



INDRAPRASTHA INSTITUTE *of*  
INFORMATION TECHNOLOGY **DELHI**

# Techniques for Automating Quality Assessment of Context-specific Content on Social Media Services

By  
Prateek Dewan

Under the Supervision of Dr. Ponnurangam Kumaraguru

Indraprastha Institute of Information Technology - Delhi

August, 2017





INDRAPRASTHA INSTITUTE *of*  
INFORMATION TECHNOLOGY **DELHI**

# Techniques for Automating Quality Assessment of Context-specific Content on Social Media Services

By  
Prateek Dewan

Submitted  
in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy

to the

Indraprastha Institute of Information Technology Delhi  
August, 2017

# Certificate

This is to certify that the thesis titled “**Techniques for Automating Quality Assessment of Context-specific Content on Social Media Services.**” being submitted by **Prateek Dewan** 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 him 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.

Supervisor Name: Dr. Ponnurangam Kumaraguru, “PK”

August, 2017

Department of Computer Science

Indraprastha Institute of Information Technology Delhi

New Delhi 110 020

Keywords: Online Social Networks, Machine learning, poor quality content, real-time system

## Abstract

Online Social Networks have become a cornerstone of Web 2.0 era. Internet users around the world use Online Social Networks as primary sources to consume news, updates, and information about events around the world. However, given the enormous volume and veracity, it is hard to manually moderate all content that is generated and shared on these networks. This phenomenon enables hostile entities to generate and promote various types of poor quality content (including but not limited to scams, fake news, false information, rumors, untrustworthy or unreliable information) and pollute the information stream for monetary gains, hinder user experience, or to compromise system reputation. We aim to address this challenge of automatically identifying poor quality content on Online Social Networks. We focus our work on Facebook, which is currently the biggest Online Social Network.

We provide an in-depth analysis of poor quality context-specific content published on Facebook. In particular, we concentrate on content generated in the context of news-making events. We propose and evaluate automated techniques to identify and mitigate the spread of such poor quality content on Facebook in real-time.

The main contributions of this work are: (a) we characterized and analyzed poor quality, context-specific content generated and disseminated on Facebook during news-making events, with the purpose of identifying characteristics that differentiate it from benign content, (b) we showed the effectiveness of our automated techniques to identify poor quality content on Facebook using content-level features combined with metadata, and temporal activity, and (c) we developed and deployed a real-world solution for identifying poor quality context-specific content published on Facebook. We evaluated the efficiency of this real-time system with a live deployment used by actual Facebook users.

First, we analyzed Facebook data for 19 global news-making events from 2013-2015 for the spread of untrustworthy content, scams, self-promotion posts, fake information, adult content, etc. Some of the prominent events we analyzed are the Paris Attacks (2015), FIFA World Cup (2014), Boston Marathon Blasts (2013), Death of Nelson Mandela (2013), Birth of the first Royal Baby (2013). We identified over 11 thousand Facebook posts promoting untrustworthy information, child-unsafe content, scams, hate speech, and spam. We performed an in-depth analysis of how this poor quality content differs from benign content and identified characteristics that differentiate entities posting poor quality content from entities posting benign content. Second, we showed how features from user-generated content, combined with meta information, and temporal behavior can be used to identify poor quality content during events effectively. Third, we proposed and evaluated automated techniques to identify poor quality content and entities using supervised learning techniques. We

developed and deployed Facebook Inspector (FbI), a real-time system to identify poor quality content on Facebook during events. Facebook Inspector is available as a browser plug-in, and has been downloaded more than 5,000 times. The system has a daily audience of over 250 Facebook users. During 20 months of its deployment, Facebook Inspector has received over 7.4 million requests and has evaluated over 2.8 million public Facebook posts, allowing us to evaluate its performance and usability.

*Anyone can do my job, but no one can be me.*

~ Harvey Specter, Pearson Specter Litt



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Context-specific content . . . . .	1
1.2	Social Media Service . . . . .	2
1.3	Thesis statement . . . . .	4
1.4	Thesis contribution . . . . .	4
1.4.1	Theoretical . . . . .	5
1.4.2	Practical . . . . .	5
1.5	Thesis roadmap . . . . .	6
<b>2</b>	<b>Background and Literature Review</b>	<b>7</b>
2.1	Facebook: Terminology, building blocks, and characterization . . . . .	7
2.1.1	Terminology and Building Blocks . . . . .	7
2.1.2	Characterization of the Facebook network . . . . .	10
2.1.3	Facebook crawling . . . . .	11
2.2	Poor quality content on Facebook . . . . .	12
2.2.1	Types of poor quality content on Facebook . . . . .	13
2.2.2	Techniques for identifying poor quality content on Facebook . . . . .	15
2.3	Context-specific content analysis on other social networks . . . . .	18
2.3.1	Event analysis on Twitter . . . . .	19
2.3.2	Event analysis on Facebook, Google+, and other online social networks . . . . .	21
2.4	Facebook Limitations and Challenges . . . . .	22
2.4.1	Fine grained privacy settings . . . . .	22

2.4.2	Technical limitations . . . . .	24
2.5	Discussion and Research Gaps . . . . .	26
<b>3</b>	<b>Towards Automatic Real Time Identification of Malicious Content on Facebook</b>	<b>28</b>
3.1	Introduction . . . . .	28
3.2	Methodology . . . . .	30
3.2.1	Data collection . . . . .	31
3.2.2	Labeled dataset creation . . . . .	31
3.3	Analysis: Dataset I . . . . .	36
3.3.1	Efficiency of Facebook’s current techniques . . . . .	36
3.3.2	Key characteristics of posts containing malicious URLs . . . . .	37
3.4	Analysis: Dataset II . . . . .	41
3.4.1	Textual content . . . . .	42
3.4.2	Entities posting malicious content . . . . .	43
3.4.3	Metadata . . . . .	43
3.5	Detecting malicious content automatically . . . . .	44
3.5.1	Dataset I . . . . .	45
3.5.2	Dataset II . . . . .	52
3.5.3	Dataset I versus Dataset II . . . . .	54
3.6	Discussion, Limitations, and Future work . . . . .	55
<b>4</b>	<b>Facebook Inspector: Implementation and Evaluation</b>	<b>57</b>
4.1	Implementation . . . . .	57
4.1.1	Back-end . . . . .	57
4.1.2	Front-end . . . . .	60
4.2	Evaluation . . . . .	61
4.2.1	Response Time . . . . .	61
4.2.2	Performance . . . . .	62
4.2.3	Usability . . . . .	63
4.3	Discussion, Limitations and Future Work . . . . .	64

<b>5</b>	<b>Hiding in Plain Sight: The Anatomy of Malicious <i>Pages</i> on Facebook</b>	<b>66</b>
5.1	Introduction . . . . .	66
5.2	Scope and data collection . . . . .	69
5.2.1	Scope . . . . .	69
5.2.2	Establishing ground truth . . . . .	70
5.2.3	Dataset . . . . .	70
5.3	Malicious pages on Facebook . . . . .	73
5.3.1	Spatial behavior . . . . .	74
5.3.2	Temporal behavior . . . . .	81
5.4	Automatic detection of malicious pages . . . . .	86
5.4.1	Supervised learning with <i>page</i> and post features . . . . .	87
5.4.2	Supervised learning with bag-of-words . . . . .	88
5.5	Future work . . . . .	92
5.6	Conclusion . . . . .	93
<b>6</b>	<b>Towards Understanding Crisis Events On Online Social Networks Through Pictures</b>	<b>95</b>
6.1	Introduction . . . . .	95
6.2	Related Work . . . . .	97
6.2.1	Images on OSNs during crisis events . . . . .	97
6.2.2	Crisis event related studies on OSNs . . . . .	98
6.3	Methodology . . . . .	99
6.3.1	Data collection . . . . .	99
6.3.2	Image characterization . . . . .	100
6.4	Analysis and Results . . . . .	104
6.4.1	Top visual themes featured misinformative images . . . . .	105
6.4.2	Text embedded in images featured sensitive topics and reflected negative sentiment . . . . .	106
6.4.3	Images inspired positive sentiment . . . . .	109
6.5	Case Studies: Instances of Misinformation . . . . .	112

