140 Characters of @Hate and #Protest

Ashish Sureka

Indraprastha Institute of Information Technology, Delhi (IIIT-D)
New Delhi - 110078 (India)
Email: ashish@iiitd.ac.in

***Abstract*—Research shows that cyber-hate, illegal or malicious form of cyber-protest and cyber-activism in online social media and Web 2.0 platforms has become a common phenomenon. This is a growing concern and hence automated techniques to counter such forms of online propaganda and identification of users and virtual communities is an area which has recently attracted a lot of research attention. In this paper, we present a simple and effective method to mine Twitter (a very popular and largest micro-blogging service) data for automatically identifying users and communities having a shared agenda. We propose a generic framework that consists of a systematic and focused traversal of the follower-network on Twitter and a user profile classifier based on content-based features. We customize the proposed framework for a specific domain and demonstrate that the proposed approach is effective. We perform empirical analysis on data crawled from Twitter and the experimental results on the test dataset reveals that the proposed features and framework can successfully identify twitterers and hidden communities having a common agenda or shared interest.**

***Keywords*-Information Retrieval, User Modeling, Text Classification, Twitter, Social Media Analytics, Cyber-Activism, Cyber-Hate, Cyber-Protest**

I. Research Motivation and Aim

Cyber-Terrorism, Online Radicalization and Extremism, Cyber-Racism, Cyber-Hate is widespread and has become a major and growing concern to the society, governments and law enforcement agencies around the world [3][4][5][12]. Research shows that social media websites and platforms such as YouTube (a popular video sharing website), Twitter (an online micro-blogging service), FaceBook (a popular social networking website), online discussion forums and blogosphere are being misused for malicious intent by hate groups, racist communities, extremists and terrorists to spread their agenda, incite anger or violence, recruit members and create virtual organizations and communities [2][13]. Such Web 2.0 platforms (which has very low barrier to publish content, allows anonymity and a very quick and widespread spread of message) are also being widely used for online social activism and Cyber-Protests [8][14].

Twitter (defined as real-time information network by the company[1]) is a very popular and fast growing micro-blogging service which allows users to post stream of messages (of up-to 140 characters in length called as tweets) and follow or subscribe to tweets of other users (called as Twitterers). Twitter has created a phenomenon social impact and according to the numbers mentioned in the company blog[2], there were 140 million average number of tweets posted per day during the month of February 2011 and 460,000 average number of new accounts created per day during the month of February 2011.

There are several noticeable incidents where Twitter (the focus area of this paper) is employed as a powerful and effective tool for online activism and cyber-protest. Due to limited space in this paper, we present two recent (and highly impactful) examples. The book by Guobin Yang on the power of online activism describes an incident where two tweets on Twitter by a person arrested unfairly on alleged defamation charges in China triggered such a quick, widespread and massive reaction around the world (quoted as: this communication revolution is a social revolution) that it finally resulted in freeing and saving the life of the Twitterer [7]. Gaffney et al. perform a retrospective analysis of the Iran national election protests on the Twitter platform and describes how Twitter played a very important role in allowing dissidents communicated with the international audience, news services and amongst themselves [6].

While freedom of expression and speech is a fundamental right of every human, several political activism and expressions takes the shape of propaganda, dissemination of hate, racism, inciting anger and violence. Solutions to counter cyber-crime related to promotion of hate and radicalization on the Internet is an area which has recently attracted a lot of research attention. Furthermore, techniques to analyze and understand various aspects of legal online activism and cyber protest is a field which is useful and timely in current context. The *broad research motivation* of the work presented in this paper is following:

- To investigate techniques for mining user-generated content in online social media platforms for deriving useful insights, actionable information and patterns for law enforcement, security analyst and a political

[1] http://twitter.com/about

[2] http://blog.twitter.com/2011/03/numbers.html

analyst.

