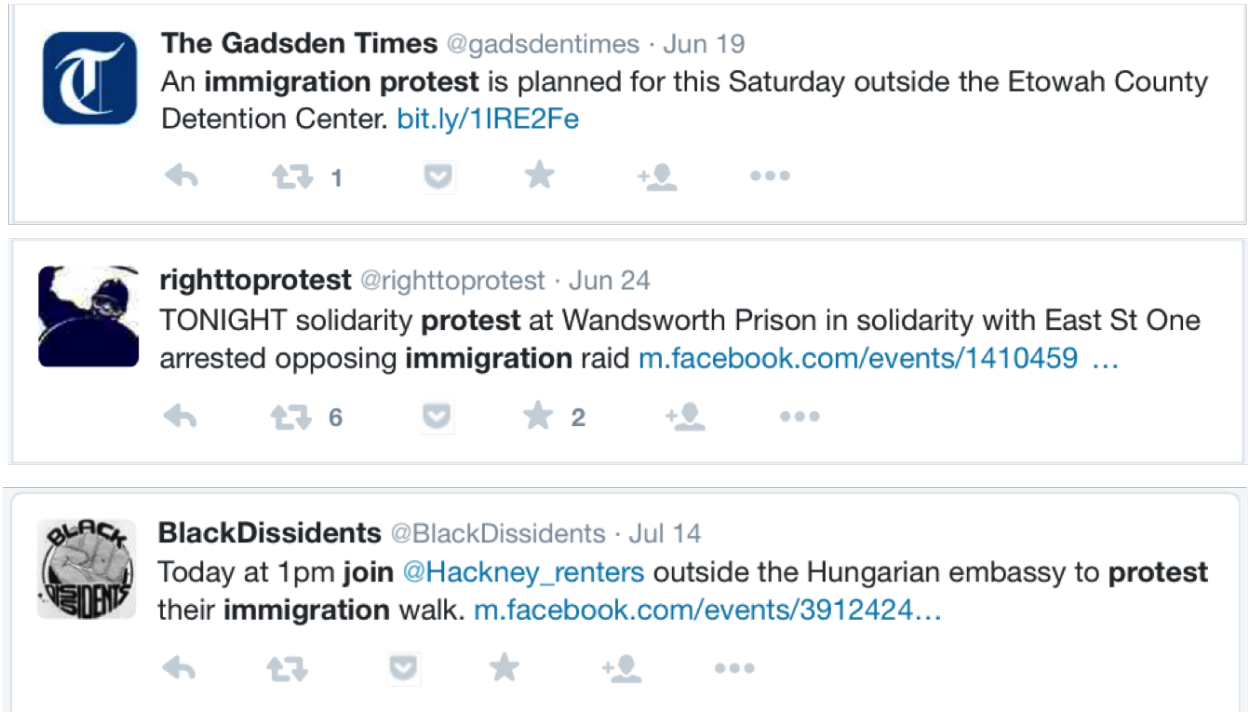


Figure 6.1: Concrete Examples of Posts Discussing About the Mobilization and Planning of Immigration Related Events on Twitter

[187] [190] [191]. We discuss a summary of closely related literature to the civil unrest event prediction performed on various micro-blogging platforms such as Twitter and Tumblr. Naren et. al. [187] build a sequential probabilistic model based data analytics platform (called as EMBERS) that mine the data from various sources (tweets, news, blogs, web search, and Wikipedia) and generate warnings for civil unrest related events across 10 countries of Latin America. In extension to the study presented in Naren et. al. [187], Muthiah et. al. [188] describe planned protest model- one of the probabilistic models used in EMBERS for civil protest forecast. They propose a keyword based flagging and probabilistic model approach to learning about the date and location of the event. Zhao et. al. [190] use a dynamic query expansion technique and propose a local modularity spatial scan (LMSS) algorithm to identify general-interest and targeted domain events. Compton et. al. [191] propose a method for early detection of events (Latin America civil unrest, sports, public functions) based on the direct extraction of relevant and highly important tweets. In their proposed approach, they focus on four key phases: keyword based flagging to find relevant tweets, mention of future dates, event geocoding for identifying the location and logistic regression method to classify tweets for event detection. In extension to the study presented in Compton et. al. [191], Xu et. al [192] conduct the similar study on Tumblr website data for predicting civil disobedience related events. Hua et. al. [186] present an approach to collect tweets related to an event of interest instead of predicting events from tweets. They use a keyword based flagging approach and connect the dots between news reports and tweets based upon topic keyword matching. They divide their approach into two steps and first identify the key terms related to an event and the topic being discussed in the news. Using ranking algorithm, they find 200 top ranked topic keywords and in the second step of their approach they collect tweets published around the event date and containing those topic keywords.

Based on the literature survey, we find that existing studies use a variety of machine learning and data

mining techniques for predicting upcoming protest events. However, the prior research is conducted on the data or tweets posted during the event or after the event has happened increasing the possibility of bias in the dataset. Further, due to the high velocity and massive size of data uploaded on social media data, mining each and every post for building predictive model impedes the performance of the model. Motivated by the previous research and gaps, our aim is to address the challenge of noisy content present in the real time stream and building a model for investigating the potential of Twitter data as an open-source precursor for anticipating and predicting civil protests.

6.2 Research Contributions

In context to the existing work, the study presented in this chapter makes the following novel contributions:

1. We perform a content-based characterization and semantic enrichment on raw tweets to classify crowd-buzz & commentary and mobilization & planning microposts related to a given a protest or civil disobedience.[**Characterization**]
2. To the best of our knowledge, our work is the first case study on immigration dataset that presents a frequency based model for an early forecasting of events. We investigate the application of trend analysis (captured along the sliding window) for early detection of an event. [**Predictive Modeling**]
3. We conduct two case studies on the Twitter dataset by defining two events (Christmas- Island Hunger Strike- January 16, 2014 and Fast for Families hunger strike- December 12, 2013) and examine the effectiveness of our proposed solution approach. [**Validation**]

6.3 Proposed Solution Approach

In this Section, we present the general research framework and methodology of our proposed approach for an early prediction of civil unrest related events. We use Twitter microblogging website as a data source for conducting our experiments. Figure 6.2 illustrates the high level design and architecture of proposed method that primarily consists of three phases: Data Collection, Semantic Enrichment and Event Forecasting as labeled in the solution framework. We discuss all three phases in the following subsections.

6.3.1 Experimental Setup

In order to conduct our experiments, we download an open source Twitter dataset available on Southampton Web Observatory[4]. Online Web Observatory is global data resource created for the Web Science research community [193]. The Web Observatory facilitates a distributed archive of data as well as the tools and mechanism to explore the evolution of web observatory. As discussed in Section 1.2.2, there are several web observatories created by large social media analytics groups from various universities. In the work presented in this Chapter, we conduct our experiments on the data acquired from one of such web observatories organized by Southampton University. We download *Immigration Tweets*[5] dataset from the observatory consisting of approximately 2 millions of tweets spanning in a time duration of 5 months (from October 1, 2013 to February 28, 2014). As illustrated in Figure 6.2, this dataset is stored in MongoDB format. We query the web observatory data from our local terminal and convert the dataset from MongoDB to MySQL

[4] `https://web-001.ecs.soton.ac.uk/new/datasets`

[5] `https://web-001.ecs.soton.ac.uk/datasets/kxwiPxvKLXSEWkgsW`

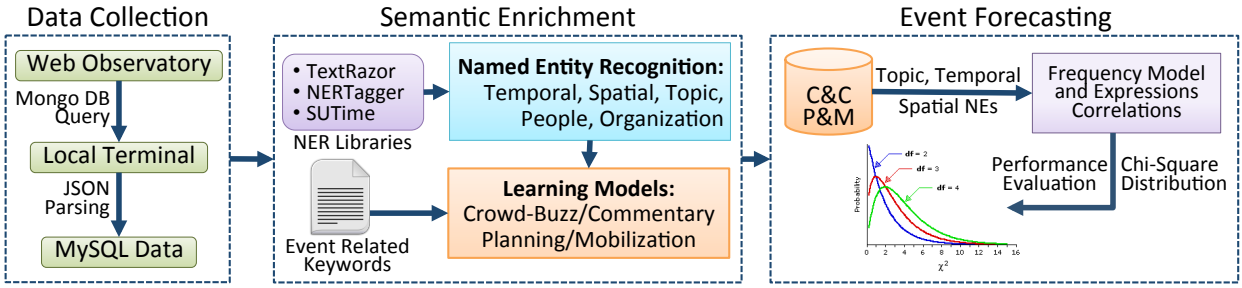


Figure 6.2: A High-level Demonstration of Proposed Research Framework Primarily Consisting of 3 Phases: 1) Acquiring "Immigration Tweets" Data from Online Web Observatory, 2) Performing Semantic Enrichment on Tweets and Extracting Crowd-buzz & Commentary and Planning & Mobilization Tweets, and 3) Building A Frequency and Expression Correlation Based Model for Early Forecast of Civil Protests and Unrest Events.

(a) Available Locations of User Profiles Posting Tweets About Event E1

(b) Available Locations of User Profiles Posting Tweets About Event E2

Figure 6.3: Distribution of Locations of Users Discussing Event Related Tweets Collected During the Sliding Window Time Frame (7 Days)

by parsing the response given in JSON format. In the present study, we conduct experiments on English language tweets only. Therefore, we identify the language of the tweet posts using Java Language Detection Library[6] and filter all records identified as non-English language tweets. The downloaded dataset is a collection of posts where each tweet consists of at least one of the following words: 'immigration', 'migration', 'immigrant' and 'migrant'. Since this dataset is restricted to 'immigration' related tweets, we identified the civil protest events related to immigration and happened during the period of data collection. We search popular news media and articles and find 2 such events. We provide a brief background about these events in the following subsections:

[6] `https://code.google.com/p/language-detection/`

Table 6.1: A Sample of Related Keywords for Events 1 ('Fast for Families' at National Mall, USA) and 2 ('Christmas Island Hunger Strike' at Australian Detention Center)

Event 1	Event 2
Washington, ActFast, Timeisnow, Greencard, National Mall, Fast4Families, Breakfastnews, CIR, Hunger, Families, Mayor	Border, Hunger, Detention, Eyelids, DentalFloss, Deportation, Obama, Sew, Refugees, Suicide, Israel, Children

6.3.1.1 E1- Fast for Families Hunger Strike, USA

A group of 4 advocates (Eliseo Medina, Dae Joong Yoon of NAKASEC, Lisa Sharon Harper of Sojourners, and Cristian Avila of Mi Familia Vota), all from different backgrounds set a tent in National Mall of USA on [**December 12, 2013**] and protested against new immigration reform bill[7]. Later, in a month, 200 ordinary people joined them in the protest and approximately $10,000$ people fasted across the country. This protest took place to show the urgency of new immigration reform for American families.

6.3.1.2 E2 Christmas island hunger Strike, Australia

The protest was initially sparked on [**January 16, 2014**] by the separation of some asylum seekers (almost 2000) from family members[8]. In this protest, nine Iranian men stitch up their mouth with dental floss and threaten to sew up their eyes. This protest took place at Australian Detention Center.

In our proposed approach, we create a model for early detection of a civil unrest event, and hence we conduct experiments on the tweets posted before the events happened. We perform a manual analysis on Twitter and observe that the organizations planning for such protest or demonstrations mostly start publishing tweets a week before the protest date. Therefore, for each event, we use trend based sliding window of 7 days and extract tweets that are published during that time frame. For example, December, 5 to 11 for Event E1 and January, 9 to 15 for Event E2. In order to minimize the bias in our dataset, we extract the locations of users who posted the tweets present in our experimental dataset. Figures 6.3a and 6.3b shows the available locations of users who posted tweets related to Event 1 (December 5 to 11) and Event 2 (January 9 to 15) in 7 days time frame before the events happened. Figures 6.3a and 6.3b reveal that the dataset selected by sliding window is posted by the users belonging to different locations across the world and not specific to the locations of events- USA and Australia. The diversity of user location shows that our experimental dataset contains no bias. For Event E1 and E2, we were able to extract a total of 527 and 428 user locations respectively; among which 74 and 66 locations were discarded due to the presence of meaningless or non-informative locations. For example, 'YouTube', 'Facebook', 'Where hope floats' and 'Anywhere and Everywhere'.

6.3.2 Lead Indicator Classifiers

In this phase of proposed method, we perform various pre-processing and classification techniques on raw tweets and make them semantically rich to achieve better accuracy in event prediction. Based on our inspection of event related posts on social media platforms, we create a hypothesis that there are two types of posts that are shared before the event takes place.

[7] https://www.americanprogress.org/issues/immigration/news/2013/12/12/81112/notes-from-the-fast-for-families-tent/

[8] https://www.theguardian.com/world/2014/jan/16/christmas-island-hunger-strike-spreads-to-family-compound

@Refugees how many asylum seekers are on hunger strike and other forms of self harm protest in Australian immigration detention centres?

http://t.co/ABrATCsPNJ Several detainees at the Christmas Island immigration detention centre have sewn their lip... http://t.co/L4nenwD7YW

@MinoWarrior @SherronShabazz We met with the organizers of the African migrant protests today in Tel Aviv

Topic Location Temporal

Figure 6.4: Examples of Civil Unrest Related Tweets Annotated with Temporal, Topic and Location Based Expressions

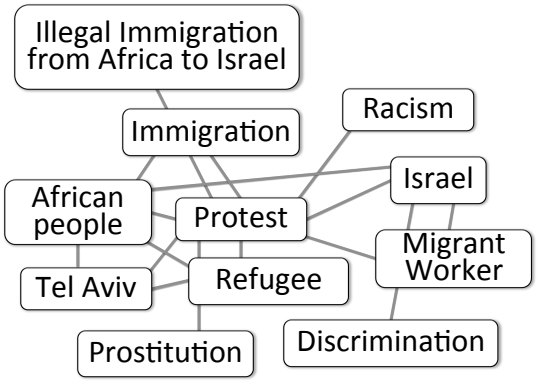


Figure 6.5: An Example of Semantic Relations Between Locations and Topics in Event Related Tweets

1. **CrowdBuzz and Commentary (C&C):** In such posts people discussing about the event and spreading the word among other people who might participate in the event.
2. **Planning and Mobilization (P&M):** Group of people organizing the protest and sharing the posts for mobilizing and planning the event.

Due to the high velocity and massive size of data being posted on Twitter in real time, it is highly likely to have noisy and irrelevant posts that are not related to the particular event. Based on this hypothesis, we train a multi-class classifier that classifies tweets into above mentioned three categories. We use C&C and P&M tweets as two leading indicators (signals which help in early detection) for forecasting an event. These indicators filter relevant tweets to our problem of civil unrest event forecasting from all other tweets and hence reduces the traffic and extra computation of irrelevant tweets. We discuss the classification methods of these indicators in the following subsections:

6.3.2.1 CrowdBuzz and Commentary Tweets Classification

As discussed above, the crowdbuzz and commentary tweets are the posts that discuss the topic of the event. Therefore, for classifying such tweets, we identify the presence of pre-defined terms (lexicon-based approach) that are relevant to the event which is being monitored. We initially create a list of words that are important and commonly used in the events information. We further apply a bootstrapping method and extend the list by extracting hashtag and keywords from tweets posted in seven days sliding window time frame. We extract event related information by mining "Google News" media websites and articles related to the event and further enrich our list by using TF-IDF (Term Frequency-Inverse Document Frequency) based approach. Table 6.1 shows the list of related keywords for event 1 (Fast For Families- USA) and event 2 (Christmas Island Hunger Strike, Australia). If a tweet present in the selected time frame of the event contains any of the event-related keywords, then we classify the tweet as crowdbuzz and commentary tweet.