6.6	Discussion, Limitations, and Future work . . . . .	116
<b>7</b>	<b>Analyzing Social and Stylometric Features to Identify Spear phishing Emails</b>	<b>118</b>
7.1	Introduction . . . . .	118
7.2	Background and Related work . . . . .	122
7.3	Data collection methodology . . . . .	124
7.3.1	Email dataset . . . . .	124
7.3.2	Email Sample Dataset Description . . . . .	128
7.3.3	LinkedIn profile dataset . . . . .	131
7.4	Analysis and results . . . . .	134
7.4.1	Feature set description . . . . .	134
7.4.2	SPEAR versus SPAM emails from Symantec . . . . .	137
7.4.3	SPEAR emails versus BENIGN emails . . . . .	139
7.4.4	SPEAR versus a mixture of BENIGN and SPAM . . . . .	141
7.5	Discussion . . . . .	143
7.6	Conclusion . . . . .	144
<b>8</b>	<b>Conclusion</b>	<b>146</b>
8.1	Summary . . . . .	146
8.1.1	Characterizing poor quality context-specific content . . . . .	147
8.1.2	Effectiveness of automated techniques for identification of poor quality content	148
8.1.3	Deployment and evaluation of a real-world solution for automated real-time assessment of content on Facebook . . . . .	148
8.2	Limitations and Future work . . . . .	149

# List of Figures

1.1	Street artist in New York, USA puts up road signs telling mobile users to pay attention while walking. . . . .	3
1.2	High level schema covering the attack scenarios, nature of poor quality content, and techniques used to study this content. . . . .	6
2.1	A pictorial representation of the Facebook Graph. . . . .	8
2.2	Facebook terminology and building blocks. . . . .	9
2.3	Example of a real Facebook post containing a link which looks like a video from an external source. . . . .	13
2.4	A fake Facebook post which looks visually similar to a genuine Facebook video. . . .	14
2.5	High level design diagram of the immune system deployed by Facebook. . . . .	15
2.6	A snapshot of the privacy settings page on Facebook. Users can control the visibility level of all their information and content independently. . . . .	23
2.7	A snapshot of the Graph API explorer page offered by Facebook’s Developers Platform. Clicking on the “Get Token” button on the right generates an access token in the “Access token” field in the middle. . . . .	24
3.1	High level flow diagram capturing the methodology followed in this work. . . . .	31
3.2	One of the 7,296 malicious posts from our dataset which were not deleted by Facebook. . . . .	36
3.3	Warning page supposed to be shown by Facebook whenever a user clicks on a link reported as abusive on WOT. . . . .	37
3.4	Snapshot of the web interface presented to the annotators. . . . .	38
3.5	Sources of malicious content, legitimate content with URLs, and all legitimate content. . . . .	41
3.6	The <i>link</i> field present in the post object returned by the Facebook Graph API. . . .	43
3.7	Distribution of platforms used to post malicious and legitimate content in our dataset. . . . .	44

3.8	Accuracy values and ROC curve for Random Forest Classifier trained on Dataset I. . . . .	48
3.9	Tag clouds of top 75 most frequently occurring terms in Dataset I and Dataset III. . . . .	49
3.10	True positive rates of all models generated from Dataset I over time. . . . .	50
4.1	Architecture / flow diagram for Facebook Inspector. . . . .	58
4.2	Sample Facebook post marked as malicious by Facebook Inspector. . . . .	60
4.3	CDF of the response time (in seconds) of Facebook Inspector. . . . .	62
5.1	Fake British Air Facebook page offering free flights for a year in return for liking, commenting on, and sharing their post. . . . .	67
5.2	Distribution of the total number of posts published, and number of page <i>likes</i> gathered by pages in our dataset. . . . .	72
5.3	Linguistic analysis of content produced by politically polarized groups of pages in our dataset. . . . .	76
5.4	Number of malicious posts versus all domains published by all 627 <i>pages</i> in our dataset. . . . .	77
5.5	Types of content published by malicious and benign pages in our dataset. . . . .	80
5.6	Network graphs capturing intra- <i>page</i> activity of malicious <i>pages</i> in our dataset. . . . .	82
5.7	Daily, hourly, and weekly temporal activity of <i>pages</i> in our dataset. . . . .	83
5.8	Percentage change in <i>page likes</i> ( <i>gain factor</i> ) over one year for all <i>pages</i> in our dataset. . . . .	84
5.9	Distribution of the popularity gradients (in degrees) for malicious and benign pages in our dataset. . . . .	85
5.10	Top 20 attributes across malicious and benign pages that were changed at least once during one year of observation. . . . .	85
5.11	ROC area under curve values for Logistic Regression classifier corresponding to different sizes of post history. . . . .	88
5.12	ROC AUC values obtained by neural networks trained on a bag of words for different sizes of bag of words. . . . .	90
5.13	ROC AUC values obtained by neural networks trained on a bag of words for different sizes of post history. . . . .	91
5.14	ROC curve for Neural Networks trained on trigrams. . . . .	92
6.1	Example of a Facebook post where sentiment associated with the post text is in contrast with the sentiment associated with the text embedded in the image. . . . .	96

6.2	Architecture of our 3-tier pipeline used to extract human understandable descriptors from images. . . . .	101
6.3	Visual similarity between a bolo tie and the famous ‘Peace for Paris’ symbol. . . . .	102
6.4	Example of text embedded in an image posted on Facebook during the Paris Attacks in 2015. . . . .	103
6.5	Rumors spread on Facebook in the form of images during the Paris Attacks in 2015.	107
6.6	Distribution of positive, negative, and neutral text emotion in our dataset of images containing text, and posts. . . . .	109
6.7	Example of a post published during the Paris attacks, showing conflicting sentiments across image and text. . . . .	111
6.8	Sentiment values across post text, image text, and images over time. . . . .	112
6.9	Rumors spread on Facebook in the form of images, during the Paris Attacks in 2015. We used modern image analysis techniques to identify these rumors. . . . .	114
7.1	Attack scenario where attacker leverages context-specific information from one platform (social network) to attack victim with poor quality content (targeted spear-phishing) on another platform. . . . .	119
7.2	Example of a malicious PDF attachment sent via a spear phishing email. . . . .	120
7.3	Time line of the number of spear phishing and spam / phishing emails in our dataset.	127
7.4	Tag clouds of the 100 most frequently occurring words in the subjects and bodies of our SPEAR, SPAM, and BENIGN datasets. . . . .	132
7.5	Flow diagram describing the data collection process we used to collect LinkedIn data, and create our final feature vector containing stylometric features from emails, and social features from LinkedIn profiles. . . . .	133
7.6	Number of SPEAR and SPAM emails received by employees in the top 25 locations extracted from their LinkedIn profiles. . . . .	139
7.7	Number of LinkedIn connections of the recipients of SPEAR and SPAM emails. . . . .	139
8.1	High level schema covering the attack scenarios, nature of poor quality content, and techniques used to study this content. . . . .	147

# List of Tables

3.1	Event name, keywords used as search queries, number of posts, and description for the 17 events in our dataset of public Facebook posts. . . . .	32
3.2	Descriptive statistics of complete dataset collected over April 2013 - July 2014. . . .	34
3.3	Category labels and descriptions returned by the WOT API. . . . .	34
3.4	Top 10 most common posts in our dataset of malicious posts. . . . .	39
3.5	Campaigns found in Dataset II. Two of the three campaigns existed across more than one event. . . . .	42
3.6	Features used for machine learning experiments. We extracted features from four sources, viz. entity, content, metadata, and link. . . . .	45
3.7	Source and feature importance value (normalized) of the top 10 features. . . . .	47
3.8	Ten-fold cross validation accuracies for four classifiers over six different feature sets.	47
3.9	Crisis and non-crisis events in our dataset. We used the Oxford Dictionary definition of <i>crisis</i> to mark events as crisis or non-crisis. . . . .	52
3.10	Ten-fold accuracy scores averaged across ten experiments for each classifier. . . . .	53
3.11	Source and feature importance value (normalized) of the top 10 features in Dataset II.	54
4.1	Label and Confidence level corresponding to probability scores produced by Model I and Model II. . . . .	59
4.2	Summary statistics for the usage of Facebook Inspector. . . . .	61
4.3	Breakdown of time consumed by the main components of Facebook Inspector. . . . .	63
5.1	Category labels and descriptions returned by WOT API. Source: WOT API Wiki ( <a href="https://www.mywot.com/wiki/API">https://www.mywot.com/wiki/API</a> ). . . . .	71
5.2	Descriptive statistics of our dataset of Facebook <i>pages</i> . . . . .	72