The information need of an analyst (law-enforcement, security, policy makers, government, intelligence) is to identify users, content and virtual communities (hidden social networks of people with the shared agenda and interest) promoting their ideologies and agenda. Keeping in mind the stated information need, the *specific research aim* of the work presented in this paper is the following:

- To investigate textual content based information retrieval models and algorithms for mining Twitter data for identifying cyber-activism and hate-promoting Twitterers and hidden communities for a pre-defined topic or domain.

The stated problem (research aim) is a technically challenging task due to the vastness of Twitter repository, noisy text, inherent natural language ambiguities, dynamic nature of the website, presence of a large number of users profiles, interests and various kinds of relationships between users.

II. Research Contributions

In this Section, we compare and contrast the research presented in this paper with closely related work and list the key differences. Furthermore, we explicitly state the novel research contributions of this work. Sureka et al. present a method to discover hate videos, users and virtual hidden communities in YouTube (largest video sharing website on Internet) [13]. The main difference between the work by Sureka et al. and this paper is the Web 2.0 platform under study (previous work is on YouTube and this paper is on Twitter). We believe that there are several fundamental differences between YouTube and Twitter posing different and unique platform-specific technical challenges. Chau et al. present a semi-automated approach to analyze and discover virtual hate communities in blogosphere [2]. Purohit et al. study user engagement, network topology, human social dynamics, and user communities around topics on Twitter [10]. Michelson present a method to discover Twitter users' topics of interest by examining the entities they mention in their Tweets [9]. Banerjee et al. explore how users express interests in real-time through micro-blogs and apply text mining techniques to interpret real-time context of a user based on her tweets [1].

In context to closely related work, this paper makes the following novel and *unique contributions*. To the best of our knowledge, this paper presents the *first study* (on Twitter) of mining Tweets to automatically identify cyber-activists and protestors (legal) and/or hate-promoters (illegal). The proposed framework (described in detail in subsequent section) consisting of user intent classification based on interplay of features such as presence of certain pre-defined hashtags, frequent world-level n-grams and user mentions is novel in context to the body of research work on Twitter and application of information retrieval models for intelligence and security related applications. This paper presents an empirical study on dataset crawled from Twitter and demonstrates that the proposed method is effective. The research presented in this paper offers fresh perspective and throws light on the topic of user and community detection for intelligence and security application on Twitter and thus adds to the growing body work on the topic.

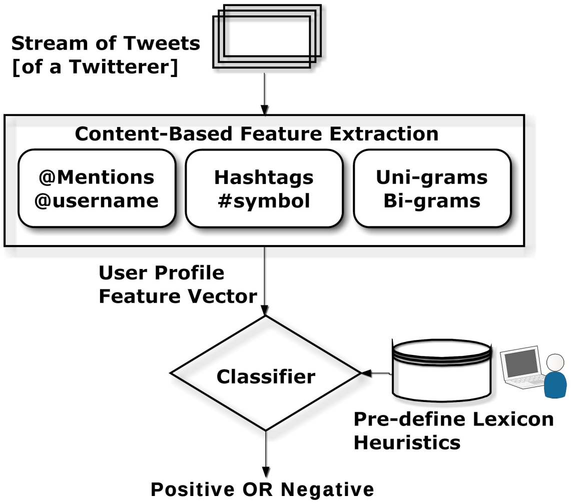

Figure 1. Component diagram of user intent classifier showing the content-based discriminatory features and heuristics module used for the given categorization task

III. Solution Approach

Figure 1 and 2 describes the proposed solution framework. The two key components of the proposed system are: user intent classifier based on content-based discriminatory features and breadth-first traversal or expansion (on the follower network) of a given seed node (twitterer categorized as belonging to the positive class). Since one of the goals is to devise a classifier to categorize cyber-activist, cyber-protestor or hate-promoter, we perform a careful manual and visual inspection of such user profiles and derive discriminatory features. In this work, we do not discriminate between a legal or illegal remarks or protest (as it can be very subjective and there can be disagreements between the experts also) and define the target class as a set of users having a clear agenda and interest of disseminating their strong opinions on a given issue. As shown in Figure 1, the classifier takes a stream of tweets (document collection) for a particular user and first performs the step of feature extraction. The classifier then applies hand-crafted heuristic or rule (derived or induced from the dataset and observations) to label the user as positive or negative. Identifying distinguishing features (having discriminatory power) is central to the effective of the classifier and we propose the following features after careful manual inspection of user tweets belonging to both