6.3.2.2 Planning and Mobilization Tweets Classification

As discussed above, the planning and mobilization tweets are used to organize and plan an event and hence contains the information about the location and time information about where and when the event is going to take place. Therefore, for classifying P&M tweets, we propose to extract the spatiotemporal

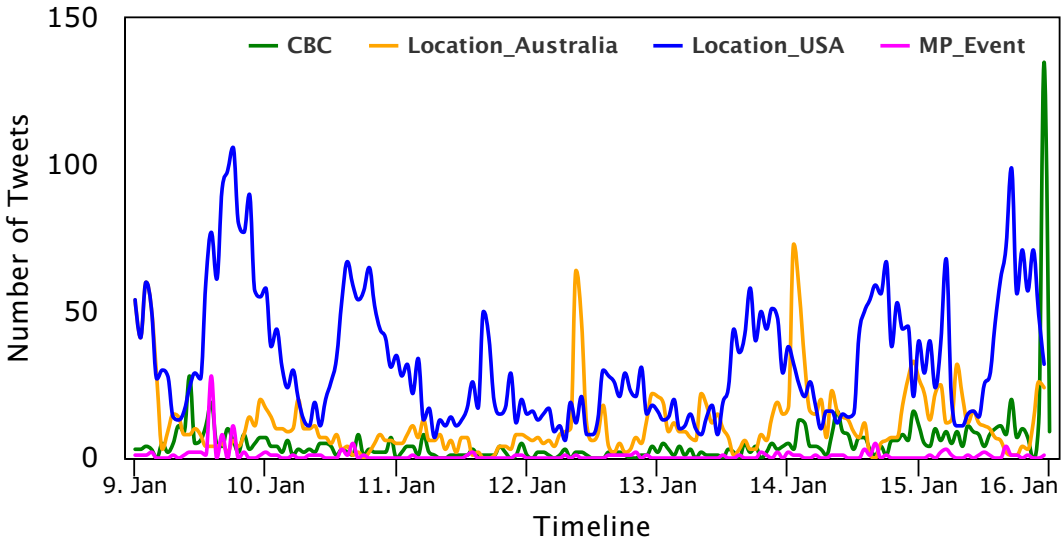


Figure 6.6: Trend of Crowd-Buzz, Mobilization and Location Based Tweets Posted in 7 Days Before Christmas Island Hunger Strike (Event 2)

feature from tweets. We also observe that P&M tweets contain reply or direct mention to other users ['@' character followed by a Twitter username] due to the exchange of messages between users for coordinating their activities. We extract the person, location and temporal features from the tweets using an ensemble based learning on a combination of Named Entity Recognizers. We use three open source NER APIs for the purpose of features extraction: Java NER[9], SUTime[10] and TextRazor[11]. We observe that due to the presence of free-form text and user-generated data, there is no defined structure of the tweets and hence contain inconsistency. For example, while some tweets mention the date of the event, some tweets contain the day information while some tweets mention future terms (tomorrow or day after tomorrow) for temporal information. Therefore, to maintain the consistency in temporal expressions, we convert the extracted expressions into the day of the week based upon the timestamp of the original tweet. For example- if a tweet contains the entity 'tomorrow' and is posted on 'Wednesday, January 15, 2014', then we convert it into 'Thursday'. Another feature that we come across to is that the tweets posted for planning and mobilizing consist of phrases like *join us, spread the word* and future tense related words. We annotate 500 P&M tweets and train a machine learning classifier for filtering the P&M tweets. We discard the remaining (not identified as C&C or p&M) as irrelevant tweets. Figure 6.4 shows concrete examples of location, temporal and topic expressions in tweets related to civil protest events.

Figure 6.6 illustrates the trend of tweets being posted in every hour for 7 days sliding window time frame. Figure 6.6 reveals that in comparison to the total number of tweets posted in a time frame, only a very small fraction of those tweets are C&C and P&M which increase the significance of filtering these posts. Figure 6.6 shows that the number of C&C tweets has similar pattern for five days before the event happened and sudden peak for last two days, unlike P&M tweets which are higher in initial days of planning. We also observe that the event happened in Australia. However, the maximum number of tweets during this time frame are posted from various locations of USA. This trend shows that it is not a good idea to predict the location of an event based upon the user profile locations with the maximum number of tweets posted.

[9] `http://nlp.stanford.edu/software/CRF-NER.shtml`
[10] `http://nlp.stanford.edu/software/sutime.shtml`
[11] `https://www.textrazor.com/docs/java`

Figure 6.7: A Word-Cloud Presentation of 48 Topics Discussed in Tweets in 7 Days Sliding Window Selected for Event E2

6.3.3 Event Forecasting Model

We define the problem of civil unrest event forecasting as applying OSSMInt for extracting three primary insights (TTL) from the event-related tweets: Temporal expression Ti- time or date of the event, Spatial expression lo- location of the event, and Topic expression To- root cause or objective of the protest. As shown in Figure 6.2, in order to build an event forecasting model, we take an input of crowdbuzz-commentary and planning-mobilization tweets that are enriched with TTL information extracted using ensemble based named entity recognizers and TF-IDF approaches. In our proposed approach, we implement an adaptive version of the algorithm proposed in Budak et. al. [194]. In addition to the spatiotemporal features proposed in Budak et. al. [194], we add topic as another discriminatory feature for event prediction. Figure 6.7 shows the name of all topics identified from the tweets present in our experimental dataset of Event 2 (Christmas Island Hunger Strike, Australia) sampled for a time frame of 7 days of sliding window.

A tweet might contain more than one named entity expressions. Therefore, we define a bipartite graph arrangement of all location, time and topic expressions. A bipartite graph is an arrangement of vertices (V) and edges (E) where V is split into two disjoint sets of vertices $V1$ and $V2$. An edge in a bipartite graph can be represented as $e \in E$=$\{v_1, v_2\}$ where $v_1 \in V1$ and $v_2 \in V2$. For each topic $To_i \in To$, location $Lo_j \in Lo$ and temporal expression $Ti_k \in Ti$, we create distinct set of all possible combinations of expressions pairs: location-topic $<Lo_j, To_i>$, topic-time $<To_i, Ti_k>$ and location-time $<Lo_j, Ti_k>$. We use these named entities pairs to compute the correlation among a location, time and topic. To keep only significant results, we select only those expressions that are at least θ frequent (occurs at least in the θ number of tweets) where θ varies for all 3 entities. We define θ for each entity as the average number of named entity expressions present in the dataset (only C&C and P&M tweets). Figure 6.5 reveals the significance of selecting θ frequent entities for event prediction. The network graph demonstrated in Figure 6.5 shows a sample of topic entities collected from the C&C and P&M tweets for Event E2, posted on the first day of the sliding window. Figure 6.5 reveals that the terms 'protest', 'refugee', 'African people', 'immigration', 'Tel Aviv' are the terms which are significantly correlated and highly frequent in tweets in comparison to 'racism' or 'prostitution'. To find these significant pairs, we extract all expressions of x entity that are at least θ frequent in the dataset

$$F(x) > \theta$$

and further extract all those expressions of paired entity y that are ψ frequent for at least one x and vice-versa.

$$F(x,y) > \lceil \psi F(x) \rceil \qquad F(y,x) > \lceil \varphi F(y) \rceil$$

Here, F is the frequency of an entity expression in the dataset. We select the pairs of named entities which satisfy the above conditions, and it reduces the number of pairs for further examination. We compute the

Table 6.2: Confusion Matrix for Crowd-Buzz Tweets Classification and Planning and Mobilization Tweets Classification of Events 1 (Fast For Families in National Mall, US) and 2 (Christmas Island Hunger Strike)

(a) Event 1

		Predicted	
		C&C	**NA**
Actual	**C&C**	1,322	524
	NA	362	77,223

(b) Event 2

		Predicted	
		C&C	**NA**
Actual	**C&C**	719	109
	NA	137	82,007

(c) Event 1

		Predicted		
		MPE	**MPG**	**NA**
Actual	**MPE**	140	87	47
	MPG	6	2,028	187
	NA	4	218	77,495

(d) Event 2

		Predicted		
		MPE	**MPG**	**NA**
Actual	**MPE**	127	6	9
	MPG	17	1,236	22
	NA	13	18	81,541

frequency of each pair in a bipartite manner and only look for the pairs that are highly correlated and has no decrement in the frequency in 7 consecutive days of the sliding window. Since these named entities are categorical attributes, we use Chi-Squared distribution to find the correlation between two entity expressions. We define a pair of expressions to be significantly correlated if their p-value$<$0.05 for their respective χ^2 and degree of freedom.

6.4 Experimental Results

6.4.1 Crowd-buzz and Mobilization Tweets Classification

To evaluate the performance of our proposed approach, we use standard measures of information retrieval and machine learning. We asked 50 graduate students to manually annotate each tweet of the sliding window data. We provided them specific guidelines for annotating crowd-buzz, planning & mobilizing for the general event and domain specific event (Fast for Families Protest- December 12, 2013 and Christmas Island hunger strike- January 16, 2014). Based on their decisions we validate our results and compute accuracy of semantic enrichment phase using precision, recall, and F1-score. Tables 6.2 shows the standard confusion matrix for both C&C and P&M classifiers for Events $E1$ and $E2$. NA denotes the number of tweets classified as irrelevant or unknown. Table 6.2a reveals that out of $1,846$ tweets (2.3% of sampled tweets for Event 1- $79,431$) annotated as C&C, we were able to predict $1,322$ tweets correctly while $77,223$ tweets as irrelevant with a total misclassification of 1.1%. Similarly, Table 6.2b reveals that out of 828 tweets (less than 1% of sampled tweets for Event 2- $82,962$) annotated as crowdbuzz, we were able to predict 719 tweets correctly with a total misclassification of less than 0.5%. Tables 6.2c and 6.2d shows the confusion matrix for planning and mobilization tweets identification. As demonstrated in Tables 6.2c and 6.2d, since the presence of spatiotemporal features is not specific to civil unrest events; we not only identify the P&M tweets that belong to protest-related events but also any general events. For example, music concerts, seminars, conferences, official government event. Tables 6.2c and 6.2d reveal that for both events $E1$ (274 P&M tweets) and $E2$ (142 P&M tweets), we were able to correctly classify 140 and 127 tweets respectively. Table 6.3 shows the accuracy results of crowdbuzz & commentary and planning & mobilization classifiers

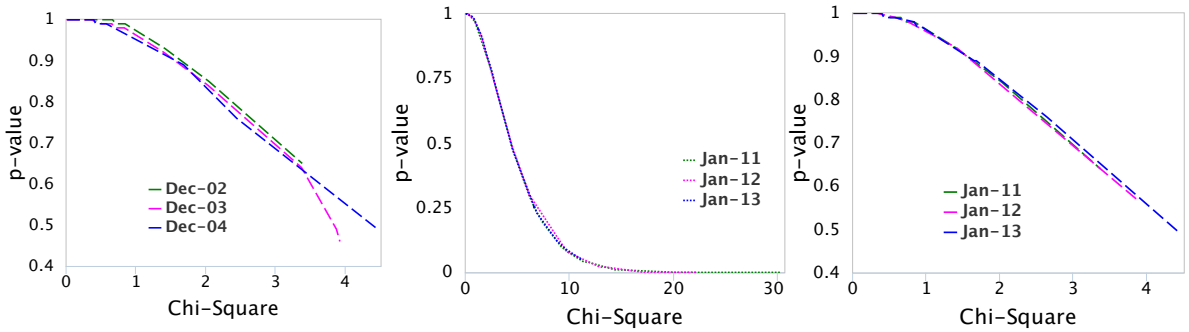
Table 6.3: Accuracy Results for Semantic Enrichment Classifiers (Crowd-Buzz and Mobilization Tweets)

(a) Fast For Families

Classifier	Precision	Recall	F1 Score
C&C	0.78	0.71	0.75
	0.51	0.93	0.66
P&M	0.91	0.86	0.89
	0.99	0.99	0.99

(b) Christmas Island Hunger Strike

Classifier	Precision	Recall	F1 Score
C&C	0.84	0.86	0.85
	0.89	0.81	0.85
P&M	0.96	0.98	0.97
	0.99	0.99	0.99



(a) US-F3P (b) Australia-CIHS (c) US-CIHS

Figure 6.8: Distribution of χ^2 and p-value for Frequent Pairs of Locations and Topics for 3 consecutive Days in Sliding Window- Christmas Island Hunger Strike (CIHS) and Fast For Families Protest (F3P)

for events E1 and E2. Table 6.3 reveals that the precision, recall, and f1-score for both the classifiers are reasonably high and we are able to classify crowdbuzz and mobilization tweets for events E1 and E2 with the f1-score of 0.75 and 0.85 respectively.

6.4.2 Chi-Squared Distribution

6.4.2.1 Event 1: Fast For Families

For "fast for families protest in national mall" event related tweets, we find 8 unique and non-overlapping topics, 3 unique locations and 3 unique temporal expressions that have a relatively high non-decreasing frequency in tweets posted during sliding window time frame (7 days). In order to identify the significant pairs of entities, we compute χ^2 and p-value distribution for each frequent pair. Chi-squared distribution (χ^2) is a statistical test applied to the categorical data to compute the frequency distribution of certain events observed in a sample data [195]. Here, the events are defined as the entities (spatiotemporal and topic entities). The p-value is computed as an evidence of the statistical model and defined as the probability of obtaining the results same or more extreme than the real observed results [195]. We compute χ^2 distribution of all frequent pairs and discard the ones that are less frequent than ψ threshold value. We further normalized all at least ψ frequent pairs of topic-time-location entity expressions between 0 and 1 (similar to the first event). Based upon the distribution, we observe that the pairs containing location "Washington (0.034)" has

very less p-value in comparison to "Europe (0.23)" and "California (0.014)". Figure 6.8a shows the variance between p-value of location "Washington D.C." and all frequent topics for three consecutive days (December 2 to 4).

6.4.2.2 Event 2: Christmas Island hunger strike

For "Christmas Island hunger strike" event, we find 10 unique topics, 6 unique locations and 5 unique temporal expressions in our dataset. Based upon the frequency model (discussed in Section 6.3.3) we find only 3 unique topics ('right of asylum', 'migrant worker' and 'refugee'), 4 unique locations ('Australia' and 'US') and 2 unique temporal expressions ('Wednesday' and 'Thursday') with no decreasing frequently in sliding window. We further compute p-value of each pair based upon the χ^2 and degree of freedom and observe that the pairs containing 'U.S.' have p-value greater than 0.05 while Australia has p-value to be very small (approx 0.000002). Figure 6.8 illustrates the variance between p-value of 'Australia' and 'US' for all topics for three days in sliding window (p-values are normalized between 0 and 1). Similarly, for (time, location) and (time, topic) pairs we keep highly correlated terms with significant p-values and are able to predict events with a high confidence score.