5.3	Number of malicious posts and <i>pages</i> in each category in our dataset. . . . .	73
5.4	Word frequency of the top 30 terms appearing in page names in our dataset. . . . .	74
5.5	Top 10 malicious domains in our dataset with their Web of Trust classification, Facebook audience, and Alexa world rank. . . . .	78
5.6	Network analysis of <i>likes</i> , <i>comments</i> and <i>shares</i> networks within and between <i>pages</i> in our dataset. . . . .	81
5.7	Mean values for standard error of estimated gradient and correlation p-values for linear model. . . . .	84
5.8	Classification accuracy and ROC AUC values for automatically detecting malicious Facebook pages. . . . .	88
5.9	Page and post level features used for training supervised learning models. . . . .	89
5.10	Classification accuracy and ROC AUC values for automatically detecting malicious Facebook pages using bag-of-words. . . . .	92
6.1	Descriptive statistics of our dataset we collected from Facebook during the Paris Attacks in 2015. . . . .	99
6.2	List of labels generated by the Inception-v3 model, and the labels they are renamed with. . . . .	102
6.3	Top 20 most common image labels in our dataset. . . . .	105
6.4	Mutually exclusive set of 20 most frequently occurring relevant keywords in post and image text, with their normalized frequency. . . . .	108
6.5	Statistical summary of all rumors we identified. . . . .	115
7.1	Top 20 most frequently occurring attachment names, and their corresponding percentage share in our spear phishing and spam / phishing datasets. . . . .	125
7.2	Top 15 most frequently occurring attachment types, and their corresponding percentage share in our spear phishing and spam / phishing datasets. . . . .	126
7.3	Top 20 most frequently occurring subjects, and their corresponding percentage share in our spear phishing, and spam / phishing email datasets. . . . .	128
7.4	Detailed description of our dataset of LinkedIn profiles and emails across 15 organizations including Enron. . . . .	129
7.5	A spear phishing email from our SPEAR dataset. . . . .	130

7.6	Examples of <i>subject</i> and <i>attachment</i> names of two spam emails from our SPAM dataset. . . . .	131
7.7	List of features used in our analysis. . . . .	135
7.8	Accuracy and weighted false positive rates for SPEAR versus SPAM emails. . . . .	137
7.9	Information gain, mean and standard deviation of the 10 most informative features from SPEAR and SPAM emails. . . . .	138
7.10	Accuracy and weighed false positive rates for SPEAR emails versus BENIGN emails. . . . .	140
7.11	Information gain, mean and standard deviation of the 10 most informative features from SPEAR and BENIGN emails. . . . .	141
7.12	Accuracy and weighed false positive rates for SPEAR emails versus mix of SPAM emails and BENIGN emails. . . . .	142
7.13	Information gain, mean and standard deviation of the 10 most informative features from SPEAR and a combination of BENIGN and SPAM emails. . . . .	142

# Chapter 1

## Introduction

Online Social Networks (OSNs) are an integral part of the modern Internet. OSNs first came into existence with the advent of Web 2.0 in the early 2000's. The social aspect introduced by OSN services caught immediate attention made them immensely popular among Internet users all around the world in a very short span of time. Today, close to two billion users around the world use at least one OSN service.<sup>1</sup> Facebook (1.59 billion) and Twitter (320 million) lead the way in terms of the number of monthly active users for a single OSN.<sup>2 3</sup> Such widespread reach and popularity make OSNs a powerful tool for communication, especially during national and international events of interest, like sports, natural calamities, political events, etc. Users around the world use OSNs as primary sources to consume news, updates, and information about events around the world. A majority of Twitter and Facebook users, for example, say that each of these platforms serves as a source for news about events and issues outside the domain of family and friends [108].

### 1.1 Context-specific content

The enormous user base of OSNs gives birth to an unending stream of user-generated content traveling across the globe with lightning speed. Especially during news-making events like natural calamities, sports, and politics, the rate of generation of content on OSNs witnesses a sharp rise [148]. For example, the 2014 FIFA World Cup final inspired more than 618,000 tweets per minute, a new record for Twitter. Facebook also saw 350 million users generating over 3 billion posts, comments, and likes during the 32 days of the world cup.<sup>4</sup> The presidential elections that took place in the USA in 2016 witnessed a similar phenomenon. Twitter said that more than 75

---

<sup>1</sup><http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>

<sup>2</sup><https://about.twitter.com/company>

<sup>3</sup><http://newsroom.fb.com/company-info/>

<sup>4</sup><http://edition.cnn.com/2014/07/14/tech/social-media/world-cup-social-media/>

million tweets related to the election had been sent by 3 am Eastern Time on November 9, 2016, the moment Trump claimed victory. More than 115 million people discussed the election on Facebook, generating more than 716 million likes, posts, comments, and shares related to the vote, the social media platform said.<sup>5</sup>

Such content generated during news-making events is different from general user-generated content like updates, conversations, and personal experiences in the sense that it has the **context** of an event attached to it. Given the volume and variety of such context-specific content, it is difficult to moderate and verify all such information generated and shared. This lack of control and inability to monitor content enables hostile entities to exploit the context of popular news-making events and generate and promote various sorts of poor quality context-specific content, including (but not limited to) untrustworthy information, fake content, scams, rumors, and hoaxes. Such activity pollutes the information stream, making veracity of content a challenging task.

For our research, we refer to content generated during, and in the context of news-making events as *context-specific content*.

## 1.2 Social Media Service

We focus our work on Facebook, which is the largest online social network in the world with more than 1.79 billion monthly active users.<sup>6</sup> Over the past decade, Facebook’s popularity has seen such a monumental surge, that “checking” their Facebook accounts has become an addiction for Internet users. Figure 1.1 depicts a road sign in New York prompting users to avoid using Facebook on their mobile devices while walking on the street.<sup>7</sup> Researchers have even proposed a “Facebook addiction scale” to measure the level of users’ obsession with Facebook [6]. Apart from keeping in touch with friends and family, a big proportion of users also resorts to Facebook for getting their daily dose of news and updates about what is going on around the world. A recent survey of 5,173 adults suggested that 30% of people get their news from Facebook, while only 8% receive news from Twitter and 4% from Google Plus [77].

The mere volume of the public Facebook content (approx. 1.33 billion posts per day [48]) makes it a potentially rich source of information. In addition, introduction of features like hashtag support [99], Graph search for posts [131], and trending topics [115] have largely increased the level of visibility of public content on Facebook, either directly or indirectly. Users can now *search* for topics and hashtags to look for content, thus making the public Facebook content more visible and consumable by its users. This increasing public visibility, and an enormous user-base, potentially

---

<sup>5</sup><http://money.cnn.com/2016/11/09/technology/election-trump-social-media-records/>

<sup>6</sup><http://newsroom.fb.com/company-info/>, as recorded on September 30, 2016.

<sup>7</sup>Picture taken from a news article on DailyMail <http://www.dailymail.co.uk/news/article-2146282/Street-artist-puts-road-signs-telling-mobile-users-pay-attention-walking--notice.html>



Figure 1.1: Street artist in New York, USA puts up road signs telling mobile users to pay attention while walking.

makes Facebook one of the largest and most widespread sources of information on the Internet, especially during real-world events, when social media activity swells significantly.

**What makes Facebook vulnerable to poor quality content?** Facebook is a friendship-oriented OSN which helps users to connect to family, friends, and other known people on the Internet. After registering to use the site, users can create a profile, add other users as “friends”, exchange messages, post status updates and photos, share videos, use various apps and receive notifications when others update their profiles. All content posted by a user’s friends is visible in the user’s “newsfeed” area. Additionally, users may join common-interest user groups, organized by workplace, school or college, or other characteristics, and categorize their friends into lists such as “People From Work” or “Close Friends”. These characteristics are unlike some of the other famous OSNs like Twitter, Instagram, Pinterest, etc., where connections are unidirectional

(follower-following) as opposed to bidirectional (friendship), and users can “follow” other users without requiring their consent (unless the account is set to “private”). In contrast, connections between two users on Facebook need prior consent from both the users. This characteristic ensures that all people in a user’s network are known to the user (in most cases), introducing a certain degree of trust and believability in the content a user sees in her newsfeed. This makes Facebook users potentially more vulnerable to fall for scams, untrustworthy information, etc. when they come across such a piece of content, especially if it is context-specific. Recently, cybercriminals exploited the context of various news events to spread hoaxes and misinformation on Facebook, luring victims into scams, phishing attacks, malware infections, etc. [105,172]. It has been claimed that Facebook spammers make \$200 million just by posting links [149]. Such activity not only degrades user experience but also violates Facebook’s terms of service. Facebook has acknowledged misinformation, spam, and hoaxes as serious issues, and taken steps to reduce poor quality content on the platform [44,110,116,117,125].