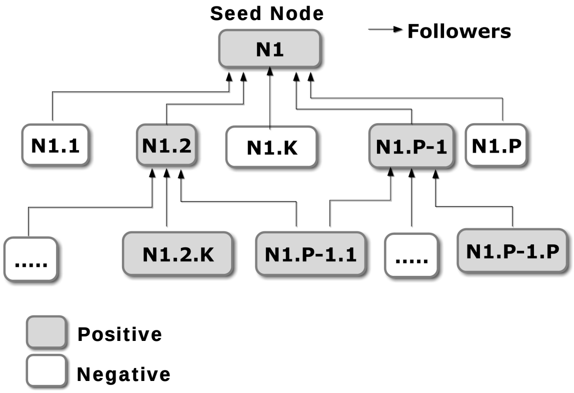

Figure 2. An illustrative example of a users follower-network (follower, follower of a follower) denoting nodes belonging to the target class (shaded)

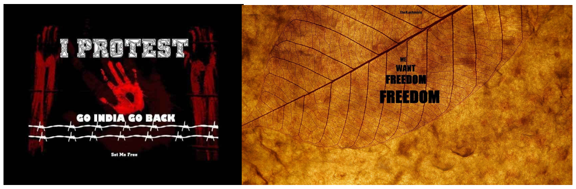

Figure 3. Profile background of twitterers classified as cyber-protestor, cyber-activist or hate-promoter by the proposed system. The picture and wording (I Protest, Go India Go Back, Freedom, I Want Freedom) in the user profile background image clearly indicates the users agenda

the positive as well as the negative class.

A. Frequent word-level n-grams

We derive frequent world-level unigram and bigrams from a collection of tweets for a given user. This design decision is based on the premise that the terms used by a user are a reflection of the user's interest and topic. We filter stop words (such as and, the, for, of etc) from the list of frequent unigrams as they are not content-bearing. The list of Top N (where N is a tuning parameter of the solution framework) frequent unigrams and bi-grams represents one of the features for the classifier.

B. Hashtags

Twitter defines hashtags[3] as keywords or topic marked by a user in a tweet. Hashtags are also called as # symbol (and hence the symbol # in the title of this paper) and is used to categorize a message by the twitterer. Hashtags can occur anywhere within the message body and this feature enables convenient searching as users can click on a hashtagged word in a message to retrieve other tweets in that category. Based on our manual inspection of the tweets belonging to the domain of cyber protest and hate, we notice that hashtags have discriminatory power. For example, we notice frequent usage of hashtags like #protest, #freedom, #democracy, #shame (reflecting activism sentiments) and country, religion, region or person names in tweets which belong to cyber-protest, cyber-activism and cyber-hate domain.

C. Mentions

Twitter defines mentions[4] as any Twitter update that contains @username (and hence the @ symbol in the title of this paper) anywhere in the body of the Tweet. Replies (inserting @username of the person replying to) are also considered mentions. More than one person's name can be mentioned in a single tweet and based on our manual inspection of tweets belonging to positive as well as negative class, we hypothesize that mentions can be used as indicators for the classification task. For example, the mention @facebook in a tweet signals that the content of the tweet has some relation with Facebook.

D. Follower-Network Traversal

Twitter has a feature called as following[5] where a user can subscribe to the tweets of another user (reflecting interest in the updates of a user by the follower). In Twitter, the following relation is not mutual and can be one-way. Figure 2 illustrates the breadth-first traversal strategy employed by the proposed approach. Figure 2 displays a network where the node $N1$ belongs to the positive class (seed node provided by the user) and consists of P followers. Each node (user profile represented as a feature vector) is categorized as positive or negative by the classifier (in Figure 2, shaded nodes are members of the positive-class and non-shaded nodes belong to the negative class). Each node is then further expanded in the same fashion and the procedure is applied up-to a depth d (pre-defined by the analyst) from the root node or terminated when there are no nodes that needs to be expanded. We hypothesize that a community of users having common interest can be discovered by applying the mentioned procedure.