6.5 Conclusions & Future Work

In this work, we present an approach for early detection and prediction of civil unrest related events. We conduct experiments on real-world dataset (open source dataset downloaded from Southampton Web Observatory) consisting of tweets on immigration and migration. We use named entity recognition and term-frequency based approaches to capture various discriminatory features from raw tweets such as time, topic and location of the event. We perform single and multi-class classifications to filter event related tweets (crowd buzz and mobilization). Experimental results reveal the high accuracy of classifiers with an f1-score (0.75 and 0.85 for "Fast for Families protest" and "Christmas Island hunger strike" events respectively). We develop a frequency based model on these semantically rich tweets and find those pairs of location, topics, and temporal based named entities/expressions that are significantly correlated. We further compute the χ^2 and p-value distribution of these entities and their correlation over a time frame. From this distribution, we conclude that by detecting trend analysis of these entities in the tweets posted during a sliding window time frame, we can predict civil unrest related events with a high confidence score. We also conclude that early identification crowdbuzz and mobilization tweets are value added in event forecasting.

Future work includes the extension of proposed method for detecting protest related events in real time and events with an overlapping sliding window (multiple events occurring on the same day or in the same week). We plan to investigate the application of Ensemble Learning approach for event forecasting and semantic enrichment. Furthermore, future work includes the timeline-based visualization of tweets to provide end-users an updated and rich information (highlights or summary) about events in real time. Our future work also includes the early prediction of a protest or civil disobedience based on the analysis of a chain of similar incidents already happened in past or currently happening and tends to happen at another location or time for the similar root causes.

Chapter 7

Conclusions and Future Directions

We present conclusions and future work at the end of each of the five chapters (Chapters 2 to 5) describing five broad applications on open-source social media intelligence in government domain. In this chapter, we summarize our work, present general conclusions, specific conclusions already covered at the end of each chapter and provide future directions.

7.1 Abstract and General Conclusions

In this dissertation, we demonstrate how open-source social media intelligence (OSSMInt) can be applied in government domain. While there has been a lot of work done in the area of social media intelligence by private companies and private sector for developing consumer-based applications, our dissertation makes a unique contribution in the relatively unexplored area of OSSMInt in government and security informatics domain. We categorize our proposed applications into three broad classes: identification, prediction and response applications. We propose and implement five broad applications and usage scenarios spanning these three categories: (1) citizen complaints and grievance redressal [response] (2) religious conflict detection [identification] (3) hate and extremism promoting content, users and community detection [identification] (4) secret message and communication detection [identification] (5) early civil protest forecasting [prediction].

We formulate these applications as text analytics based problems and propose information retrieval (IR), machine learning (ML) and natural language processing (NLP) based solutions to develop these applications. However, due to the free-form nature of social media data, building such applications is a technically challenging problem. The presence of noisy content such as spelling errors, incorrect grammar, unstructured text, use of slang and abbreviations makes it challenging to develop applications without ignoring the false positives. Further, due to the presence of disguised emotions and ambiguity in the content, it's hard to label the data by human annotation as well. Our work demonstrates that government can gather a lot of intelligence from publicly available user-generated data on social media platforms. We show how government can automatically analyze and monitor social media data in real time for improving their decision making. The chapters in the dissertation are organized in the form of applications and case-studies and show how useful intelligence can be gathered without invading into anybodys privacy as open-source social media data is public. Hence one of the major and broad conclusions of our work is the following:

> There are several untapped opportunities as well as technical challenges in exploiting open-source social media data by the government for intelligence gathering.

We present general technical challenges [Chapter 1 - Introduction] which covers all the five broad applications described in this dissertation and also includes the unique technical challenges within the context of each application [respective chapters]. Another significant contribution of our thesis is a demonstration of how the open-source social media data can be systematically collected and processed by applying different IR, ML and NLP based techniques. Crawling, filtering, cleaning, annotating and pre-processing (unlike stopwords removal and text tokenization) social media data is non-trivial. We present solutions and approaches to enhance this that varies for different platforms. The text enhancement and enrichment of user-generated data includes language detection and translation in multilingual scripts, removal of non-informative content, expanding joint keywords and slang without affecting the problem specific terms and handling incorrect spelling and grammar.

Advanced multi-step pre-processing of social media data is mandatory before it can be used for analysis or statistical model building. There are no one-size-fits-all pre-processing tools, and each step in the pre-processing pipeline requires problem and domain specific solutions.

Despite having a lot of similarity in terms of the usability and reachability of the websites, several online social media (OSM) platforms facilitate unique features to their users. One of such features is the amount of text that can be uploaded to the website at once (micropost, short text or long text). NLP based techniques are the integrated part of building government applications by mining social media data. The performance of NLP-based approaches varies based on the amount of input text. The variation in length of each data input impedes the accuracy of linguistic features.

In addition to the enrichment of raw text, in order to make content identification approaches generalized, it is mandatory to identify platform independent features that are generalized for all sizes and structures (formal, informal, unstructured) of the input text.

OSSMInt in Security Informatics and government domain covers the applications that require experts to annotate the results and evaluate the performance of proposed methodologies. However, due to the high velocity and massive size of available social media data, creating ground truth for each record is overwhelmingly impractical. Further, the lack of annotated data impacts the performance of feature extraction and supervised learning based classification techniques. Based on our proposed approaches we conclude that the unsupervised, semi-supervised and active learning based approaches are able to learn the features on the way. The use of such classification techniques efficiently categorizes the content reducing the efforts taken into annotation of each record. Furthermore, due to the excess of irrelevant content on social media, the data is highly imbalanced, and a very small percentage (0.000001%) of records belong to the topic and relevant class (hate promoting, complaint reports, religion and race-targeted posts, discussion on protest and unrest events). Therefore, it is impractical to identify the class of each record and extract discriminatory features for them.

To identify the relevant content by applying OSSMInt in government domain, unsupervised and semi-supervised learning based techniques are reasonably effective. While, training the classifier on the features of targeted class (one-class classification, identifying relevant records and classifying other inputs as unknown) are practically feasible and reduce the noise in the classification.

OSSMInt based applications for government and law enforcement agencies include the identification of many such contents on social media that includes sensitive topics like religion, race and anti-national posts. While it is important for the security analysts to monitor each and every post that might lead to the violence

and unrest; it is also important to identify the intent and objective of the author before removing the content or account from social media. Due to the presence of ambiguity, references, and sarcasm, the use of shallow keyword-spotting techniques is discouraging and requires deeper analysis.

> Identification of authors' personality and psychological behaviors including social tendencies, emotions, social references, and writing style is mandatory for capturing the dependencies between content and intent of a post.

Online radicalization, planning of civil protests, conflicts of religious beliefs, online terrorism activities are not specific to only a one or a few countries or specific regions but are the topics of concern for the security analysts at a global level. Social media websites allow their users to post textual content in a variety of languages and scripts. Several groups and communities of like-minded people and individuals take leverage of such features and post ill-disposed content in the regional language of targeted audiences. Further, the users expressing their religious opinions and sentiments in their community on social media use their regional language for communications and discussions.

> In order to remove the bias from data collection and geographical independent analysis, it is mandatory to analyze the multilingual texts and scripts (by translating them into a base language) on social media and not just analyzing the content posted in the defined base language.

In addition to the users publishing sensitive and harmful content on social media, there are several users who do not actively post any malicious content on their channel but silently contribute to the community. Online social media platforms allow their users to see the different activity feeds of their followings and subscriptions varying for different platforms. For example, YouTube and Facebook allow users to see the *'like', 'favorite'* and *'comment'* activities unless kept as private. Therefore, users can disseminate information and spread the already published content among their followers without actively posting them on their channel.

> It is required to monitor and analyze the activity feeds of the users and not just the uploaded content on their channels for identifying the right communities and reduce the number of false negatives (silent contributors identified as naive bloggers).

7.2 Specific Conclusions

In addition to the general take aways from this thesis, we also discuss some of the specific conclusions of the work presented in each chapter of this dissertation.

Mining Twitter to Identify Citizen Centric Complaints and Grievances

1. In the first application of OSSMInt for government discussed in Chapter 2, we perform a case study on 4 different Twitter accounts of Indian public service agencies for identifying complaints and grievances of citizens. Based on our results, we conclude that computational linguistic and natural language processing-based techniques are an efficient method for identifying discriminatory features from complaints reports.

2. We also conclude that a prior identification of tweets that are certainly not the complaint reports (appreciation, information sharing, and promotional tweets) improves the accuracy of complaint tweet classification.
3. In Chapter 2, we also conclude that for the same features and training data, the accuracy of baseline SVM classification method is improved by ensembling different kernels functions of SVM. In the first case-study conducted in Chapter 2, we were able to boost the accuracy of our classifier up to 20%.
4. In addition to the identification of response-seeking based complaint reports, in Chapter 2, we also conduct a case-study on mining Twitter for identifying complaints on bad road conditions. Based on the performance of proposed approach, we conclude that identification of pinpoint location (or landmark), region (or city), and key issue reported in the complaints are the three primary features for identifying the road related complaints.
5. Based on our experiments conducted on the Twitter dataset, we also conclude that there are several complaints which are incomplete and do not contain all three primary features. However, based on the type of missing information (such as missing region or city information) these tweets can be enriched and used for further analysis.
6. We conduct a characterization study on the tweets identified as complaint reports and conclude that a maximum number of complaints are reported about the risky, dangerous and accident prone roads while most of them are due to the poor condition of amenities.

Mining Public Opinions on Tumblr for Identifying Religious Conflicts

1. In Chapter 3, we conduct a case-study on open source data of Tumblr website for identifying religious conflicts within society. Based on our experimental results, we conclude that social media is a rich source of information for identifying religious beliefs of people and mining these religious beliefs and sentiments on OSM websites fills the gaps of offline surveys for religious conflicts identification.
2. Based on our survey conducted among people from different groups (graduate students, Tumblr bloggers and users, and random individuals from society), we conclude that the presence of positive and negative sentiments in religion based posts do not reveal the religious beliefs of the author. But it requires deeper analysis of the context in terms of the emotions present in a post. For example, disgust, insult, disappointment, defensive, ashamed and disbelief.
3. In order to identify the religious conflicts on social media and fill the gaps of offline studies, it requires a significant number of posts from different groups of users. However, it is overwhelmingly impractical to annotate each post individually. In the work presented in Chapter 3, based on our experimental results we conclude that unsupervised and semi-supervised learning based approaches are the efficient techniques to process and large-scale data.

Identifying Extremist Content, Users, and Communities on Social Media

1. In the first case-study (Section 4.4) presented in Chapter 4, we conduct experiments on Twitter data for identifying hate promoting content posted in form of tweets. We propose to use an n-gram based model and investigate the efficiency of various features like presence of war-related terms, religious mention, negative emotions and use of swear & offensive words. Our results reveal that due to the sparse and highly imbalanced nature of dataset, SVM outperforms K-NN with a margin of 23% in F-score.
2. To investigate the efficiency of selected features, we apply leave-p-out strategy and measure the performance of classifiers. Our results reveal that while, the presence of war-related terms, negative emotions and swear words are discriminatory features for identifying extremist content, classification based on the presence of certain hate promoting hashtags impedes the performance of classifiers.

3. In the second case study (Section 4.5) presented in Chapter 4, we propose to use computational linguistic based approach for identifying the intent of the author in radicalized and racist posts. Based on our results, we conclude that shallow NLP and keyword spotting techniques are inefficient for determining the intent of the radicalized posts. Whereas, the use of author's psychological behavior such as personality traits, social tendencies, writing cues and semantic role of terms used in the post are strong indicators for identifying the radicalized or racist intent-based posts.
4. We conduct our experiments on open source data of Tumblr website, and our results reveal that prior identification of topics (such as religion and race) in the posts and filtering non-topic posts improves the accuracy of intent classification.
5. In addition to the identification of radicalized, racist posts, in Chapter 4, we also present our work on identifying hate promoting users and uncover their hidden and virtual communities on social media. We conduct our case studies on YouTube (Section 4.6) and Tumblr (Section 4.7) websites. We propose to use link analysis and topical crawling based approaches for navigating through the website and identify the channels and bloggers promoting extremist content on the website. Our results reveal that while conducting our experiments on YouTube, Shark Search based graph traversal algorithm outperform Best First search with a margin of 5% (accuracy 74%). Whereas, Random Walk based algorithm on Tumblr gives an accuracy of 77%.
6. We perform social network analysis on the extremist users on YouTube and Tumblr website for identifying the central users playing major role in the community. Our results reveal that in cluster-based networks of YouTube users, shark search navigation is able to extract large number strongly connected users and communities in comparison to best first search navigation. While based on the social network analysis performed on the Tumblr bloggers, we conclude that in comparison to the like feature, re-blogging is a strong indicator and a discriminatory feature to capture strongly connected communities on the website.

Detecting Word Obfuscation in an Adversarial and Secret Message Communication

1. In the work presented in Chapter 5, we propose a conceptual similarity based approach to detect term obfuscation in adversarial communication. To evaluate the performance of proposed methodology, we conduct our experiments on three different datasets: Enron email corpus (EMC), Brown news corpus (BNC) and the examples used in previous literature. Our experimental results reveal an accuracy of 72.72%, 77.4% and 62.0% for EMC, BNC and examples dataset respectively.
2. Our results reveal that computing the conceptual and semantic similarity between the terms is an effective method for identifying out-of-context words in the sentences. Empirical evaluation and validation show that commonsense knowledgebase is an efficient lexical resource for identifying the obfuscated terms in formal as well as informally structured sentences.
3. Based on our results, we conclude that the proposed approach is also able to detect term obfuscation in long sentences containing more than $5-6$ concepts. Furthermore, we conclude that the proposed approach is generalizable as we conduct experiments on nearly 2000 sentences belonging two three different datasets and diverse domains.

Mining Twitter Data for an Early Forecast of Civil Protest and Unrest

1. In the work presented in Chapter 6, we propose a trend analysis and time sliding window based approach for an early prediction of civil unrest and protest related events. We conduct our experiments on immigration-related tweets downloaded from publicly available data on Web Observatory. Based on

our experimental results, we conclude that identification of crowdbuzz & commentary and mobilization & planning tweets posted prior to the event date are lead indicators for predicting a protest event.

2. We propose a frequency-based model and compute the correlation between topic, spatial location and temporal features for identifying the significant pairs of entities that co-occur together over a time period (for example, from a week before the event). Our experimental results reveal that identification of such significant pairs in lead indicator tweets and analyzing their trend in sliding window frame improves the performance and confidence score of the predictive model.

7.3 Future Directions

The work presented in this dissertation is interdisciplinary as it at the intersection of fields such as computer science (particularly machine learning, information retrieval, natural language processing), security, social media, e-governance, e-participation, counter-terrorism and law enforcement. We propose the following future directions of our work which draws from multiple domains.