### 1.3 Thesis statement

Given the above mentioned challenges, we address the issue of identifying poor quality context-specific content on the Facebook social media platform. The thesis statement is as follows:

**Poor quality context-specific content generated on Social Media Services during news-making events can be identified within seconds, and with over 80% accuracy via automated techniques using only publicly available information.**

### 1.4 Thesis contribution

This thesis is a timely contribution to the area of computer science, needed to mitigate the spread of poor quality content on OSNs in the present day scenario. Results from this thesis can be used to build real-world solutions to mitigate the spread of untrustworthy information, fake content, scams, etc. on different OSNs in the future. The insights obtained and system built as part of this thesis are effectively used by hundreds of Facebook users every day to make informed and intelligent decisions about the content they view on the OSN. We present how automated computational techniques can be used to deploy a real-world system for ordinary users to differentiate poor quality content from benign content.

### 1.4.1 Theoretical

- Analyzed and compared the characteristics of poor quality and benign content (including posts, users, and pages) on Facebook, with the purpose of developing automated techniques to differentiate poor quality content from benign content during news-making events.
  - We found multiple characteristics at the content, metadata, and temporal level which segregated poor quality content from benign content. For instance, entities posting poor quality content were temporally more active.
- Evaluated the effectiveness of automated computational methods to detect poor quality content on Facebook using a combination of content, meta information, and temporal features.
  - Ensemble methods (Random Forest) turned out to be the most effective for differentiating poor quality content from benign content at the metadata level; artificial neural networks performed the best at the content level.
- Developed and deployed a novel framework for providing indication for potentially poor quality content posted during news-making events. Evaluated the framework for performance, response time, and usability.
  - This framework (Facebook Inspector) is able to distinguish between poor quality and benign content without relying on popular Facebook metrics (*likes*, *comments*, and *shares*) with over 80% accuracy in under 3 seconds on average.

### 1.4.2 Practical

- The real-time system Facebook Inspector (FbI) built as part of this thesis to identify poor quality posts on Facebook is used by over 250 real Facebook users every day. It has been downloaded over 5,000 times, and has analyzed over 2 million public Facebook posts.
- A unique dataset of thousands of poor quality Facebook posts, users, and pages across 20+ news making events collected over 4 years. This is one of the largest datasets of public Facebook data in literature, and has been made available for future research.

Figure 1.2 presents a high level schema of the topics covered in this thesis. Specifically, the diagram describes the nature of poor quality content studied under the attack scenarios covered in this thesis, and the techniques used to characterize, study, and analyze this content.

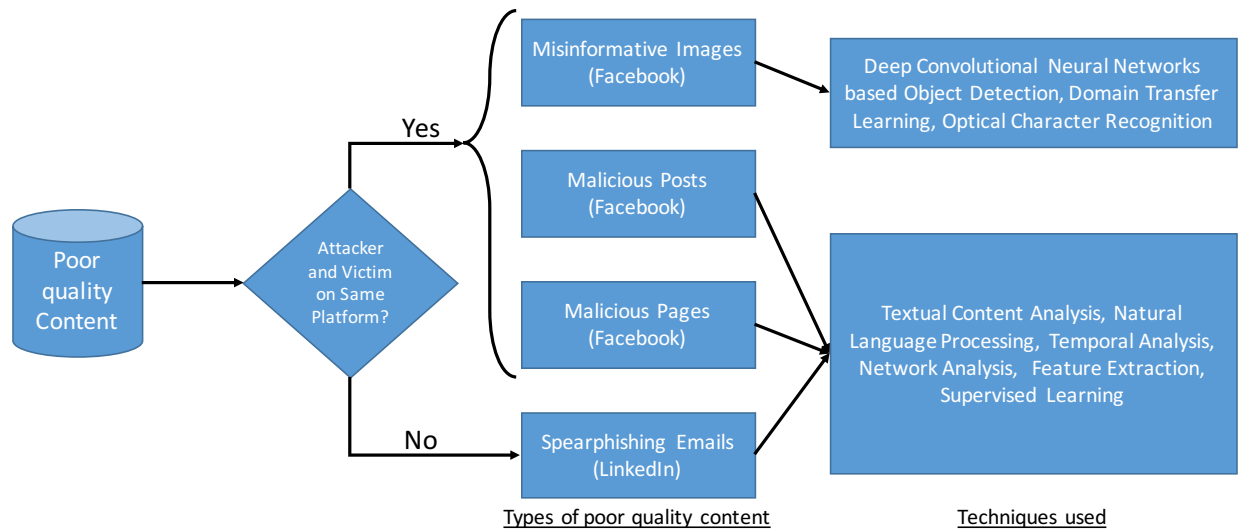


Figure 1.2: High level schema covering the attack scenarios, nature of poor quality content, and techniques used to study this content.

## 1.5 Thesis roadmap

The rest of the thesis is organized as follows. Chapter 2 discusses the literature review in the space of exploring the Facebook network, existing research in the area of poor quality content identification on Facebook and other social network platforms, and studies related to context-specific (real-world event related) content analysis on Facebook and other social network platforms. Chapter 3 describes our work on identification, characterization, analysis and automatic real-time detection of malicious posts on Facebook. Chapter 4 discusses the implementation, public deployment and evaluation of Facebook Inspector. Chapter 5 contains our work describing the anatomy of malicious pages on Facebook, and techniques for automatic identification of such pages. Chapter 6 describes our work on deciphering crisis events through images, and misinformation present on the Facebook network in the form of images. Chapter 7 extends our attack scenario beyond Facebook, and explores the possibility of exploiting context-specific information about victims to better identify poor quality content in the form of targeted spear phishing emails. We conclude our work and discuss the limitations, implications and future directions in Chapter 8.



## Chapter 2

# Background and Literature Review

In this chapter, we look at the basic building blocks of the Facebook social network and review existing work done in the space of identification and analysis of poor quality content on Facebook. We also review literature in the space of analysis of context-specific content in the form of news-making events on online social networks. The aim of this chapter is to look at a range of research attempts which would help to explore the various types of poor quality content spread on Facebook. Then, we look at the various limitations that Facebook poses, which makes event analysis, and context-specific content quality assessment on this network a hard problem. Towards the end, we discuss the implications and research gaps in identifying and analyzing poor quality user-generated content on Facebook during news-making events.

### 2.1 Facebook: Terminology, building blocks, and characterization

Facebook is a bidirectional network, where two users cannot connect without mutual consent. Each “object” on this network such as a user, picture, post, video, album etc. is modeled as a node in a graph, and each “action”, for example, a *like*, *comment*, *share*, *friendship connection* etc. is an edge which connects two nodes. Figure 2.1 depicts a pictorial representation of this graph. This graph structure followed by Facebook is the primary reason behind the name of Facebook’s API, the Graph API.

#### 2.1.1 Terminology and Building Blocks

Every user who registers on Facebook gets their own *user profile*. According to Facebook’s terms of service, an individual can have only one user profile in her name. A user profile is a collection of photos, stories, and experiences that tell the story of that user. It includes the user’s Timeline,