E. Worked-Out Example

Kashmir is a region in the Indian subcontinent which had faced a very high magnitude of terrorism, human rights violations, genocides, militant extremism, border tensions, conflicts, wars, disputes, hate-crime, protest, activism and crisis to the extent that it has resulted in deaths of hundreds and thousands of people [11]. Table I displays profile description, hashtags, user mentions, frequent n-grams of some Twitterers belonging to the positive (cyber-protestor, cyber-activist or hate-promoter) and negative class. The root (labeled as N1 in Table I) is manually searched and acts as the seed for the breadth-first search on the follower-network. The presence of terms (which are dominant and frequent) like killed, forces, human rights, army, troops,

[3] http://support.twitter.com/entries/49309-what-are-hashtags-symbols

[4] http://support.twitter.com/entries/14023-what-are-replies-and-mentions

[5] http://support.twitter.com/articles/14019-what-is-following

Table I

TABLE DISPLAYS PROFILE DESCRIPTION, HASHTAGS, USER MENTIONS, FREQUENT N-GRAMS OF SOME TWITTERERS BELONGING TO THE POSITIVE (CYBER-PROTESTOR, CYBER-ACTIVIST OR HATE-PROMOTER) AND NEGATIVE CLASS. NOTE: KASHMIR IS A REGION IN INDIAN SUBCONTINENT WHICH HAS BEEN UNDER DISPUTE FOR PAST SEVERAL DECADES AND HAS FACED LOT OF TERRORISM, GENOCIDE AND EXTREMISM

Node ID	**Description**	**N-Grams**	**Hashtags**	**Mentions**	**Category**
N1	Indian Occupied Kashmir, Dreamer of Independent Kashmir and Palestine	india, indian, people, ppl, kashmir, killed, govt, forces, tera bharat, human rights indian army, indian troops, indian forces	srinagar, shame, sad, revolution, protest, press, palestine, onekashmir, kashmiri, kashmiris, india, freekashmir, freedom, fail	thekashmiris, kashmirtimes, kashmirfollow, kashmirdispatch, irkashmir, iluvkashmir, _kashmir	Positive
N1.1 [Follower Of N1]	-	india, indian, kashmir, kashmiris, truth, killed, media, freedom, army, forces, indian media, indian forces, youth killed	Srinagar, protest, Kashmir, pakistan, india, Indian, freedom, democracy	Killthebillion, kashmirtimes, kashmirdispatch	Positive
N1.1.1 [Follower of N1.1.]	For the Kashmir, Of the Kashmir,by the Kashmiris	kashmir, india, Indian, people, govt, india Pakistan, atrocities Kashmir, hindu atrocities, human rights, Indian army	unfair, propaganda, political, palestine, pakistan, lawlesslaw, kashmiri, Kashmir, india, failed, srinagar, shame	thekashmiris, kashmirfirst, freekashmir, kashmirdispatch	Positive
N1.2 [Follower Of N1]	Daily Health, Weight loss tips Weight, loss, health, tips	skin, diet, food, fat, body, skin care, back pain, junk food	weightloss, obesity, nutrition, health, happiness, fitness, beautytip, antiaging	womenshealthmag, totalbeauty, dailyhealthtips	Negative

freedom, media, govt, truth, india, indian, atrocitites and kashmir clearly indicates the interest and agenda of the twitterer. We performed a manual and visual inspection of all the tweets of the user and the messages clearly shows activism, protest, anger and hate. The node labeled N1.1 (refer to Table I) is follower of the user labeled as N1. The feature vector or profile of the node N1.1 signals positive class. Similarly, node N1.1.1 also belongs to the positive class.