7.3.1 Integration with Other Domains

1. **Information Visualization-** The end-users of the proposed applications and usage scenarios in this dissertation are the government agencies and security analysts. It is important for them to not only able to build such applications that can detect relevant content on social media, but it is also important to visually analyze the current state-of-the-art in the form of dynamic and interactive visualizations [196]. We have not explored sophisticated visualization techniques and advanced topics in information visualization in our work. We believe that visual analytics and interactive visualization can further improve the value and usability of the applications presented in our work [197].
2. **Deep Learning-** We have applied several machine learning based techniques for statistical model building in our work. However, we have not used deep learning based methods which are recent and advanced developments in machine learning research and algorithms [198]. Deep learning has shown encouraging results in areas such as speech and image recognition, and we believe that the application of deep learning for solving problems defined in this dissertation can be interesting [199].
3. **Sensemaking-** The user-generated data on social media is written in free-form text and hence does not have a specific structure, vocabulary or defined terminology while reporting for complaint tweets. Further, due to the excessive use of sarcasm in the reports, the standard NLP techniques become shallow and are not able to capture the issues reported in the complaint reports. Our future work includes the application of sensemaking in identifying hidden and ambiguous complaints in micropost. Sensemaking is an area that comprises of developing theories, applying psychological research, developing interaction between human sensors and information technology and making sense of observed data [200] [201].

7.3.2 Novel Research Applications of OSSMInt for Government

1. **Intent vs. Impact-** In addition to the identification of posts and channels created with the intention of promoting hatred and a shared propaganda, we propose a novel use-case of OSSMInt in government domain. The aim of proposed application is to identify the posts and users who cause an extreme and grave impact on their viewers irrespective of the intent of the post. The proposed application not only detect the intent and content-based posts but it also aims to analyze the amount of attention received

on the post. The impact of such posts can be measured in terms of identifying the user engagement, discussion, and activities performed on and regarding the post.

2. **Detection of Online Recruitment in Radical Groups-** The aim of online radicalized communities on social media is to not only spread hatred and promote their ideologies, but it also includes recruiting young people in their groups. The aim the proposed application is to perform a real-time and dynamic analysis of the extremist communities on social media and capture the phenomena of variation in new participating nodes in the community and existing nodes leaving the community. We propose to use the advanced and interactive visualizations for analyzing the periodic changes in the community and users moving back and forth within multiple communities.
3. **Corruption Barometer-** Social Media platforms are gradually being used by government agencies to combat corruption by empowering citizens to report cases of bribery [202] [203]. Popular social media websites like Facebook and Twitter provides an effective and quick two-way communication between citizens and government wherein citizens can engage with the government. Our future work includes the automatic identification and extraction of useful information from bribery reports posted by citizens that can help the anti-corruption and law enforcement agencies in understating the extent of corruption in various departments, location and the volume of bribes being paid.
4. **Fake News Detection-**Prior literature shows that while there has been a lot of work done in the area of rumor, spam and phishing content detection, the spread of fake news has recently gained the attention of the researchers [204]. However, due to the lack of facts and ground truth, automatic identification of fake news is a technically challenging problem [205]. Our future work includes investigating the applications of contrast in contradictions, cross-platform data mining and parallel corpora for identifying the fake news on social media.
5. **NLP Systems for Security Informatics-** Existing NER tools such as Stanford CORE NLP Parser [91] and Indico [101] only extract the specific entities like people, location, money, date, time, and cardinal numbers. However, such NER tools are not able to capture the named entities when applied in security informatics domain where named entities have semantic meanings. For example, in a terrorist attack news, standard NER tools can identify the entities like a person, location, and organization. While the semantics of these entities such as 'defense organization', 'group who caused the attack', 'place where the attack happened' are not captured. Similarly, standard NERs can identify the cardinal number while the information like the number of people injured or died need to extracted. Since not all reports are structured and written by different teams; a dictionary based approach cannot be used to identify all relevant entities and requires a customized named entity recognizer for the government and security informatics based applications.

7.3.3 Modality and Dimensions

1. **Multimedia Content-** In the work presented in this dissertation, we have investigated the application of only contextual and linguistic metadata for building OSSMInt based applications. Social media websites allow users to post multimedia content (images with text, gifs, videos) to express their opinions and disseminate information. We believe that mining such content and extracting features from them can be used to improve the accuracy of classification. Mining such content includes the advanced techniques such as image processing [206] and optical character recognition [207] based methods.
2. **Cross-Platform-** to develop the proposed applications in this thesis, we have explored different and various social media platform for each case-study. However, we believe that despite the different structure and user behavior on the websites, analyzing data on multiple platforms can improve the performance of base approaches [208]. The future work includes the identification of similar or extended radicalized communities on multiple social media platforms.

Appendix A

Bibliometric and Scientometric Analysis

In Appendix of this thesis, we present some of the work that is a part of the research published during my Ph.D. However, due to non-social media and government applications work, it is not a part of the chapters discussed in this dissertation. Since mining open source social media and making intelligence from social media data for government applications is a multidisciplinary area; as a part of my comprehensive examination and domain assessment, we conducted a bibliometric and scientometric analysis on the conferences, publication records and previous research conducted in this area. We also performed a publication mining on the research published in well known and reputed conferences in Computer Science domain. In the appendix of this thesis, we provide a supplementary information and a brief introduction of these studies.

A.1 ACM Hypertext and Web Conferences

ACM Special Interest Group on Hypertext and the Web (ACM SIGWEB)[1] is a community of scientists, researchers, scholars and professionals in academia and industry. ACM SIGWEB studies various forms of hypertext, document engineering, social networks, information and knowledge management, digital libraries, hypermedia, web science, data mining, web search and user modeling. ACM SIGWEB conducts several activities such as sponsoring and organizing conferences, delivering awards, coordinating education and research development activities. One of the major activities of ACM SIGWEB is organizing and sponsoring symposia, workshops, and conferences aimed at providing platforms for the exchange of ideas and information. ACM SIGWEB (formerly known as SIGLINK) is now in its third decade and sponsors seven annual conferences spanning a broad range of topics. The seven conferences sponsored by ACM SIGWEB (as of 8 May 2016) are: 1) Hypertext and Social Media (HT), 2) Joint Conference on Digital Libraries (JCDL), 3) Document Engineering (DOCENG), 4) Web Science (WEBSCI), 5) Information and Knowledge Management (CIKM), 6) Web Search and Data Mining (WSDM), 7) User Modeling, Adaptation and Personalization (UMAP). These seven ACM SIGWEB conferences are prestigious in the broad area of the Web. They are long-running conferences that are convened across the world and, together, they provide an international forum for the presentation of research results and the exchange of ideas between academic and industry professionals. Assessment and evaluation of the quality, impact, status and evolution of the various SIGWEB conferences are important for the ACM SIGWEB community and conference sponsors, steering committees, science & technology policy makers and government bodies. We believe that a scientometric and bibliometric

[1] `http://www.sigweb.org/`

analysis provides a scientific approach to explore and evaluate the state of these conferences systematically. The Digital Bibliography & Library Project (DBLP) Computer Science Bibliography[2] provides accessible bibliographic information on the major computer science journals and proceedings and is a popular and widely used service. As of 8 May, 2016, the DBLP dataset contains 3,337,182 articles by 1,717,004 distinct authors published in various proceedings, including 4,731 conferences, 1,474 journals and other sources such as Ph.D. theses and technical reports. DBLP provides an XML dump of the DBLP records and entries containing the metadata and attributes of indexed publications. The specific research goal of this study is to conduct a bibliometric analysis based on the DBLP data and proceedings metadata (metadata of articles, authors, affiliations, and conferences) extracted from the ACM digital library. We parsed the ACM digital library pages of each article and analyzed various aspects of the seven ACM SIGWEB-sponsored Conferences. The research objective of this study is to present a reflection and a historical or general overview of the seven SIGWEB conferences. Furthermore, to accomplish this goal, our aim is to analyze the metadata for all papers published from the beginning of the conferences through September 17, 2015 that are available from the DBLP dataset, the ACM digital library, and from data extracted using other open source APIs. In particular, we aim to conduct an analysis on the following facets of the SIGWEB conferences:

1. Paper selectivity and yearly trend in terms of the number of accepted and submitted papers
2. Average number of authors per paper
3. Prolific and most productive authors
4. University and industry collaboration as indicated by joint authorship
5. Scholarly output of various countries in the world that have contributed to SIGWEB conferences
6. Cross-country collaboration (as indicated by co-authorship from researchers from different countries)
7. Female participation as authors, and an investigation of gender imbalances or gaps
8. Number and percentage of papers published by authors from the country hosting the conference
9. Common research topics and themes as indicated by the author-generated keywords
10. Topic evolution and trends across years
11. Funding agencies that have sponsored research published in SIGWEB papers

A.1.1 Experimental Setup

We conduct our study on the publication records extracted for SIGWEB conferences (9,141 individual scholarly entries for 162 conference proceedings) from the DBLP bibliographical database. We formulated 11 research questions that cover above facets and dimensions and analyzed several aspects of the SIGWEB community and the SIGWEB conferences. The DBLP database provides only the publication records of a conference or an article such as the conference name, authors' names, year published and publisher (ACM, IEEE, and Springer). However, these records are insufficient to conduct a comprehensive study on the ACM SIGWEB community. Therefore, we also extracted additional metadata for each SIGWEB conference by parsing the contents of the ACM Digital Library. The aim of ACM metadata collection is to enrich the information of four attributes of the bibliographical entries in the DBLP database. These are primarily named Articles, Authors, Affiliations, and Conferences.

1. Conference Metadata Enrichment includes mining ACM digital library for the extraction of dignitaries' profiles (authors of a conference proceeding), conference location, and conference acceptance rate per year.

2. Articles' Metadata Enrichment includes the extraction of ACM concepts and author tags associated with the article, unique authors, citations of the article and references used in the article, publication type (short, full, poster, and demo) and funding source information.

[2] `http://dblp.uni-trier.de/`

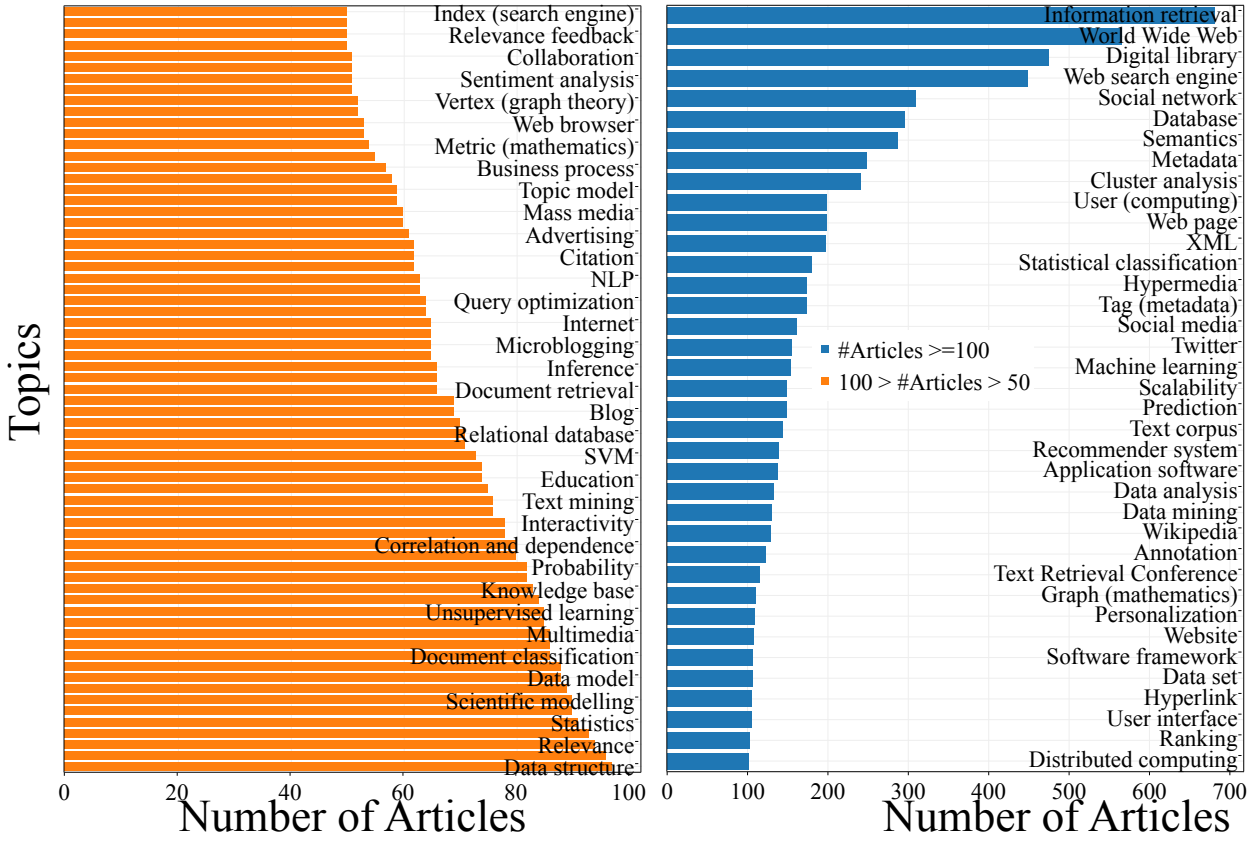


Figure A.1: Distribution of ACM Concepts for more than 50 Articles Published in 7 ACM SIGWEB Conferences.

3. Authors Metadata Enrichment includes mining authors' profile in ACM digital library for extracting contextual metadata such as the number of publications, the number of citations, colleagues, and co-authors, affiliation information (all and unique to articles). We also used Genderize.io API[3] to determine the gender of the authors in our database. We enriched the affiliation information by identifying the type of organization such as industries and academic institutions. We used four different open source APIs (OpenStreetMap (OSM) API[4], Alchemy API [114], Google Map API[5] and the Bing Geocoding API[6]) arranged in a cascaded ensemble manner to identify the type of each affiliation.

Since, the work presented in the appendix is not a part of the work discussed in thesis chapters, we present the results and statistics for only one research question proposed in our study and present the summary of results obtained for the remaining questions. Open source social media intelligence (OSSMInt) for government and law enforcement agencies is a multi-disciplinary domain comprising of three domains: 'social media analytics', 'security informatics' and 'data mining and machine learning'. Therefore, we discuss the results obtained

[3] https://genderize.io/

[4] http://wiki.openstreetmap.org/wiki/

[5] https://developers.google.com/maps/documentation/geocoding/intro

[6] https://msdn.microsoft.com/en-us/library/ff701713.aspx

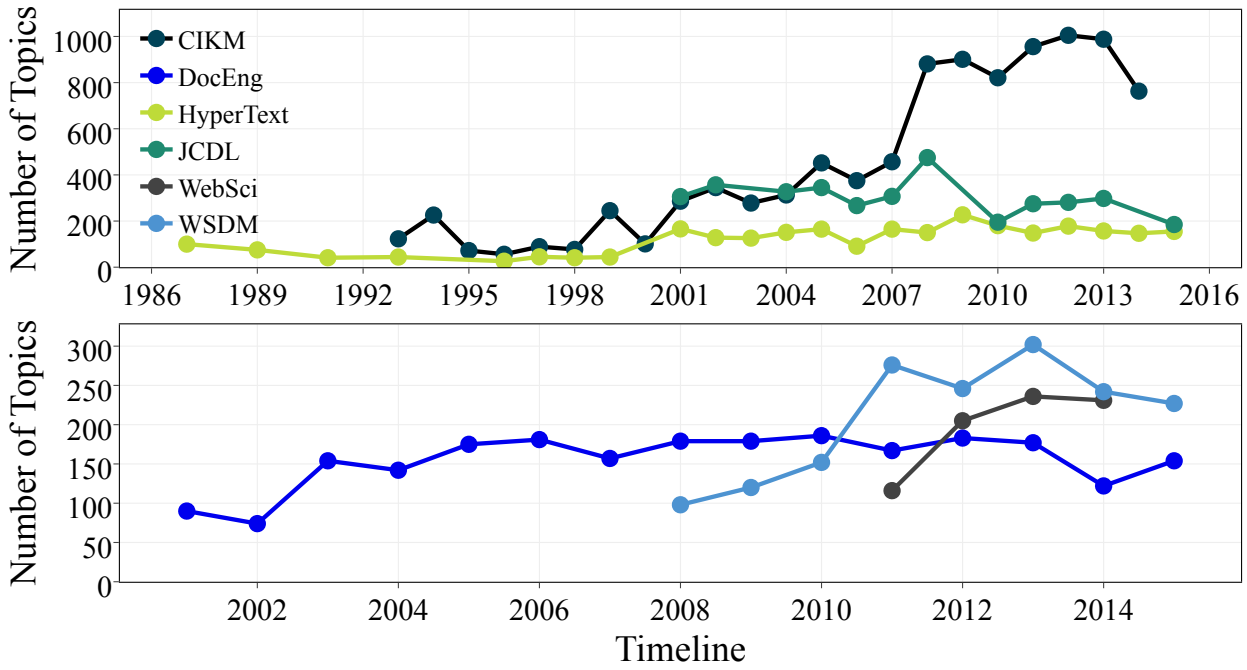


Figure A.2: Evolution of the Number of Topics in SIGWEB Conferences over the Years.

for our research question on the common topics of interest in ACM SIGWEB conferences and evolution of various fundamental and multi-disciplinary topics in these conferences.