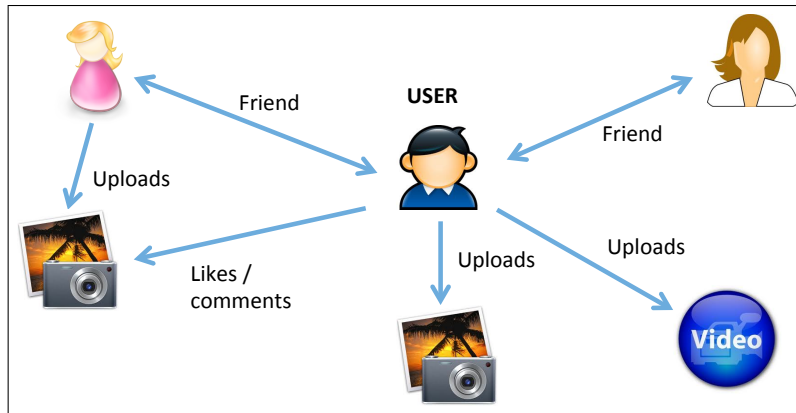


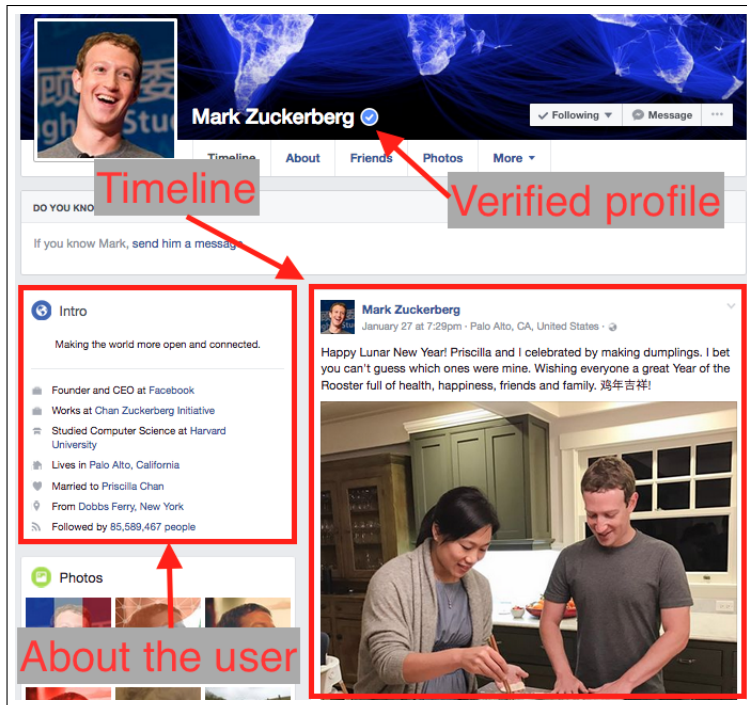
Figure 2.1: A pictorial representation of the Facebook Graph. Users, photos, videos etc. are nodes, and likes, comments, shares, etc. represent edges.

profile picture, biography, and personal information. The *Timeline* is where users can see their posts or posts they have been tagged in, displayed by date. Each Facebook user sees their *News Feed* by default when they log into Facebook. The News Feed is a constantly updating list of stories in the middle of the default Facebook homepage. It includes status updates, photos, videos, links, App activities, and Likes from the people, Pages, and Groups that the user is associated with. Facebook also allows users to send instant messages to their friends using the *chat* feature. The News Feed also shows a list of popular topics that are *Trending*. Trending is a list of topics and hashtags that have recently spiked in popularity on Facebook. Until January 2017, this was a personalized list based on the user's location, Pages they had liked, and what was trending across Facebook.

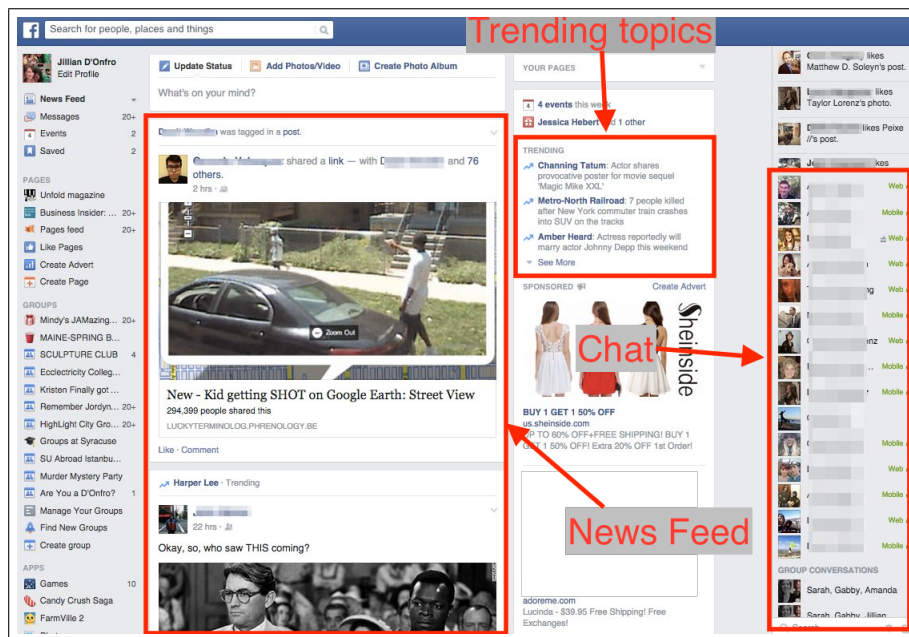
User profiles can be public or private but are only for non-commercial use. Facebook provides *pages* for commercial use. Facebook Pages help businesses, organizations, and brands share their stories and connect with people. Like profiles, Pages can be customized by posting stories, hosting events, adding apps, and more. People who like a Page can get updates in their News Feeds. Pages are controlled by *Page Admins*. When a user creates a Page, they automatically become the Page's admin, which means only they themselves can change how the Page looks and post as the Page. Users can then assign roles to other people to help them manage their Page.

Some Pages and profiles are *verified* by Facebook to let people know that they are authentic. These can include celebrities and public figures, global brands and businesses, and media. Once verified, a blue badge appears next to the Page's or profile's name, and is visible to everyone.<sup>1</sup> Figure 2.2 shows an example of a Facebook user profile and News Feed and the various building blocks as discussed.

<sup>1</sup><http://sproutsocial.com/insights/facebook-terminology-glossary/>



(a) Example of a Facebook User Profile



(b) Example of Facebook News Feed

Figure 2.2: Facebook terminology and building blocks.

### 2.1.2 Characterization of the Facebook network

Facebook was made publicly available in September 2006.<sup>2</sup> Within a couple of years, the rapid growth rate of Facebook’s user base made it a center of attraction for research around the world. Researchers started off by collecting and trying to understand subsets of the Facebook network. Lewis et al. [98] provided a first of its kind dataset of Facebook users, and made it publicly available. Authors downloaded and characterized the profile and network data of 1,640 freshmen students enrolled at a diverse private college in the Northeast U.S. in 2009 by requesting permission from Facebook, and the university in question. Their findings revealed some interesting characteristics about these students, for example, the average pair of students whether or not they shared ties or demographics were found to have a higher percentage of favorite books / authors in common (2.%) than movies (1.5%) or music (1.5%). Similarly, they observed the highest similarity among friends who both appeared in each other’s photo albums. Closely related to this work was Amanda et al.’s work [153], where authors studied the structure of social networks of students by examining the graphs of Facebook “friendships” at five American universities at a single point in time. The primary aim of this paper was to use an unsupervised algorithm to compute the community structure consisting of clusters of nodes of these universities and to determine how well the demographic labels included in the data correspond to algorithmically computed clusters. Following up, authors extended their study to one hundred American colleges and universities, and examined homophily and community structure for each of the networks, and compared the community structure to partitions based on the given categorical data [154]. Both these studies also obtained their datasets from Facebook directly.

All the aforementioned studies showed promising results, derived by using sound methodologies, but suffered from intrinsic sampling bias. The sample datasets obtained for all these studies were fairly small and represented a confined set of Facebook users, which were American college students. Although these datasets were obtained directly or indirectly from Facebook itself and were complete in most aspects, the results obtained by these studies cannot be extended and generalized to the common Facebook audience. To overcome this sampling bias, researchers either needed to analyze the entire Facebook network at once, or to employ better techniques for sampling, to obtain more representable subsets of the Facebook graph. We discuss the former in the next paragraph, and the latter in Section 2.1.3.

By the year 2011, over 500 million users around the world had become a part of the Facebook network.<sup>3</sup> This was the time when Ugander et al. [157] first studied the complete Facebook graph. Authors of this work brought out some major insights about Facebook’s network by conducting a large scale analysis on the entire Facebook network. They confirmed the ‘six degrees of separa-

---

<sup>2</sup><http://blog.facebook.com/blog.php?post=2210227130>

<sup>3</sup><http://www.digitalbuzzblog.com/facebook-statistics-stats-facts-2011/>

tion’ phenomenon on a global scale, and found that the social network is nearly fully connected, with 99.91% of individuals belonging to a single large connected component. In addition, authors observed a strong effect of age on friendship preferences as well as a globally modular community structure driven by nationality but did not find any strong gender homophily. In a follow-up study, Backstrom et al. [8] found the average distance between two Facebook users to be 4.74, corresponding to 3.74 intermediaries or “degrees of separation” instead of six. Prior to this work, most of the research involving Facebook studied it from either from a privacy standpoint [3, 14, 74, 100], or a social science perspective, looking at why users use Facebook [45, 86, 123, 132], and what are the mental, social, and emotional affects of Facebook usage on individuals [60, 89].