It is not necessary that a follower or subscriber will have the same interest as the user getting subscribed. Based on our manual inspection, we notice that there are several users who don't have any common interest with the user to whom they have subscribed. For example, the node labeled as N1.2 (follower of N1) in Table has absolutely no interest in topics related to the Node N1. Twitterer N1.2 describes his profile as daily health and weight loss tips having frequent n-grams, hashtags and mentions as skin, diet, food, fat, body, weightloss, obesity, nutrition, health, happiness, fitness, beautytip and antiaging. We visually inspected the tweets, URLs, entity mentions of these users (actual usernames are not mentioned in this paper due to privacy issues) and our manual inspection (first experiment) confirms our hypothesis that features like frequent world-level n-grams, presence of pre-defined terms in hashtags and mentions have discriminatory power and are characteristics of the positive class in this context. Furthermore, the profile description (refer to Table I) and the background user profile image (refer to Figure 3 and the Fifure caption) also indicates the motive of the twitterer

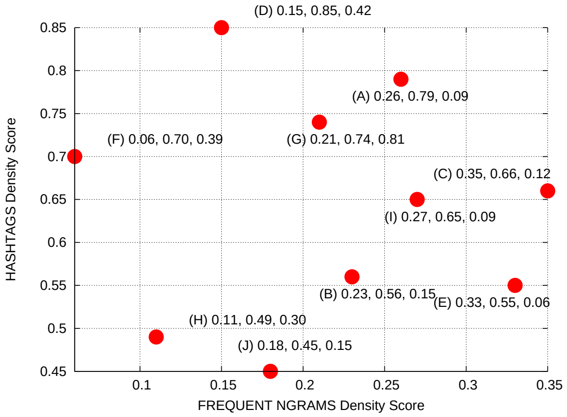

Figure 4. Plot of twitterers belonging to the target class along the axis of density score for frequent word-level n-grams in a users tweets and density score of hashtags given a pre-defined lexicon. Each user (represented as a feature vector having three dimensions) is labeled with the numerical value of each of the three features.

IV. EMPIRICAL ANALYSIS AND RESULTS

We implemented the proposed solution strategy (expanding a seed node along the follower network and applying a node classifier based on discriminatory features and pre-defined lexicon to detect cyber-protestors, cyber-activist and hate-promoting users and communities) using Twitter developers APIs [6] and client libraries[7] for Twitter. We

[6] http://dev.twitter.com/doc

[7] http://twitter4j.org/en/index.html

hand-crafted three simple lexicons (refer to the solution architecture diagram in Figure 1) one each for frequent world-level n-grams in tweets, hashtags and user mentions. The lexicon contains terms which are relevant to the specific topic (such as words like India, Pakistan, Kashmir, army, protest, freedom and killed). While our proposed framework is generic, we test our hypothesis by performing a case-study on activism and protests around the Kashmir conflict. The seed user and lexicon (one each for n-grams, hashtags and mentions) is an input parameter to the system.

We measure the density of the terms present in the lexicon in the user profile vector. The density is simply the percentage of terms in a user profile vector present in the lexicon. For example, let us say that the user profile vector consists of 10 n-grams, 10 hashtags and 10 user mentions (Top K where K=10). If out of 10, there are 5 hashtags which are present in the pre-defined lexicon then the density is 50% or 0.50. We defined a rule that if the sum (threshold value) of these densities is above a pre-defined percentage (a user defined parameter derived from experimenting with the system or can be learned using machine learning techniques). The threshold value is also dependent on the number and kind of terms in the lexicon (how restricted or relaxed the terms are with respect to a topic) and in this work we hard-code the classification function or heuristics (derived based on trial and error and making observations) as: (density of n-grams) + (density of hashtags) + 3*(density of mentions) ≥ 0.7. In the future, we plan to devise a more rigorous and scientific method of deriving the classification function.