RQ: What is the trend of the evolution of topics in SIGWEB conferences?

Based on the ACM concepts and author tags collected in Section A.1.1, we find that the collective distribution of SIGWEB articles for each tag varies between 1 and 681. Our database reveals that among 5,743 unique concepts, a total of 5,100 concepts appear as labels for fewer than ten articles, and 2,875 tags are associated with only one article. Most of these outlier tags are for nontechnical research-specific articles, and are not related to the usual topics seen in the conference. For example, several such tags are "spectral theory of ordinary differential equations", "personal knowledge management", and "computerized physician order entry". Figure A.1 shows the overall distribution of various ACM concepts that appear as labels for 50 to <100 articles (left) and more than 100 articles (right). Figure A.1 reveals that the majority of the concepts with fewer than 100 articles are either application-specific concepts or proposed metrics and solution-approach-specific key terms. Some examples include "query optimization" (64), "business process" (57), "blog" (69), "support vector machine" (73), "text mining" (76) and "user-generated content" (51). In contrast, concepts that are generalized and are common to all SIGWEB conferences are associated with large numbers of articles. Some examples include "information retrieval" (681), "world wide web" (565), "digital library" (475), "web search engine" (449), "social network" (309), "database" (295), "semantics" (287), "metadata" (248), "cluster analysis" (241), "machine learning" (154), "data analysis" (133), "data mining" (130) and "data set" (106).

Research shows that certain themes and topics in a conference can later evolve into more advanced and important research topics as new topics emerge [209]. Some of these topics arise over the years, includ-

ing the topics related to various interdisciplinary domains. Coulter et. al. [209] presented a study based on 13 years of publication records from Software Engineering conferences (1982–1994). They performed a co-word analysis on the index terms of these publications extracted from ACM-CCS. Their study revealed that the software engineering research field is growing rapidly; several new topics are introduced over the years such as "object oriented themes", "user interfaces", and "software reuse". They also identified the core themes of these conferences and the topics that remained constant over a span of 13 years such as "requirement and specification", "quality assurance", "software development" and "verification and validation". In contrast to the existing study, and as an extension to the previous research question (RQ9), we analyzed the evolution of various topics and research themes from ACM SIGWEB conferences over the years.

Figure A.2 shows a line graph of the distribution of topics in each SIGWEB conference over the past 2.5 decades. As Figure A.2 shows, compared to all the other SIGWEB conferences, CIKM has the maximum variation in the number of unique topics associated with the articles published in the conference in every year. Despite being the oldest and most popular SIGWEB conference, HyperText has had relatively fewer new topics introduced over the years; the majority of publications in the HyperText conference are consistent with the common and core topics of the conference. For example, hypermedia and hyperlinks have been the core research topics in HT since 1989. During the first few years of the conference, HT included articles labeled with approximately 100 unique topics. Unlike other SIGWEB conferences, the number of topics in HT decreased continually for 13 years. However, as social media, blogs, and forums emerged, the number research topics for HT also increased—at a rate of 75% (refer to Figure A.2). For example, blog, social network, and multimedia topics were introduced in HT in 2002. Furthermore, with the emergence of microblogging websites, topics such as "Twitter", "user generated content", and "microblogging" were introduced in 2010–2011. Figure A.2 reveals that DocEng has had a relatively constant rate for the number of distinct topics in the conference each year. We found that, despite including new topics in the conference every year, the number of unique topics per year remains approximately the same. For example, in 2009, published articles were labeled with 143 topics that were different and unique from 2008, but the total number of unique topics in 2008 and 2009 was the same (i.e., 179). Figure A.2 reveals that there has been a rapid change in the number of unique topics and concepts in the WSDM conference over the years. The distribution of unique topics in WSDM varies from 98 to 302 per year. These topics cover a wide range of concepts. Web mining, web crawlers, user-generated data, social media, information retrieval and ranking are some examples of the concepts that have been core and major topics of research in WSDM since 2008. While some topics that were either research-specific or that have risen over time occur significantly less often for WSDM articles. For example, concepts such as Wordnet and Yago are associated with only very few research papers in WSDM. Similarly, semantic web, ontology, Facebook, big data and Twitter are concepts that have been introduced in the conference since the emergence of Web 3.0 and microblogging portals.

A.1.2 Summary of Results Obtained for other Research Questions

Our results revealed that the overall acceptance rate of each conference has been decreasing over the years and varies between 20% to 40%. Each year that the number of submissions decreases for a conference, we see an increasing acceptance trend. Our findings reveal that the average number of co-authors per article in both older and newer SIGWEB conferences varies up to maximum values of 2.7 and 3.87, respectively. Over the past decade there has been an upward trend in the degree of author collaborations. The scholarly output of each author revealed that the researchers publishing the most in SIGWEB conference proceedings are those authors who have been involved in the conference since the beginning. Researchers publishing from various industries and academic institutions have also been active participants in the conference. We conclude that the degree of collective collaboration among various industries and universities is relatively lower in the SIGWEB conferences than in other CSR conferences. CIKM and WSDM are the two conferences that have recently had a large number of publications co-authored by researchers from different types of organizations.

The SIGWEB group is a widespread community that stretches across the globe and includes publications from various countries of the world. However, we conclude that the USA, United Kingdom, China and a few European countries contribute the most to the community.

Our empirical analysis also revealed that despite an upward trend in the degree of co-authorship, the overall collaboration of authors from distinct countries is relatively low; only those countries with the highest scholarly output in SIGWEB conferences have been collaborating with each other. We performed a gender analysis on the SIGWEB-published authors and conclude that unlike other CSR conferences the participation rate of female authors is significantly high in SIGWEB conferences ($\sim$ 50% of the total number of authors participating in the conference); however, there is a huge imbalance among researchers in leadership positions (general chairs and editors). Our analysis revealed that female authors who have served as editors or as general chairs in the conference have been part of the conference since the beginning. Similar to the trends for scholarly output of a country, only a few countries such as the USA, some European countries and China have higher participation rates in SIGWEB conferences from local communities when they are the host country of a conference. Our empirical analysis on articles' metadata revealed that certain topics and themes of the conference remain the same over the years and some evolve into major and advanced topics (core themes) as new topics emerge in the community. We conclude that SIGWEB conferences have been growing rapidly as new topics and research themes have arisen. With the emergence of new studies in the domain of web science, many third party organizations (both government and industry) have provided funding resources for the published studies. SIGWEB conferences metadata reveals that most of the organizations supporting web science and data analytics studies are based in the USA, UK, China and Europe. In this study, we conclude that SIGWEB is a rapidly emerging and expanding community from many aspects, including articles, conferences, and number of authors. However, researchers affiliated with some of the more advanced developed and developing countries participate in the conference at low rates.

A.2 Security Informatics Conferences and Journals

Intelligence and Security Informatics (ISI) is a field of study concerning investigation and development of counter terrorism, national and international security support systems and applications. ISI is a fast growing interdisciplinary area which has attracted several researchers and practitioners (form academia, government and industry) attention spanning across multiple disciplines like computer science, social sciences, political science, law and even medical sciences. The studies presented in this thesis demonstrate various application of computer science in security informatics domain. The aim of the study presented in this work is to conduct a systematic literature survey of those previous techniques as documented in scholarly articles. Our goal is to do a comprehensive analysis of these articles to better understand the state-of-the-art, research gaps, and techniques used in existing literature. We conducted an in-depth and rigorous literature survey on two sub-topics within ISI (online radicalization and civil protest forecasting). Our literature review is comprehensive and provides insights useful to the ISI research community. To the best of our knowledge this work is the first such survey of existing literature on social media in the domain of automated techniques for Online Radicalization detection and generating early Warning or predicting civil unrest related events. We propose a one class classification architecture across several dimensions (social media platforms being used as a data source, countering issues of online radicalization and civil unrest, machine learning techniques) to classify scholarly articles that fall under the scope of focus of our research (social media platforms, text analysis based machine learning techniques, security informatics area). We perform an in-depth characterization and classification based upon meta analysis of existing researches. We analyze every article and inspect commonly used techniques, identify trends and find research gaps.

Since, the aim of this study is to conduct bibliometric analysis based on the study presented in the articles, unlike our previous study, we do not conduct this analysis on DBLP publication database. We create a list of key-terms representing our topic or problem area. For example, some of the search key-terms to retrieve relevant papers are: 'civil unrest/protest', 'event forecasting', 'early warning', 'extremism detection', 'online

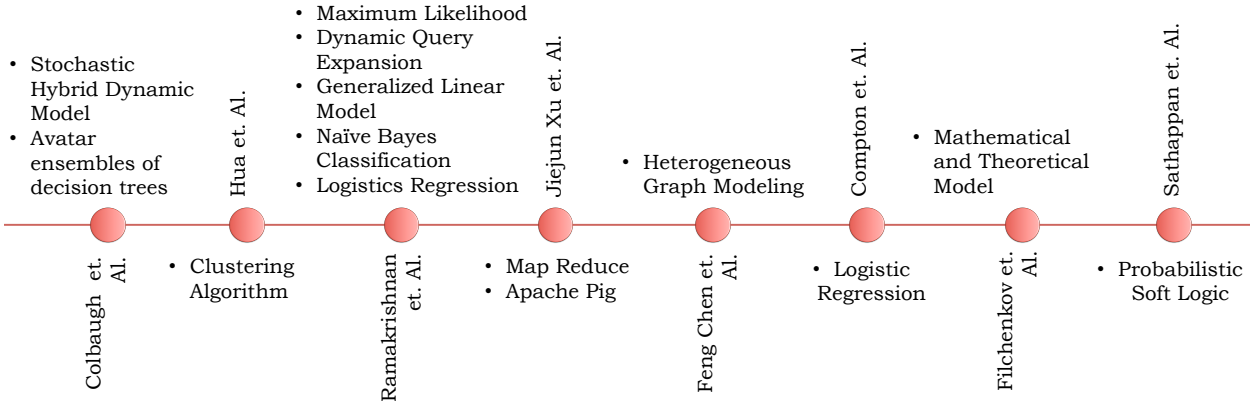


Figure A.3: Machine Learning Techniques Used By Researchers in Existing Literature of Event Forecasting in the Domain of Civil Disobedience.

radicalizing community detection'. We search relevant articles using Google Scholar[7] which is a well-known and widely used web-based search engine for finding scholarly literature. We use Google Scholar as it has a good coverage and index of articles and also provides a powerful mechanism to explore related work, citations and authors. Google Scholar also provides data on how often and how recently a paper has been cited which is also a useful metric (judging the impact) for conducting our literature survey. We go through the Title, Keywords and Abstract of the articles retrieved as a result of key-term based search as well as through related article and cited by links provided for each article in the search result. We perform a meta analysis on article and determine the relevance of the article to our topic or focus area. We perform a one class classification (Rule Based Classifier) on each article and check if it meets the scope and focus of our study. During classification, we typically mine the Title, Keywords and Abstract and accordingly save the article in our database of collection of closely related work.

A.2.1 Machine Learning and Data Mining Techniques

Figures A.3 and A.4 illustrate the machine learning and information retrieval techniques used in previous papers of civil unrest event prediction and online radicalization respectively. Figure A.3 illustrates the proposed methods and techniques used in each paper individually. Ramakrishnan [187] proposed five different techniques for event forecasting for five different kinds of data and models (volume based, opinion based, tracking of activities, distribution of events and cause of protest). Unlike Ramakrishnan [187], in Figure A.3, we find that most of the researchers have applied ensemble learning on multiple data mining and machine learning techniques to achieve better accuracy in prediction [188] [192] [210]. Clustering, logistic regression, and dynamic query expansion are the commonly used techniques to predict upcoming events related to civil unrest or protest. We go through the methodology Section of each paper and observe that named entity recognition is a common phase for all approaches illustrated in Figure A.3. In named entity recognition, they extract several entities present in the contextual metadata (tweets, Facebook comment, News article) [211]. These entities could be spatiotemporal expressions, the topic being discussed in posts (refer to Table A.1 for the complete list of entities and features). List of all entities and keywords is dynamically expanded by using Dynamic Query Expansion method where they find similar and relevant keywords/entities using external

[7]https://scholar.google.co.in/

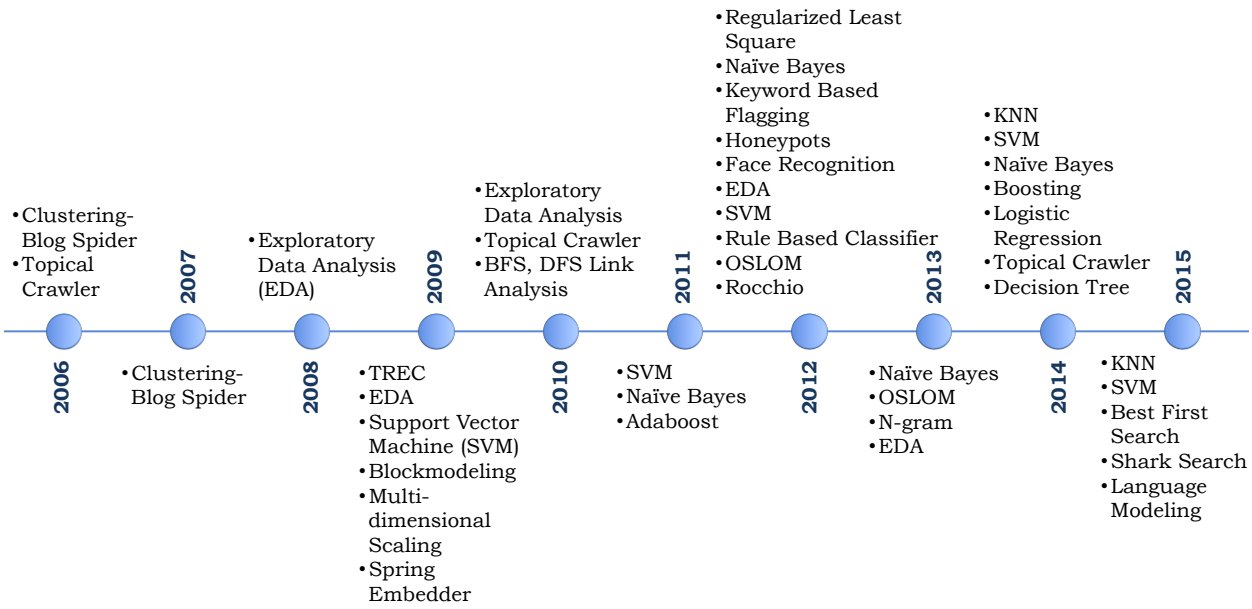


Figure A.4: Machine Learning Techniques Used By Researchers in Existing Literature of Online Radicalization and Hate Promotion Detection.

lexical source (for example- WordNet[8], VerbOcean[9]). Dynamic Query Expansion is an iterative process and converges once the keywords are stable. They further perform several clustering and classification techniques on these entities and text to predict upcoming events. Another popular technique used for event forecasting is graph modeling. Feng Chen et.al. [212] uses a heterogeneous graph modeling as keyword enrichment and pre-processing. To achieve accurate results in event forecasting, they use only filtered entities for NPHGS approach. According to Feng Chen et.al. [212], a heterogeneous graph is defined as a network consisting of nodes, edges, and relations where nodes are the entities extracted using named entity recognition. There can be multiple types of nodes equivalent to the number of entities extracted (topic, temporal, spatial, organization, etc). Edges are the link between two entities and relation defines the feature vector between two entities. For example, one relation between a Twitter user U (entity: people) and a term T (entity: topic) can be the number of tweets by user U on topic T. Entities having high relevance and polarity score above certain threshold are filtered and used in next phase of NPHSG.