### 2.1.3 Facebook crawling

As discussed in Section 2.1.2, results obtained from studies conducted on the Facebook network subsets were non generalizable, and needed better sampling techniques. Researchers in some more studies crawled the Facebook network to study user interactions, instead of fetching selective data from Facebook [161, 166]. In 2008, Wilson et al. [166] exploited Facebook’s partitioning of the user population into networks to perform a complete crawl of its subsets iteratively. Their primary data set comprised of the profile, Wall and photo data crawled from the 22 largest regional networks on Facebook between March and May of 2008. In all, this dataset was composed of full profiles of over 10 million Facebook users. Vishwanath et al. [161] used a similar technique to crawl a partial subset of the New Orleans network. Authors were able to gather information of about 90,269 users and 3,646,662 friendship links between those users. This accounted for 52% of the users in the New Orleans network based on the statistics provided by Facebook at that time.

With large-scale crawls of the Facebook network, and bigger datasets, researchers were able to reduce sampling biases to some extent. However, there was still scope for better sampling techniques to enhance representativeness and generalizability. To this end, Gjoka et al. [59] implemented several crawling techniques to obtain a representative and unbiased sample of the Facebook network. Authors found the Metropolis-Hasting random walk and a re-weighted random walk to work well, whereas the traditional Breadth-First-Search and Random Walk were found to perform quite poorly, producing substantially biased results. The collected samples were validated against a true uniform sample obtained using Facebook user IDs, as well as via formal convergence diagnostics, and were shown to have good statistical properties. The technique used by the authors for collecting a true uniform sample was then utilized by Catanese et al. [21], who performed a comparative analysis of two large crawls of Facebook; one using the Breadth-First-Search technique, and the other using the true uniform sampling technique mentioned previously. Authors of this work highlighted some distinct differences between the two datasets, including degree distribution, clustering coefficient, Eigenvector centrality etc.

In this section, we looked at the building blocks of the Facebook network, and existing work done on characterizing the Facebook graph and its subsets. We observed that quite a few pieces of early experiments conducted using Facebook’s data suffer from sampling bias, and the datasets used are not representative of the entire Facebook population. To counter this sampling bias, we then looked at literature which made use of large-scale crawls of the Facebook network, and found how some crawling techniques worked better than some others.

## 2.2 Poor quality content on Facebook

The popularity and reach of Facebook have also attracted a lot of spam, phishing, malware, and other types of malicious activity. Attackers lure victims into clicking on malicious links pointing to external sources and infiltrate their network. These links can be spread either through personal messages (chats) or through wall posts. To achieve maximum visibility, attackers prefer to post links publicly. Typically, an attacker initiates the attack by posting memes with attention grabbing previews, which prompt users to *like*, *share*, or *comment* on them in order to view them. The actions of *liking*, *commenting* or *sharing* spread these memes into the victim’s network. Once the meme is spread, the victim is redirected to a malicious website, which can further infect her computer, or friends network through phishing, malware, or spyware. Figure 2.3 shows an example of such a malicious post, which appears to be a video. Clicking on the link redirects the user to a phishing page as shown in Figure 2.4, which looks very similar to a genuine Facebook post. This phishing page asks the victim to share this video with their friends in order to view it. However, once the victim shares this video, the page redirects to a random advertisement page. The video corresponding to the preview / thumbnail shown in the post does not actually exist.

Multiple other sources have cited such examples of scams and malicious posts on Facebook in the past few years.<sup>4,5</sup> In addition to phishing scams, other malicious activity on Facebook includes unsolicited mass mentions, photo tagging, post tagging, private / chat messages etc. Intuitively, a user is more likely to respond to a message or post from a Facebook friend than from a stranger, thus making this *social* spam a more effective distribution mechanism than traditional email. This increased susceptibility to such kind of spam has prompted researchers to study, and combat social spam and other malicious activity on Facebook. We now look at the various attack and detection techniques that have been used in the past to identify and spread malicious content on Facebook respectively.

---

<sup>4</sup><http://online.wsj.com/news/articles/SB10001424052970203686204577112942734977800>

<sup>5</sup>[http://allfacebook.com/facebook-warning-amazon\\_b74943](http://allfacebook.com/facebook-warning-amazon_b74943)



Figure 2.3: An example of a real Facebook post from 2014, containing a link which looks like a video from an external source. Clicking on the link asks users to share the link before they can watch the video, which does not actually exist.

### 2.2.1 Types of poor quality content on Facebook

To identify and contain poor quality content on Facebook (or any social media service), it is essential to explore and understand the techniques that are or can potentially be deployed by attackers to spread such content. Patsakis et al. [122] described how Facebook could be exploited and converted into an attack platform, in order to gain some sensitive data, which can complete a perfect attacking profile against a user. Authors created a Facebook application for demonstration purposes that on the surface was a simple application, but on the background, it collected useful data. This app executed malicious code on the victim's browser and collected the IP address of the user-victim, the browser version, the OS platform and whether some specific ports are open or closed. This data was then transmitted to the authors over email. Authors also pointed out that their app was indexed on the main list of Facebook applications, despite the fact that the description of app clearly stated that it was generating malicious traffic, and had been created for penetration testing purposes. Huber et al. presented a *friend-in-the-middle* attack through hijacking session cookies. Authors explained how it was possible to impersonate the victim using this technique and interact with the network without proper authorization. However, this technique was proposed in



Figure 2.4: A fake Facebook post which looks visually similar to a genuine Facebook video. The URL in the address bar depicts that the page is a fake.

2011, when using HTTPS to connect to the website was optional.<sup>6</sup> Post 2013, all communication on Facebook uses encryption (HTTPS) by default,<sup>7</sup> which means that such attacks are no more possible.

Fan et al. [53] proposed a virus model based on the application network of Facebook. Authors also modeled the virus propagation with an email virus model and compared the behaviors of virus spreading in Facebook and email network. Their findings revealed that while Facebook provides a platform for application developers, it also provides the same chance for virus spreading. In fact, the virus was found to spread faster on the Facebook network if users spend more time on it. The result of their simulation showed that, even though a malicious Facebook application attracts only a few users in the beginning, it can still spread rapidly. That is because users may trust their friends of Facebook and install the malicious application.

It is important to understand that in addition to the techniques described above, a large proportion

<sup>6</sup><https://www.facebook.com/notes/facebook/a-continued-commitment-to-security/486790652130>

<sup>7</sup><https://www.facebook.com/notes/facebook-engineering/secure-browsing-by-default/10151590414803920>





Interestingly, despite this complex immune system deployed by Facebook, unwanted spam, phishing, and other malicious content continues to exist and thrive on Facebook. Although the immune system deployed by Facebook utilizes a variety of techniques to safeguard its users, authors did not present an evaluation of the system in terms of accuracy and efficiency in detecting anomalies.

Gao et al. [58], in 2010, presented an initial study to quantify and characterize spam campaigns launched using accounts on Facebook. They studied a large anonymized dataset of 187 million asynchronous “wall” messages between Facebook users and used a set of automated techniques to detect and characterize coordinated spam campaigns. Authors detected roughly 200,000 malicious wall posts with embedded URLs, originating from more than 57,000 user accounts. They also found that more than 70% of all malicious wall posts advertised phishing sites. Further, their findings revealed that more than 97% of the accounts they analyzed, were compromised accounts, rather than “fake” accounts created solely for the purpose of spamming. To the best of our knowledge, this is the only study which addresses the problem of detecting and analyzing malicious content on Facebook via automated means.

Following up their work, Gao et al. [57] presented an online spam filtering system that could be deployed as a component of the OSN platform to inspect messages generated by users in real-time. Their approach focused on reconstructing spam messages into campaigns for classification rather than examining each post individually. They were able to achieve a true positive rate of slightly over 80% using this technique and achieved an average throughput of 1,580 messages/sec with an average processing latency of 21.5ms on their Facebook dataset of 187 million wall posts. Although the technique of campaign identification has previously been used for offline spam detection, authors claimed to be able to achieve real-time detection using this technique with sufficiently low overhead. Their model stayed accurate for over 9 months after initial training, overcoming the need for frequent re-training.

Stringhini et al. [145] utilized a honeypot model to collect information about spammers on Facebook. Authors first crawled a set of 2,000 random profiles each, across 16 different regional networks to collect a representative sample which could help them create a representative honey profile for the network. They monitored this profile over a duration of one year, and manually identified 173 spam profiles among a total of 3,831 friendship requests they received. The 173 spam profiles were then crawled to extract features like URL ratio (number of URLs posted per message), message similarity, number of friends, number of messages sent, friend choice (ratio of total number of friend names to the total number of distinct friend names). These features were fed to a classifier along with features extracted from 1,000 legitimate profiles. Authors reported a low false positive rate of 2% and a low false negative rate of 1% but did not report the accuracy measure. Using this model trained on 173 spam, and 1,000 legitimate profiles, authors tested 790,951 profiles from Los Angeles and New York networks and identified 130 more spam profiles. Seven out of these 130 profiles detected, however, were marked as false positive upon manual inspection.