Figure 4 displays a plot of 10 twitterers along the axis of density score for frequent n-grams and hastags. The 10 twitterers (we plot only 10 for the purpose of clarity and argument) were classified as positive (using the heuristic that the combined score of the three pre-defined feature value ≥ 0.7) by the proposed system for a given seed. We manually checked the profile and tweets of the users and validate that the output produced by the system is accurate. The classification of a user as cyber-protestors, cyber-activist or hate-promoter has naturally some degree of subjectivity but our manual inspection of the output produced by the system for various seeds validate that the three proposed features are reliable indicators. Each user in Figure 4 is labeled with all the three scores and we notice that a combination of all the three features is a robust and reliable indicator for the specific classification task. We also confirm the presence of a phenomenon where certain users are influential or authoritative nodes (followed by several users sharing the same interest and agenda) and we also observe a phenomenon where nodes belonging to target class are present 3 to 4 level down the seed node in the follower network.

V. Conclusions

We describe a generic framework to automatically identify cyber-protestors, cyber-activist and hate-promoters (motivating application is intelligence and security informatics, monitoring and surveillance) in Twitter by mining a twitterers tweets and using the follower relationship for discovering other twitters with a shared agenda or common interest. Applying the proposed method on a sample dataset reveals that the technique is effective. We hypothesize presence of certain characteristics and features of the target class and perform an empirical study to test the proposed hypothesis. Our findings indicate that attributes such presence of pre-defined terms in the frequent world-level n-grams of users tweets, hashtags and mentions are reliable indicators for the given categorization task.

References

[1] Nilanjan Banerjee, Dipanjan Chakraborty, Koustuv Dasgupta, Sumit Mittal, Anupam Joshi, Seema Nagar, Angshu Rai, and Sameer Madan. User interests in social media sites: an exploration with micro-blogs. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1823–1826, New York, NY, USA, 2009. ACM.

[2] Michael Chau and Jennifer Xu. Mining communities and their relationships in blogs: A study of online hate groups. *Int. J. Hum.-Comput. Stud.*, 65:57–70, January 2007.

[3] Jessie Daniels. *Cyber Racism: White Supremacy Online and the New Attack on Civil Rights (Pespectives on a Multiracial America Series).* Rowman & Littlefield Publishers, 2009.

[4] H.-L. Dienel, Y. Sharan, C. Rapp, and N. Ahituv. *Terrorism and the Internet: Threats - Target Groups - Deradicalisation Strategies.* IOS Press, 2010.

[5] Karen M. Douglas, Craig McGarty, Ana-Maria Bliuc, and Girish Lala. Understanding cyberhate: social competition and social creativity in online white supremacist groups. *Soc. Sci. Comput. Rev.*, 23:68–76, March 2005.

[6] Devin Gaffney. iranelection: quantifying online activism. *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, 2010.

[7] Theresa Loong. Guobin yang: The power of the internet in china: Citizen activism online. *Publishing Research Quarterly*, 26:305–307, 2010.

[8] MARTHA MCCAUGHEY and Michael D. Ayers. *Cyberactivism: Online Activism in Theory and Practice*. Routledge, 2003.

[9] Matthew Michelson and Sofus A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, AND '10, pages 73–80, New York, NY, USA, 2010. ACM.

[10] Hemant Purohit, Yiye Ruan, Amruta Joshi, Srinivasan Parthasarathy, and Amit Sheth. Understanding user-community engagement by multi-faceted features: A case study on twitter. *WWW 2011 workshop on Social Media Engagement (SoME)*, 2011.

[11] Victoria Schofield. Kashmir in conflict: India, pakistan and the unending war. *I. B. Tauris*, 2010.

[12] Philip Seib and Dana M. Janbek. *Global Terrorism and New Media: The Post-Al Qaeda Generation (Media, War and Security)*. Routledge, 2010.

[13] Ashish Sureka, Ponnurangam Kumaraguru, Atul Goyal, and Sidharth Chhabra. Mining youtube to discover extremist videos, users and hidden communities. 6458:13–24, 2010.

[14] Wim van de Donk, Brian D. Loader, Paul G. Nixon, and Dieter Rucht. *Cyberprotest: New Media, Citizens and Social Movements*. Routledge, 2004.