Unlike Figure A.3, due to a large number of papers in online radicalization domain, we present machine learning technique over a timeline instead of individual presentation for each paper. Figure A.4 reveals that text classification {KNN (k-nearest neighbor), naive Bayes, support vector machine, rule-based classifier, decision tree}, clustering (blog spider), exploratory data analysis (EDA), topical crawler/link analysis (breadth first search, depth first search, best first search) and keyword-based flagging (KBF) are the most widely used techniques for online radicalization detection on social media websites [136] [213] [214]. Social networking websites, micro-blogging websites and video sharing websites are amongst the largest repositories of user-generated content on the web. Therefore, Text classification (automatic and semi-supervised learning), clustering (unsupervised learning), EDA and KBF approaches are very well known techniques and commonly used for identifying extremist content on social media [148] [215]. Whereas, topical crawler and link analysis are the techniques used for identifying similar users and locating hidden communities on

[8]https://wordnet.princeton.edu
[9]http://demo.patrickpantel.com/demos/verbocean/

Table A.1: List of Various Dimensions and their Associated Properties Used for Annotating Articles in the Domain of Civil Unrest Related Event Forecasting.

Dimension	Categories		Description
Features	F1	Temporal	Presence of time related expressions
	F2	Spatial	Presence of location based expressions
	F3	Topic	Presence of targeted topic or domain related expressions
	F4	Content	Mining text to extract event related information
	F5	Demographic	Other demographic and statistical based metadata
Evaluation	E1	Cross Validation	optimizing the output of forecasting by splitting data into k-samples
	E2	Precision	Evaluates the exactness of forecasting results
	E3	Recall	Evaluates the completeness of results
	E4	NA	No evaluation technique is mentioned
Analysis	A1	Content	Mining only textual content for feature extraction and event forecasting
	A2	Community	Mining user profiles and networks for extracting event related information
Language	L1	English	Conducting experiments only on English Language Tweets
	L2	Non-English	Posts consisting of any Non-English language content
	L3	Others	Testbed consisting of multiple languages content
Genre	G1	Protest-Country	Studying the protests happened in one country/region [Latin America]
	G2	Global	Event forecasting on any civil unrest related event happened worldwide

social media websites [216]. The topical crawler is a recursive process that adds and removes nodes after each iteration. It starts from a seed node, traverses in a graph navigating through some links and returns all relevant nodes to a given topic. The breadth first search, depth first search, and best first search are the different approaches to select neighbors and navigate through the external links. These links and neighbors vary for different social networking websites. For example, if user u posted a tweet t then a neighbor can be a follower of user u, or users liking and re-tweeting or re-blogging (in the case of Tumblr) that post. Similarly in YouTube, if user u posted a video v then a link can re-direct to a user channel who subscribed u or posted a comment on video v. Language modeling, n-gram, Boosting are other techniques for classifying textual data as hate promoting based upon several discriminatory features.

A.2.2 Characterization and Classification of Articles Based Upon Meta Analysis

We present a characterization based study on previous researches done in the area of civil unrest related event forecasting and online radicalization detection. We analyze each paper and create a list of all dimensions to demonstrate the statistics. We also present statistics of existing literature that use social media platforms as a data source for conducting experiments. Each dimension is further classified in sub-categories and has their properties associated with them. Tables A.1 and A.2 shows the categories and subcategories of various facets identified for bibliometric analysis of publications database of online civil unrest prediction and online radicalization detection respectively.

Based on our analysis, we observe a surge in research interest over the last 3 years on the topic of solutions for identifying and forecasting civil unrest and mobilization by mining textual content in open-source social media. The number of research papers on social media analytics for online radicalization detection is higher (about 7 times) than on civil disobedience detection. Intelligence and Security Informatics (ISI) conference and Security Informatics (SI) Journal are the two main venues for publishing papers on the topic of online radicalization detection and mining. Our analysis reveals that micro-blogging websites like Twitter and Tumblr are the two most common sources of social media data for civil unrest detection and forecast-

Table A.2: List of Various Dimensions and their Associated Properties Used for Annotating Prior Literature in the Domain of Online Radicalization Detection.

Dimension	Categories		Description
Features	F1	Text	Textual content of the posts (Video Title, Tweet, User Comments)
	F2	Link	Links between two user profile (Subscription, follower, friend)
	F3	Demographic	Other demographic and statistical based metadata
Evaluation	E1	Precision	Evaluates the exactness of forecasting results
	E2	Recall	Evaluates the completeness of results
	E3	F-Score	Weighted harmonic mean of Precision (E1) and Recall (E2)
	E4	Accuracy	Evaluates the correctness of the technique
	E5	cross Validation	optimizing the output of forecasting by splitting data into k-samples
	E6	SNA	Social Network Analysis to show the findings in community detection
	E7	User Based	Evaluation performed by external users
	E8	NA	No evaluation method is defined.
Analysis	A1	Content	Mining only textual content for feature extraction and event forecasting
	A2	User Profile	Mining user profiles metadata for extracting event related information
	A3	Community	Mining linked profiles and their communities
Language	L1	English	Conducting experiments only on English Language Tweets
	L2	Arabic	consisting of content written in Arabic (scripted or language)
	L3	Other	consisting of other Non-English language (excluding L2) (German, French)
Region	R1	US Domestic	targeting US issues or radicalization originated from US domestic regions
	R2	International	Researchers focusing on global or international radicalization or extremism
	R3	Others	Radicalization originating or targeting worldwide regions (Middle Eastern)
Genre	G1	Anti-black	Content posted by white supremacy communities targeting black people
	G2	Jihad	Groups posting content for promoting Jihad among their viewers
	G3	Terrorism	activities performed by terrorists groups
	G4	Extremism	posted in order to promote extremism among various targeted audience
	G5	Religion	Content posted against a religion (example- Anti-Islamic Tweets)

ing applications. We believe that Twitter has been very instrumental in facilitating political mobilization in comparison to other social media platforms because of its inherent characteristics of sharing short text through direct messages and follower relationship. It is interesting to observe that despite the immense popularity and penetration of YouTube as an online video-sharing website, it has not been used in any of the existing research for protest planning or prediction. On the contrary, our research reveals that YouTube is the most widely used platform for online radicalization, hate and extremism promotion as indicated by the published research papers. In comparison to Twitter, Tumblr which is also a popular micro-blogging website has not been a major focus of research attention for online radicalization detection applications.

In this survey, we categorize existing studies on various dimensions such as the use of discriminatory features, type of metadata being analyzed and evaluation techniques used by authors to examine the effectiveness of their approach. Our analysis reveals that mining contextual metadata of a post and spatiotemporal information are most commonly used features for predicting civil unrest related events. In existing studies, we find that maximum researchers evaluate the accuracy of their prediction approach by computing precision of their results. We also observe that 90% of the studies mine English language text. Since the problem of civil unrest related event prediction can be targeted to a specific country or region, we find that the methods proposed in various studies are able to mine information from multilingual texts (Dutch, Spanish, and French). Our analysis also reveals that 60% of the studies target events specific to a country or region. Maximum number of studies are conducted on the events happened in Latin America and USA.

Characterization and meta-analysis performed on the existing studies of online radicalization detection

reveal that contextual based metadata is most commonly used attribute to identify the presence of extremist content. However, demographic information and activity feed of a user profile and links between two users are discriminatory features for locating hidden communities of radical users. We observe that many of the existing techniques are capable of mining multilingual text such as Arabic and capture relevant information. Our analysis reveals that some of the articles focus on only a country or region specific radicalization detection (Latin America, Middle Eastern, North Africa). We also observe that to examine the effectiveness of results in community detection, and extremism content identification, social network analysis, and precision measures are the most commonly used evaluation methods.

Bibliography

[1] Richard Hanna, Andrew Rohm, and Victoria L Crittenden. "Were all connected: The power of the social media ecosystem". In: Business horizons 54.3 (2011), pp. 265–273.

[2] Jan H Kietzmann et al. "Social media? Get serious! Understanding the functional building blocks of social media". In: Business horizons 54.3 (2011), pp. 241–251.

[3] Jean Éric Pelet. "Using Web 2.0 social computing technologies to enhance the use of information systems in organizations". In: Social Computing Theory and Practice: Interdisciplinary Approaches: Interdisciplinary Approaches (2010), p. 101.

[4] Facebook- Social Media and Social Networking website. `http://facebook.com/`.

[5] Twitter micro-blogging website. `https://twitter.com`.

[6] Sina Weibo- Chinese micro-blogging website. `http://weibo.com/`.

[7] Tumblr micro-blogging website. `http://tumblr.com/`.

[8] YouTube- Video Sharing and Social Networking Wesite. `http://youtube.com/`.

[9] Vimeo- Video Hosting and Sharing website. `http://vimeo.com`.

[10] DailyMotion- Video Hosting and Sharing website. `http://dailymotion.com`.

[11] Imgure- Image Hosting and Sharing Website. `http://imgur.com`.

[12] Flickr- Online Image and Video Sharing Website. `http://flickr.com`.

[13] Instagram- Online Image and Video Sharing Website. `http://instagram.com`.

[14] StackOverflow- Question and Answering Website for Programmers and Developers. `http://stackoverflow.com`.

[15] Reddit- Online Social News and Media Aggregation Website. `http://reddit.com`.

[16] List of Active Social Media Website- Source: Wikipedia. Accessed on Jan 31, 2017 `https://en.wikipedia.org/wiki/List_of_social_networking_websites`.

[17] Social Media Website Statistics `http://www.wordstream.com/blog/ws/2017/01/05/social-media-marketing-statistics`. [Last Accessed on Jan 31, 2017].

[18] YouTube User Statistics `https://fortunelords.com/youtube-statistics/`. [Last Accessed on Jan 31, 2017].

[19] Twitter User and Tweets Statistics `https://about.twitter.com/company`. [Last Accessed on Jan 31, 2017].

[20] Tumblr Bloggers Statistics `https://www.tumblr.com/about`. [Last Accessed on Jan 31, 2017].

[21] Tumblr Posts Statistics `http://www.internetlivestats.com/`. [Last Accessed on Jan 31, 2017].

[22] Mirjam Wattenhofer, Roger Wattenhofer, and Zack Zhu. "The YouTube Social Network." In: ICWSM. 2012.

[23] Haewoon Kwak et al. "What is Twitter, a social network or a news media?" In: Proceedings of the 19th international conference on World wide web. ACM. 2010, pp. 591–600.

[24] Yi Chang et al. "What is tumblr: A statistical overview and comparison". In: ACM SIGKDD Explorations Newsletter 16.1 (2014), pp. 21–29.

[25] Facebook API. Facebook for Developers `https://developers.facebook.com`. [Last Accessed on Jan 31, 2017].

[26] Twitter API. Twitter Developer Documentation `https://dev.twitter.com/docs`. [Last Accessed on Jan 31, 2017].

[27] YouTube API. YouTube Developer Documentation `https://developers.google.com/youtube/documentation/`. [Last Accessed on Jan 31, 2017].

[28] Tumblr API `https://www.tumblr.com/docs/en/api/v2`. [Last Accessed on Jan 31, 2017].

[29] Sina Weibo API `http://open.weibo.com/wiki/API/en`. [Last Accessed on Jan 31, 2017].

[30] Swati Agarwal, Ashish Sureka, and Vikram Goyal. "Open source social media analytics for intelligence and security informatics applications". In: International Conference on Big Data Analytics. Springer. 2015, pp. 21–37.

[31] Daniel Zeng et al. "Social media analytics and intelligence". In: IEEE Intelligent Systems 25.6 (2010), pp. 13–16.

[32] Siqi Zhao et al. "Human as real-time sensors of social and physical events: A case study of twitter and sports games". In: arXiv preprint arXiv:1106.4300 (2011).

[33] Mike Moran, Jeff Seaman, and Hester Tinti-Kane. "Teaching, Learning, and Sharing: How Today's Higher Education Faculty Use Social Media." In: Babson Survey Research Group (2011).

[34] Ayelet Gal-Tzur et al. "The potential of social media in delivering transport policy goals". In: Transport Policy 32 (2014), pp. 115–123.

[35] Pierre R Berthon et al. "Marketing meets Web 2.0, social media, and creative consumers: Implications for international marketing strategy". In: Business horizons 55.3 (2012), pp. 261–271.

[36] Andreas M Kaplan and Michael Haenlein. "Users of the world, unite! The challenges and opportunities of Social Media". In: Business horizons 53.1 (2010), pp. 59–68.

[37] Andrea L Kavanaugh et al. "Social media use by government: From the routine to the critical". In: Government Information Quarterly 29.4 (2012), pp. 480–491.

[38] KL Courtney et al. "The use of social media in healthcare: organizational, clinical, and patient perspectives". In: Enabling health and healthcare through ICT: available, tailored and closer 183 (2013), p. 244.

[39] Foteini Alvanaki et al. "See what's enBlogue: real-time emergent topic identification in social media". In: Proceedings of the 15th International Conference on Extending Database Technology. ACM. 2012, pp. 336–347.

[40] Eugene Agichtein et al. "Finding high-quality content in social media". In: Proceedings of the 2008 international conference on web search and data mining. ACM. 2008, pp. 183–194.

[41] Alex Hai Wang. "Don't follow me: Spam detection in twitter". In: Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on. IEEE. 2010, pp. 1–10.