Ahmed et al. [5] presented a Markov Clustering (MCL) based approach for the detection of spam profiles on Facebook. Authors crawled the public content posted by 320 hand picked Facebook users, out of which, 165 were manually identified as spammers, and 155 as legitimate. Authors then extracted three features from these profiles, viz. Active friends, Page Likes, and URLs to generate a weighted graph, which served as input to the Markov Clustering model. This work, however, was also targeted at detecting spam campaigns instead of individual posts, similar to Gao et al.'s work [58]. Although the suggested approach produced reasonable results, there was no evaluation of the model on a bigger dataset.

In an attempt to protect Facebook users from malicious posts, Faloutsos [52] designed an efficient social malware detection method which took advantage of the social context of posts. The author was able to achieve a maximum true positive accuracy rate of 97%, with the algorithm requiring, on average, 46 milliseconds to classify a post. This algorithm was then used to develop MyPageKeeper<sup>8</sup>, a Facebook app to protect users from malware and malicious posts. Using data from this app, the author analyzed over 40 million posts during a period of 4 months and found that 49% of the users were exposed to at least one social malware post during this period. This work also showed how social malware significantly differed from traditional email spam or web-based malware. According to Faloutsos, website blacklists were able to identify only 3% of the posts flagged by MyPageKeeper, while 26% of flagged posts pointed to malicious apps and pages hosted on Facebook, which no antivirus or blacklist was designed to detect. In addition to malicious or compromised user accounts, there also exist multiple third party applications which enable / aid the spread of malicious posts on Facebook. To determine if a Facebook application is malicious, Rahman et al. developed FRAppE (Facebook Rigorous Application Evaluator), which was one of the first attempts focused on detecting malicious apps on Facebook [129]. FRAppE worked on a feature set that the author extracted from observing the posting behavior of approximately 111K Facebook apps across 2.2 million Facebook users. The authors found that 13% of the apps in their dataset were malicious, and were able to achieve an accuracy of 99.5% for detecting malicious apps using FRAppE. This was one of the first attempts at identifying malicious apps on Facebook via automated means.

Jin et al. [84] in 2011, proposed a scalable online social media spam detection system by utilizing a combination of image content features, text features, and social network features. Their feature sets were independent of the social media platform, but authors chose to test the system on Facebook because of its popularity. Although the feature sets proposed in this work looked more exhaustive than the techniques proposed previously, authors did not present their exact feature set or any evaluation results of their system.

---

<sup>8</sup><https://apps.facebook.com/mypagekeeper/>

**Summary** In this section, we looked at multiple research and system level contributions which identify the techniques for spreading poor quality content, and address the problem of detection of malicious content on Facebook. However, most of the work discussed above does not comprehensively address the issues in detail due to multiple reasons. One of the major reasons, as also addressed by Stringhini et al. [145], is the lack of availability of a substantial amount of data to analyze from Facebook. Large scale studies including studies conducted by Gao et al. [57,58], and Stringhini et al. [145] utilized the open nature, and geographically divided crawl-able networks of Facebook prior to October 2009. However, post 2009, the introduction of stringent privacy controls and unification of all geographical networks into one big graph have made it a challenging task to collect a sizable amount of data for analysis on Facebook. We discuss the other challenges in more detail in Section 2.4.

Most system level contributions towards detection lack vital details including detailed feature set description, the source of true positive labeled datasets, details of the algorithms used, comparison of existing spam classification techniques with proposed techniques, or lack of evaluation etc., questioning the reproduce-ability of their results. Regarding research level contributions, there has been some work done in this space, but with little follow-up studies to look at the evolution of malicious content over time. There does exist some work proposing a variety of attack techniques to propagate malicious content in the network, but the characterization and detection of this content is still an open issue.

There exist multiple other pieces of closely related work, which look at detecting fake profiles / sybil nodes on Facebook, and malicious activity on other social networking platforms like Twitter. We do not cover those pieces of work, since the focus of this survey is confined to *poor quality user-generated content* on Facebook, in particular.

## 2.3 Context-specific content analysis on other social networks

Online social network platforms like Facebook, Twitter, Google+, YouTube, Flickr, Instagram, Pinterest, etc. have provided people with a free and open platform to communicate with each other. Today, anything that happens in the real world is talked about on online social media services. From sports to storms, terrorist attacks, bomb blasts, earthquakes, and even elections, users share thoughts and information about literally everything using online social media services. A recent study revealed that social media activity increases up to 200 times during major events like elections, sports, or natural calamities [148]. This swollen context-specific activity has drawn significant attention from the computer science research community. Context-specific content and activity on Twitter, in particular, has been widely studied by researchers during events [10, 78, 95, 135, 165]. However, few studies have looked at social media platforms other than Twitter to study context-

specific content [24, 75, 114]. In this section, we look at the various attempts at analyzing context-specific content in the form of news-making events on online social network platforms. Since there has been a substantial amount of work done in this space on Twitter alone, we look at event analysis on Twitter, and event analysis on other social media services, separately.

At the broadest level, analyzing an event on online social media can be broken down into two components. The first component deals with identifying that an event has occurred, and the second component deals with collecting information specific to the event from a social media service. In an ideal scenario, the output of the first component serves as the input to the second component. Detecting that an event has occurred using a stream of user-generated content from online social media is, however, a grand challenge in itself. Thus, in practice, most of the work related to event analysis on online social media deals with the two components separately. For our survey, we do not look at the work which focuses on detection of events, we only look at work done on event-specific online social media data, where the event is already known to have occurred.

### **2.3.1 Event analysis on Twitter**

Twitter has been used widely during emergency situations, such as wildfires [32], hurricanes [79], floods [159] and earthquakes [43, 90, 135]. Journalists have hailed the immediacy of the service which allowed “to report breaking news quickly - in many cases, more rapidly than most mainstream media outlets” [126]. Sakaki et al. [135] explored the potential of the real-time nature of Twitter and proposed an algorithm to detect the occurrence of earthquakes by simply monitoring a stream of tweets in real-time. Here, the authors took advantage of the fact that users tweet about events like earthquakes as soon as they take place in the real world, and were able to detect 96% of all the earthquakes larger than a certain intensity. Their reporting mechanism was able to convey this information to users through emails, 6 minutes before the Japanese Meteorological Agency made an official announcement. In another research work, Sakaki et al. [136] analyzed tweet trends to extract events that happened during a crisis event from Twitter. They analyzed the log of user activity from Japanese tweets on all earthquakes during 2010-2011. Cheong et al. [25] performed social network analysis on Twitter data during Australian floods of 2011 to identify active players and their effectiveness in disseminating critical information. More similar work includes an attempt by researchers to identify information from Twitter, that may contribute to enhancing situational awareness during natural hazard real-world events. Authors of this work focussed on communications broadcast on Twitter by people who were “on the ground” during the Oklahoma Grassfires, and Red River Floods in the USA in 2009. They proposed an enhanced set of generic features, which could be used for building systems to improve situational awareness automatically during emergency events [159]. Hughes et al. [79] also studied four high profile, mass convergence events on Twitter and discovered that Twitter messages sent during such events reveal features of

information dissemination that support information broadcasting and brokerage.

Varol et al. [158] analyzed the Gezi Park movement in Turkey through the lens of Twitter. Authors analyzed 2.3 million tweets about the event produced over a period of 25 days, during May - June 2013. Authors of this work identified four types of users, viz. common users, rebroadcasters, influentials and hidden influentials, who tweeted during the event. Their analysis revealed that the conversation becomes more democratic as events unfold, with a redistribution of influence over time in the user population. Gupta et al. [67] studied the public Twitter stream during the Boston Marathon bombings in 2013. Their results revealed that 29% of the most viral content, during the crisis were rumors and fake content; while 51% was generic opinions and comments; and rest was true information. Besides, authors used regression prediction model, to verify that the overall impact of all users who propagate the fake content at a given time, can be used to estimate the growth of that content in future. Many malicious accounts were also created on Twitter during the Boston event, which were later suspended by Twitter. Authors were able to identify over six thousand such user profiles and observed that the creation of such profiles surged considerably right after the blasts occurred. In another research work, Gupta et al. [68] analyzed the public Twitter stream during the hurricane Sandy, which hit the USA in 2012. Here, authors performed a characterization analysis, to understand the temporal, social reputation and influence patterns for the spread of fake images. Their results showed that top thirty users (0.3%) out of 10,215 users resulted in 90% of the retweets of fake images, and network links such as follower relationships of Twitter contributed very less (only 11%) to the spread of these fake photos URLs. Classification models were used to distinguish fake images from real images spreading during Hurricane Sandy, and authors found the Decision Tree classifier to perform the best, with 97% accuracy in distinguishing fake images from real. Mendoza et al. [107] used the data from 2010 earthquake in Chile to explore the behavior of Twitter users for emergency response activity. The results showed that propagation of rumor tweets versus true news was different and automated classification techniques can be used to identify rumors. Longueville et al. [32] analyzed Twitter feeds during forest Marseille fire event in France; their results showed that in location-based social networks, spatiotemporal data could be analyzed to provide useful localized information about the event.