[42] Alexander Pak and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." In: LREc. Vol. 10. 2010. 2010.

[43] Georgios Kontaxis et al. "Detecting social network profile cloning". In: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on. IEEE. 2011, pp. 295–300.

[44] Robert P Schumaker and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system". In: ACM Transactions on Information Systems (TOIS) 27.2 (2009), p. 12.

[45] Sarah Vieweg et al. "Microblogging during two natural hazards events: what twitter may contribute to situational awareness". In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM. 2010, pp. 1079–1088.

[46] Andranik Tumasjan et al. "Predicting elections with twitter: What 140 characters reveal about political sentiment." In: ICWSM 10.1 (2010), pp. 178–185.

[47] Wilas Chamlertwat et al. "Discovering Consumer Insight from Twitter via Sentiment Analysis." In: J. UCS 18.8 (2012), pp. 973–992.

[48] Swati Agarwal and Ashish Sureka. "Topic-Specific YouTube Crawling to Detect Online Radicalization". In: 10th International workshop on Databases in Networked Information Systems (DNIS), Fukushima, Japan. 2015.

[49] Sheryl Prentice et al. "Analyzing the semantic content and persuasive composition of extremist media: A case study of texts produced during the Gaza conflict". In: Information Systems Frontiers 13.1 (2011), pp. 61–73.

[50] Swati Agarwal and Ashish Sureka. "A Topical Crawler for Uncovering Hidden Communities of Extremist Micro-Bloggers on Tumblr". In: 5th International Workshop on Making Sense of Microposts, Big things come in small packages (Microposts), Co-located with WWW, Florence, Italy. 2015.

[51] Nisha Aggarwal, Swati Agarwal, and Ashish Sureka. "Mining YouTube metadata for detecting privacy invading harassment and misdemeanor videos". In: Privacy, Security and Trust (PST). 2014, pp. 84–93.

[52] Swati Agarwal and Ashish Sureka. "Learning to Classify Hate and Extremism Promoting Tweets". In: Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint. IEEE. 2014, pp. 320–320.

[53] Vahed Qazvinian et al. "Rumor Has It: Identifying Misinformation in Microblogs". In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA, 2011, pp. 1589–1599. ISBN: 978-1-937284-11-4.

[54] OAuth- The OAuth 2.0 authorization framework. `https://oauth.net`.

[55] Swati Agarwal, Nitish Mittal, and Ashish Sureka. "Potholes and Bad Road Conditions-Mining Twitter to Extract Information on Killer Roads". In: 22nd International Conference on Database Systems for Advanced Applications (DASFAA). Springer, 2017.

[56] Nitish Mittal, Swati Agarwal, and Ashish Sureka. "Got a Complaint?- Keep Calm and Tweet It!" In: 12th International Conference on Advanced Data Mining and Applications (ADMA), Gold Coast, Australia. Springer, 2016.

[57] Swati Agarwal and Ashish Sureka. "Investigating the Dynamics of Religious Conflicts by Mining Public Opinions on Social Media". In: The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Jeju, South Korea. Springer, 2017.

[58] Swati Agarwal and Ashish Sureka. "A Collision of Beliefs: Investigating Linguistic Features for Religious Conflicts Identification on Tumblr". In: 13th International Conference on Distributed Computing and Internet Technology (ICDCIT), India. Springer, 2017.

[59] Swati Agarwal and Ashish Sureka. "Using kNN and SVM based One-Class Classifier for Detecting On-line Radicalization on Twitter". In: Proceedings of 11th International Conference on Distributed Computing and Internet Technology (ICDCIT), Odisha, India. 2015.

[60] Swati Agarwal and Ashish Sureka. "Using Common-Sense Knowledge-Base for Detecting Word Obfuscation in Adversarial Communication". In: Proceedings of Workshop on Future Information Security (FIS), co-located with COMSNETS, Bangalore, India. 2015.

[61] Swati Agarwal and Ashish Sureka. "Investigating the Potential of Aggregated Tweets as Surrogate Data for Forecasting Civil Protests". In: ACM India SIGKDD Conference on Data Sciences (CoDS), Pune, India. 2016.

[62] Nitish Mittal Swati Agarwal and Ashish Sureka. Syntactic Enhancement of Killer Road Complaint Tweets Posted on Twitter, Mendeley Data, v1, `http://dx.doi.org/10.17632/dm6s252524.1`. 2017.

[63] Swati Agarwal and Ashish Sureka. Semantically Analyzed Metadata of Tumblr Posts and Bloggers, Mendeley Data, v1, `http://dx.doi.org/10.17632/hd3b6v659v.1`. 2016.

[64] Swati Agarwal, Nitish Mittal, and Ashish Sureka. Enhanced Dataset of Citizen Centric Complaints and Grievances on Twitter, Mendeley Data, v1, `http://dx.doi.org/10.17632/w2cp7h53s5.1`. 2016.

[65] Swati Agarwal and Ashish Sureka. Religious Beliefs on Social Media: Large Dataset of Tumblr Posts and Bloggers Consisting of Religion Based Tags, `http://dx.doi.org/10.17632/8hp39rknns.1,MendeleyData,v1,`. 2016.

[66] Sumit Bhatia et al. "Specialized Research Datasets in the CiteSeer Digital Library". In: D-Lib Magazine 18.7/8 (2012).

[67] Heather A Piwowar and Wendy W Chapman. "Public sharing of research datasets: a pilot study of associations". In: Journal of informetrics 4.2 (2010), pp. 148–156.

[68] Toby Green. "We need publishing standards for datasets and data tables". In: Learned publishing 22.4 (2009), pp. 325–327.

[69] Vanessa Frias-Martinez, Abson Sae-Tang, and Enrique Frias-Martinez. "To call, or to tweet? Understanding 3-1-1 citizen complaint behaviors". In: ASE BigData/ SocialCom/ CyberSecurity Conference. 2014.

[70] Gohar Feroz Khan, Bobby Swar, and Sang Kon Lee. "Social Media Risks and Benefits A Public Sector Perspective". In: Social Science Computer Review 32.5 (2014), pp. 606–627.

[71] Euripidis Loukis, Yannis Charalabidis, and Aggeliki Androutsopoulou. "Evaluating a Passive Social Media Citizensourcing Innovation". In: Electronic Government: 14th IFIP WG 8.5 International Conference, EGOV 2015. Springer International Publishing, pp. 305–320.

[72] Thomas Heverin and Lisl Zach. "Twitter for city police department information sharing". In: Proceedings of the American Society for Information Science and Technology (2010).

[73] Megan Anderson, Kieran Lewis, and Ozgur Dedehayir. "Diffusion of innovation in the public sector: Twitter adoption by municipal police departments in the US". In: Portland International Conference on Management of Engineering and Technology, 2015.

[74] Albert Jacob Meijer and René Torenvlied. "Social Media and the New Organization of Government Communications An Empirical Analysis of Twitter Usage by the Dutch Police". In: The American Review of Public Administration (2014), p. 0275074014551381.

[75] Arthur Edwards and Dennis Kool. "Webcare in Public Services: Deliver Better with Less?" In: Social Media for Government Services. Cham: Springer International Publishing, 2015, pp. 151–166.

[76] Avinash Kumar, Miao Jiang, and Yi Fang. "Where not to go?: detecting road hazards using twitter". In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM. 2014, pp. 1223–1226.

[77] Yiming Gu, Zhen Sean Qian, and Feng Chen. "From Twitter to detector: Real-time traffic incident detection using social media data". In: Transportation Research Part C: Emerging Technologies 67 (2016), pp. 321–342.

[78] Kaiqun Fu, Rakesh Nune, and Jason X Tao. "Social media data analysis for traffic incident detection and management". In: Transportation Research Board 94th Annual Meeting. 2015.

[79] Eleonora et al. D'Andrea. "Real-time detection of traffic from twitter stream analysis". In: IEEE Transactions on Intelligent Transportation Systems 16.4 (2015), pp. 2269–2283.

[80] Eric Mai and Rob Hranac. "Twitter interactions as a data source for transportation incidents". In: Proc. Transportation Research Board 92nd Ann. Meeting. 2013.

[81] Axel Schulz, Petar Ristoski, and Heiko Paulheim. "I see a car crash: Real-time detection of small scale incidents in microblogs". In: Extended Semantic Web Conference. Springer. 2013, pp. 22–33.

[82] Napong et al. Wanichayapong. "Social-based traffic information extraction and classification". In: ITS Telecommunications (ITST), 2011 11th International Conference on. IEEE. 2011, pp. 107–112.

[83] Panos Panagiotopoulos et al. "Social media in emergency management: Twitter as a tool for communicating risks to the public". In: Technological Forecasting and Social Change 111 (2016), pp. 86–96.

[84] Martin F Porter. "An algorithm for suffix stripping. `http://snowball.tartarus.org/algorithms/porter/stemmer.html`". In: Program 14.3 (1980), pp. 130–137.

[85] Bing Search API Version 2.0. `http://www.bing.com/developers/s/APIBasics.html`.

[86] Bing Search Engine. `https://www.bing.com`.

[87] NoSlang- Internet & Text Slang Dictionary & Translator `http://www.noslang.com/dictionary/`.

[88] The Ministry of Road, Transport, and Highways, Government of India. `http://morth.nic.in`.

[89] Fitrie Handayani, Siti Dewi Sri Ratna Sari, and Wira Respati. "The use of meme as a representation of public opinion in social media: a case study of meme about Bekasi in path and Twitter". In: HUMANIORA 7.3 (2016), pp. 333–339.

[90] Shehroz S Khan and Michael G Madden. "A survey of recent trends in one class classification". In: Irish conference on Artificial Intelligence and Cognitive Science. Springer. 2009, pp. 188–197.

[91] Stanford's CORE NLP Parser. `http://stanfordnlp.github.io/CoreNLP/`.

[92] IBM Watson's Tone Analyzer. `http://www.ibm.com/watson/developercloud/tone-analyzer.html`.

[93] Official Twitter Handle of Railway Ministry of India. `https://twitter.com/RailMinIndia`.

[94] Official Twitter Handle of Delhi Traffic Police. `https://twitter.com/dtpTraffic`.

[95] Official Twitter Handle of Delhi State Police. `https://twitter.com/DelhiPolice`.

[96] Official Twitter Handle of Income Tax Department, Government of India. `https://twitter.com/IncomeTaxIndia`.

[97] WordNet- A lexical database for English. `https://wordnet.princeton.edu`.

[98] IBM Watson's Relationship and Extraction API. `http://www.ibm.com/watson/developercloud/apis/relationship-extraction-apis.html`.

[99] Google Maps Geocoding API. `https://developers.google.com/maps/documentation/geocoding/intro`.

[100] Preeti Bansal and Durga Toshniwal. "Analyzing civic complaints for proactive maintenance in smart city". In: Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on. IEEE. 2016, pp. 1–6.

[101] Indico API. Text Analysis. `https://indico.io`.

[102] OpenStreetMap API. `https://wiki.openstreetmap.org/wiki/API_v0.6`.

[103] Official Twitter Handle of Ministry of Road Transport and Highways, Government of India. `https://twitter.com/MORTHIndia`.

[104] Official Twitter Handle of Union Minister of Road Transport and Highways, Government of India. `https://twitter.com/nitin_gadkari`.

[105] Soujanya et al. Poria. "Dependency-Based Semantic Parsing for Concept-Level Text Analysis". In: Computational Linguistics and Intelligent Text Processing (2014), pp. 113–127.

[106] ConceptNet- An open, multilingual knowledge graph. `http://conceptnet.io`.

[107] Catherine Havasi, Robert Speer, and Jason Alonso. "ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge". In: Recent advances in natural language processing. Citeseer. 2007, pp. 27–29.

[108] Robert Speer and Catherine Havasi. "Representing General Relational Knowledge in ConceptNet 5." In: LREC. 2012, pp. 3679–3686.

[109] Steve Spagnola and Carl Lagoze. "Edge dependent pathway scoring for calculating semantic similarity in ConceptNet". In: Proceedings of the Ninth International Conference on Computational Semantics. Association for Computational Linguistics. 2011, pp. 385–389.

[110] Christiane Fellbaum. WordNet. Wiley Online Library, 1998.

[111] The DBpedia Knowledge Base. `http://wiki.dbpedia.org/about`.

[112] ConceptNet 5.5 REST API. `https://github.com/commonsense/conceptnet5/wiki/API`.

[113] S. Amari and S. Wu. "Improving support vector machine classifiers by modifying kernel functions". In: Neural Networks 12.6 (1999), pp. 783 –789. ISSN: 0893-6080.

[114] AlchemyLanguage- Natural language processing for advanced text analysis `https://www.ibm.com/watson/developercloud/alchemy-language.html`.

[115] Stephen Few. "Information dashboard design". In: (2006).

[116] Greg Acciaioli. "Grounds of conflict, idioms of harmony: custom, religion, and nationalism in violence avoidance at the Lindu Plain, Central Sulawesi". In: Indonesia 72 (2001), pp. 81–114.

[117] Swati Agarwal and Ashish Sureka. "But I did not Mean It!- Intent Classification of Racist Posts on Tumblr". In: 6th IEEE European Intelligence & Security Informatics Conference (EISIC), Uppsala, Sweden. IEEE, 2016.

[118] Joshua Cohen. "Freedom of expression". In: Philosophy & Public Affairs (1993), pp. 207–263.

[119] Robert A Emmons, Chi Cheung, and Keivan Tehrani. "Assessing spirituality through personal goals: Implications for research on religion and subjective well-being". In: Social indicators research 45.1-3 (1998), pp. 391–422.

[120] William R Swinyard, Ah-Keng Kau, and Hui-Yin Phua. "Happiness, materialism, and religious experience in the US and Singapore". In: Journal of happiness studies 2.1 (2001), pp. 13–32.

[121] Joshua A Wilt et al. "Anxiety predicts increases in struggles with religious/spiritual doubt over two weeks, one month, and one year". In: The International Journal for the Psychology of Religion (2016), pp. 1–9.

[122] Lai Fong Yang and Md Sidin Ahmad Ishak. "Framing interethnic conflict in malaysia: a comparative analysis of newspapers coverage on the hindu rights action force (HINDRAF)". In: International Journal of Communication 6 (2012), p. 24.

[123] Johannes Vüllers and Birte Pfeiffer. "Measuring the Ambivalence of Religion: Introducing the Religion and Conflict in Developing Countries (RCDC) Dataset". In: International Interactions 41.5 (2015), pp. 857–881.

[124] Matthias Basedau, Birte Pfeiffer, and Johannes Vüllers. "Bad religion? Religion, collective action, and the onset of armed conflict in developing countries". In: Journal of Conflict Resolution 60.2 (2016), pp. 226–255.

[125] Carlos Ivan Chesnevar et al. "Opinion Aggregation and Conflict Resolution in E-Government Platforms: Contrasting Social Media Information". In: Interdisciplinary Perspectives on Contemporary Conflict Resolution (2016), p. 183.

[126] Hannah Miller et al. "Blissfully happy or ready to fight: Varying Interpretations of Emoji". In: Proceedings of ICWSM 2016 (2016).

[127] Yla R Tausczik and James W Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods". In: Journal of language and social psychology 29.1 (2010), pp. 24–54.

[128] Ian Jolliffe. Principal component analysis. Wiley Online Library, 2002.

[129] Mark A Hall and Geoffrey Holmes. "Benchmarking attribute selection techniques for discrete class data mining". In: IEEE Transactions on Knowledge and Data engineering 15.6 (2003), pp. 1437–1447.

[130] Niamh Russell, Laura Cribbin, and Thomas Brendan Murphy. "upclass: An R Package for Updating Model-Based Classification Rules". In: (2012).

[131] Nema Dean, Thomas Brendan Murphy, and Gerard Downey. "Using unlabelled data to update classification rules with applications in food authenticity studies". In: Journal of the Royal Statistical Society: Series C (Applied Statistics) 55.1 (2006), pp. 1–14.

[132] Ying Chen et al. "Detecting offensive language in social media to protect adolescent online safety". In: Privacy, Security, Risk and Trust (PASSAT), SocialCom. IEEE. 2012, pp. 71–80.

[133] Derek O'Callaghan et al. "Uncovering the wider structure of extreme right communities spanning popular online networks". In: Proceedings of the 5th Annual ACM Web Science Conference. ACM. 2013, pp. 276–285.

[134] Irene Kwok and Yuzhou Wang. "Locate the Hate: Detecting Tweets against Blacks." In: AAAI. 2013.

[135] Swati Agarwal and Ashish Sureka. "A Focused Crawler for Mining Hate and Extremism Promoting Users, Videos and Communities on YouTube". In: Proceedings of 25th ACM Conference on Hypertext and Social Media (HT), Santiago, Chile. 2014.

[136] Tianjun Fu, Chun-Neng Huang, and Hsinchun Chen. "Identification of extremist videos in online video sharing sites". In: Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on. IEEE. 2009, pp. 179–181.

[137] Swati Agarwal and Ashish Sureka. "Learning to Classify Hate and Extremism Promoting Tweets". In: Proceedings of IEEE Joint Intelligence and Security Informatics Conference (EISIC+ ISI), the Hague, the Netherlands. 2014.

[138] Jinpeng Wang et al. "Mining New Business Opportunities: Identifying Trend related Products by Leveraging Commercial Intents from Microblogs." In: EMNLP. 2013, pp. 1337–1347.

[139] I-Hsien Ting et al. "An Approach for Hate Groups Detection in Facebook". In: The 3rd International Workshop on Intelligent Data Analysis and Management. 2013, pp. 101–106.

[140] Matthew Goodwin. The Roots of Extremism: The English Defence League and the Counter-Jihad Callenge. Chatham House, 2013.

[141] Ashish Sureka et al. "Mining YouTube to Discover Extremist Videos, Users and Hidden Communities". In: Information Retrieval Technology. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, pp. 13–24.

[142] Hsinchun Chen. "Extremist YouTube Videos". In: Dark Web. Vol. 30. Integrated Series in Information Systems. 2012, pp. 295–318.

[143] Hsinchun Chen et al. "The Dark Web Forum Portal: From multi-lingual to video." In: ISI. IEEE, 2011, pp. 7–14.

[144] Hsinchun Chen et al. "Chapter 1 - Revealing the Hidden World of the Dark Web: Social Media Forums and Videos". In: Intelligent Systems for Security Informatics. Academic Press, 2013, pp. 1 –28. ISBN: 978-0-12-404702-0.

[145] Tianjun Fu and Hsinchun Chen. "Knowledge Discovery and Text Mining". In: ().

[146] Edna Reid and Hsinchen Chen. "Internet-savvy US and Middle Eastern extremist groups". In: Mobilization: An International Quarterly 12.2 (2007), pp. 177–192.

[147] Arab Salem, Edna Reid, and Hsinchun Chen. "Content Analysis of Jihadi Extremist Groups' Videos". In: Intelligence and Security Informatics. Springer Berlin Heidelberg, 2006, pp. 615–620. URL: http://dx.doi.org/10.1007/11760146_66.

[148] Shah Mahmood. "Online social networks: The overt and covert communication channels for terrorists and beyond". In: Homeland Security (HST), 2012 IEEE Conference on Technologies for. IEEE. 2012, pp. 574–579.

[149] David Décary-Hétu and Carlo Morselli. "Gang Presence in Social Network Sites." In: International Journal of Cyber Criminology 5.2 (2011).

[150] Jinpeng Wang, Gao Cong, and et al. "Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets". In: AAAI. 2015.

[151] Hemant Purohit, Guozhu Dong, and et al. "Intent Classification of Short-Text on Social Media". In: SocialCom 2015.

[152] Xiao Ding et al. "Mining User Consumption Intention from Social Media Using Domain Adaptive Convolutional Neural Network." In: AAAI. 2015, pp. 2389–2395.

[153] P Geetha, RM Chandresh, and et al. "Feature Selection Framework for Data Analytics in Microblogs". In: 'Emerging Research in Computing, Information, Communication and Applications' ERCICA. 2014.

[154] Allison G Smith et al. "The language of violence: Distinguishing terrorist from nonterrorist groups by thematic content analysis". In: Dynamics of Asymmetric Conflict 1.2 (2008), pp. 142–163.

[155] Rui Li, Shengjie Wang, and Kevin Chen-Chuan Chang. "Towards social data platform: automatic topic-focused monitor for twitter stream". In: Proceedings of the VLDB Endowment 6.14 (2013), pp. 1966–1977.

[156] Rui Li et al. "Towards social user profiling: unified and discriminative influence model for inferring home locations". In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2012, pp. 1023–1031.

[157] Pascal Soucy and Guy W Mineau. "A simple KNN algorithm for text categorization". In: Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE. 2001, pp. 647–648.

[158] Daniel T Munroe and Michael G Madden. "Multi-class and single-class classification approaches to vehicle model recognition from images". In: Proceedings of IEEE AICS (2005).

[159] Henk Wolda. "Similarity indices, sample size and diversity". In: Oecologia 50.3 (1981), pp. 296–302.

[160] D Martinus and J Tax. "One-class classification: Concept-learning in the absence of counterexamples". PhD thesis. PhD thesis, Delft University of Technology, 2001.

[161] Larry M Manevitz and Malik Yousef. "One-class SVMs for document classification". In: Journal of Machine Learning Research 2.Dec (2001), pp. 139–154.

[162] Thorsten Joachims. Learning to classify text using support vector machines: Methods, theory and algorithms. Kluwer Academic Publishers, 2002.

[163] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A library for support vector machines". In: ACM Transactions on Intelligent Systems and Technology 2 (3 2011). Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 27:1–27:27.

[164] Sylvain Arlot, Alain Celisse, et al. "A survey of cross-validation procedures for model selection". In: Statistics surveys 4 (2010), pp. 40–79.

[165] Swati Agarwal and Ashish Sureka. "Spider and the Flies: Focused Crawling on Tumblr to Detect Hate Promoting Communities". In: arXiv preprint arXiv:1603.09164 (2016).

[166] Pete Burnap and Matthew L Williams. "Us and them: identifying cyber hate on Twitter across multiple protected characteristics". In: EPJ Data Science 5.1 (2016), p. 1.

[167] Lacy G McNamee, Brittany L Peterson, and Jorge Peña. "A call to educate, participate, invoke and indict: Understanding the communication of online hate groups". In: Communication Monographs 77.2 (2010), pp. 257–280.

[168] Fuchun Peng, Dale Schuurmans, and Shaojun Wang. "Language and task independent text categorization with simple language models". In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics. 2003, pp. 110–117.

[169] S. Rawat and D.R. Patil. "Efficient focused crawling based on best first search". In: Advance Computing Conference (IACC), 2013 IEEE 3rd International. 2013, pp. 908–911. DOI: 10.1109/IAdCC.2013.6514347.

[170] Michael Hersovici et al. "The shark-search algorithm. An application: tailored Web site mapping". In: Computer Networks and ISDN Systems 30.1 (1998), pp. 317–326.

[171] Ben Carterette. "Precision and Recall". In: Encyclopedia of Database Systems. Springer, 2009, pp. 2126–2127.

[172] Judith A Effken et al. "Using ORA to explore the relationship of nursing unit communication to patient safety and quality outcomes". In: International journal of medical informatics 80.7 (2011), pp. 507–517.

[173] W Keith Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: Biometrika 57.1 (1970), pp. 97–109.

[174] SW Fong, DB Skillicorn, and D Roussinov. "Detecting word substitution in adversarial communication". In: 6th SIAM Conference on Data Mining. Bethesda, Maryland. 2006.

[175] Dmitri Roussinov, SzeWang Fong, and David Skillicorn. "Detecting Word Substitutions: PMI vs. HMM". In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '07. 2007, pp. 885–886.

[176] SW. Fong, D. Roussinov, and D.B. Skillicorn. "Detecting Word Substitutions in Text". In: IEEE Transactions on Knowledge and Data Engineering 20.8 (2008), pp. 1067–1076.

[177] Ben Allison Sanaz Jabbari and Louise Guthrie. "Using a Probabilistic Model of Context to Detect Word Obfuscation". In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08) (2008).

[178] Sonal N. Deshmukh, Ratnadeep R. Deshmukh, and Sachin N. Deshmukh. "Performance Analysis of Different Sentence Oddity Measures Applied on Google and Google News Repository for Detection of Substitution". In: International Refereed Journal of Engineering and Science (IRJES) 3.3 (2014), pp. 20–25.

[179] Ronen Yitzhak. "The War Against Terrorism and For Stability of the Hashemite Regime: Jordanian Intelligence Challenges in the 21st Century". In: International Journal of Intelligence and CounterIntelligence 29.2 (2016), pp. 213–235.

[180] Hugo Liu and Push Singh. "ConceptNeta practical commonsense reasoning tool-kit". In: BT technology journal 22.4 (2004), pp. 211–226.

[181] Robert Speer and Catherine Havasi. "ConceptNet 5: A large semantic network for relational knowledge". In: The Peoples Web Meets NLP. Springer, 2013, pp. 161–176.

[182] Chi-En Wu and Richard Tzong-Han Tsai. "Using relation selection to improve value propagation in a ConceptNet-based sentiment dictionary". In: Knowledge-Based Systems (2014).

[183] Arbi Bouchoucha, Xiaohua Liu, and Jian-Yun Nie. "Integrating Multiple Resources for Diversified Query Expansion". In: Advances in Information Retrieval (2014), pp. 437–442.

[184] R Akileshwari, S Revathi, and A Grace Selvarani. "A Novel Approach for Similarity Based Video Annotation Utilizing Commonsense Knowledgebase". In: ().

[185] Nakatani Shuyo. Language Detection Library for Java. 2010. URL: `http://code.google.com/p/language-detection/`.

[186] Ting Hua et al. "Analyzing civil unrest through social media". In: Computer 46.12 (2013), pp. 80–84.

[187] Naren Ramakrishnan et al. "'Beating the news' with EMBERS: forecasting civil unrest using open source indicators". In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2014.

[188] Sathappan Muthiah et al. "Planned Protest Modeling in News and Social Media". In: Proceedings of the 29th Association for the Advancement of Artificial Intelligence. AAAI '15. Austin, Texas, USA, 2015.

[189] Andrey A. Filchenkov, Artur A. Azarov, and Maxim V. Abramov. "What is More Predictable in Social Media: Election Outcome or Protest Action?" In: Proceedings of the 2014 Conference on Electronic Governance and Open Society: Challenges in Eurasia. ACM, 2014, pp. 157–161.

[190] Liang Zhao et al. "Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling". In: PloS one 9.10 (2014), e110206.

[191] Ryan Compton et al. "Detecting future social unrest in unprocessed twitter data: emerging phenomena and big data". In: ISI, 2013 IEEE International Conference On. 2013.

[192] Jiejun Xu et al. "Civil unrest prediction: A tumblr-based exploration". In: Social Computing, Behavioral-Cultural Modeling and Prediction. Springer, 2014.

[193] Thanassis Tiropanis et al. "The web science observatory". In: IEEE Intelligent Systems 28.2 (2013), pp. 100–104.

[194] Ceren Budak et al. "Geoscope: Online detection of geo-correlated information trends in social networks". In: Proceedings of the VLDB Endowment 7.4 (2013), pp. 229–240.

[195] Alan Agresti and Maria Kateri. Categorical data analysis. Springer, 2011.

[196] Daniel A Keim. "Information visualization and visual data mining". In: IEEE transactions on Visualization and Computer Graphics 8.1 (2002), pp. 1–8.

[197] Jean-Daniel Fekete et al. "The value of information visualization". In: Information visualization. Springer, 2008, pp. 1–18.

[198] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: Nature 521.7553 (2015), pp. 436–444.

[199] Christine Chin and David E Brown. "Learning in science: A comparison of deep and surface approaches". In: Journal of research in science teaching 37.2 (2000), pp. 109–138.

[200] Gerardo Patriotta. "Cities of Noise: Sensemaking, Sensemakers, and Organized Worlds". In: Academy of Management Review 41.3 (2016), pp. 557–570.

[201] Heather C Vough and Brianna Barker Caza. "Where do I go from here? Sensemaking and the Construction of Growth-Based Stories in the Wake of Denied Promotions". In: Academy of Management Review 42.1 (2017), pp. 103–128.

[202] Zachary Alfred. "Tweeting against corruption: Fighting police bribery through online collective action". In: (2014).

[203] Thomas Barnebeck Andersen. "E-Government as an anti-corruption strategy". In: Information Economics and Policy 21.3 (2009), pp. 201–210.

[204] Chengcheng Shao et al. "Hoaxy: A platform for tracking online misinformation". In: Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee. 2016, pp. 745–750.

[205] Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. Tech. rep. National Bureau of Economic Research, 2017.

[206] Myron Flickner et al. "Query by image and video content: The QBIC system". In: computer 28.9 (1995), pp. 23–32.

[207] Anil K Jain and Bin Yu. "Automatic text location in images and video frames". In: Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on. Vol. 2. IEEE. 1998, pp. 1497–1499.

[208] Xiaoping Zhou et al. "Cross-platform identification of anonymous identical users in multiple social media networks". In: IEEE transactions on knowledge and data engineering 28.2 (2016), pp. 411–424.

[209] Neal Coulter, Ira Monarch, and Suresh Konda. "Software engineering as seen through its research literature: A study in co-word analysis". In: Journal of the American Society for Information Science 49.13 (1998), pp. 1206–1223.

[210] Richard Colbaugh and Kristin Glass. "Early warning analysis for social diffusion events". In: Security Informatics 1.1 (2012), pp. 1–26.

[211] Alan Ritter, Sam Clark, Oren Etzioni, et al. "Named entity recognition in tweets: an experimental study". In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2011, pp. 1524–1534.

[212] Feng Chen and Daniel B. Neill. "Non-parametric Scan Statistics for Event Detection and Forecasting in Heterogeneous Social Media Graphs". In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014.

[213] Jennifer Xu and Michael Chau. "Mining communities of bloggers: A case study on cyberhate". In: ICIS 2006 Proceedings (2006), p. 11.

[214] Ming Yang and Hsinchun Chen. "Partially supervised learning for radical opinion identification in hate group web forums". In: Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on. 2012, pp. 96–101.

[215] Anneli Botha. "Assessing the vulnerability of Kenyan youths to radicalisation and extremism". In: Institute for Security Studies Papers (2013), 28–p.

[216] Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen. “A focused crawler for Dark Web forums”. In: Journal of the American Society for Information Science and Technology 61.6 (2010), pp. 1213–1231.