The work discussed above highlights the importance of Twitter as an information sharing medium during events. Tweets have been shown to travel faster than the tremors of an earthquake, making it a potential warning system capable of saving thousands of human lives in the event of a major earthquake [135]. During the Boston Marathon blasts, Twitter came to rescue and helped the Boston Police to track down the two main suspects within hours of the blasts.<sup>9</sup> The popularity, reach, and public nature of Twitter have made it the first choice for almost everyone, including law and order agencies, journalists, and evidently, computer science researchers. However, this intense

---

<sup>9</sup><https://blog.twitter.com/2013/the-boston-bombing-how-journalists-used-twitter-to-tell-the-story>

focus on Twitter has left a gap for a similar analysis on other social media services, especially Facebook. Content posted on Facebook and other famous social media services during events has fairly been overlooked. We now look at some of the research work which analyzes events on social media services other than Twitter.

### 2.3.2 Event analysis on Facebook, Google+, and other online social networks

As discussed in Section 2.3.1, there exists little work which focuses on online social media services other than Twitter to analyze events. Osborne et al. [113] examined how Facebook, Google Plus, and Twitter report in breaking news. Authors identified 28 major events which took place in December 2013, and scanned all the three OSM services, viz. Facebook, Twitter, and Google Plus for posts related to these events. Their findings revealed that all media carried the same major events, but Twitter continued to be the preferred medium for breaking news, almost consistently leading Facebook or Google Plus. Facebook and Google Plus largely reposted newswire stories and their main research value was that they conveniently packaged multiple sources of information together. This was one of the first attempts towards studying public streams of Facebook, and Google Plus and comparing them with Twitter. Szell et al. investigated messages from five sources, viz. Facebook, Twitter, app.net, Enron email corpus, and a popular online forum, created in response to major, collectively followed events such as sports tournaments, presidential elections, or a large snow storm [148]. They related content length and message rate, and found a systematic correlation during events which can be described by a power law relation - the higher the excitation, the shorter the messages. Authors showed that on the one hand this effect could be observed in the behavior of most regular users, and on the other hand was accentuated by the engagement of additional user demographics who only post during phases of high collective activity. Palen et al. [119] studied two real-world events, to understand and characterize the wide scale interaction on social networking websites with respect to the events. Authors in this work presented a qualitative analysis of a group on a popular social networking site as a virtual destination in the aftermath of the Northern Illinois University (NIU) shootings of February 14, 2008, in relation to the related activity that happened in response to the Virginia Tech (VT) tragedy 10 months earlier.

In this section, we looked at the various research on analyzing events on Twitter, and other online social media services. Past research depicts the dominance of Twitter when it comes to analyzing events. However, we also saw evidence of similarity in the information streams between Twitter and other social media services like Facebook and Google Plus during events [113]. The dominance of work on Twitter has left a wide gap and prompts researchers to look at public streams on other social media events during events.

## 2.4 Facebook Limitations and Challenges

It is important to note that there hardly exists any work which focuses on Facebook content for analyzing events. Given that Facebook is much older and bigger than Twitter, the reasons for the lack of work on Facebook need good justification and discussion. In this section, we look at the various limitations and challenges posed by Facebook, which possibly makes it a difficult task to extract, and analyze data from this network. Intuitively, the private nature of Facebook can largely be attributed to the lack of research on Facebook content on a large scale. We now look at these, and other challenges with Facebook research in detail.

### 2.4.1 Fine grained privacy settings

Facebook provides its users with an exhaustive set of privacy settings, which enable them to control who can see what information from their profile and posts. Unlike Twitter, the majority of content on Facebook, including profile information, content, and network information, is not accessible publicly. Privacy settings at Facebook broadly offer visibility of information at four *levels*, as follows:

- Only me: Only the authorized user can see this information.
- Friends: Users who are connected to the authorized user via a “friendship” relation can see this information.
- Friends of friends: Visibility at this level is increased to one more hop in the network. i.e. users who are friends with the friends of the authorized user, can also see this information.
- Public: Anyone on the Internet can see this information.

In addition to these four basic visibility levels, Facebook also provides a custom level of visibility, where a user can choose the audience of her information and content selectively, at an individual level. These visibility levels can be applied to almost all sections of a user’s Facebook account, like profile information, pictures, albums, videos, wall posts, etc. Figure 2.6 presents a snapshot of the privacy settings page on Facebook. Users can also select who can send them friendship requests, and who can search them.

Although these privacy settings offer users tremendous control over their content, they pose serious implications from a research standpoint. No more than 28% of Facebook users share their content with an audience wider than their friends [103]. This leaves researchers with no more than a quarter of the total content on Facebook (probably even lesser), to collect and analyze. Although with a user base of over a billion users, this proportion of content may convert to millions of posts a day, it



Privacy Settings and Tools			
Who can see my stuff?	Who can see your future posts?	Public	Edit
	Review all your posts and things you're tagged in		Use Activity Log
	Limit the audience for posts you've shared with friends of friends or Public?		Limit Past Posts
Who can contact me?	Who can send you friend requests?	Everyone	Edit
	Whose messages do I want filtered into my inbox?	Basic Filtering	Edit
Who can look me up?	Who can look you up using the email address you provided?	Everyone	Edit
	Who can look you up using the phone number you provided?	Everyone	Edit
	Do you want other search engines to link to your timeline?	Yes	Edit

Figure 2.6: A snapshot of the privacy settings page on Facebook. Users can control the visibility level of all their information and content independently.

is hard to see this content as a good representation of the entire Facebook population. Especially, when it comes to identifying and analyzing malicious content, it is hard to find an effective and scalable solution using only a fraction of the content. Attackers can effectively exploit the private nature of the network to target multiple closed networks simultaneously, with no way for outsiders to identify these threats and provide any warnings or protection. The majority of attack vectors may thus, never surface, and go untraceable.

Privacy rules applied on profile and network information of users, make it even harder to analyze the sources of malicious content in Facebook. Apart from gender, name, and username, all other profile information about a user is not available publicly, unless explicitly specified by the user. Vital pieces of information like a user's work, education, description, location, account creation time, birth date, etc. are virtually impossible to extract from Facebook. This implies that even if a user is identified as malicious, it is hard to analyze and extract features, which can be used to differentiate a benign user from a malicious one. Lack of network information poses similar implications in the analysis of infected subgraphs of Facebook. Networks in Facebook are bidirectional. i.e. two users cannot connect to each other without mutual consent from both. This connection between two users is known as a "friendship" connection. Friends of a user are also non-public by default, which eliminates the possibilities of analyzing cascades, communities, infected subgraphs, and the flow of malicious pieces of content in the network. Almost all pieces of work on event analysis on Twitter (Section 2.3.1), heavily use profile and network features of its users for in-depth characterization

and analysis. The absence of these features in Facebook make event analysis on this network a tough problem for researchers.

Facebook also poses technical challenges and limitations on collection and analysis of the small proportion of data which is publicly available. We now discuss these in Section 2.4.2.

## 2.4.2 Technical limitations

Almost all online social media services, including Facebook, provide a range of Application Programming Interface (API) endpoints for users to interact with the service programmatically, and exchange data. Users can authenticate their identity with the service by providing (in most cases) a working phone number and register with the service as developers. After successful authentication, the platform grants a unique authorization token (or access token) to users which they can use to interact with the API programmatically. For Facebook, one of the easiest ways to obtain such an access token is through the Graph API explorer at <https://developers.facebook.com/tools/explorer/>. Figure 2.7 shows a screenshot of the Graph API explorer. Clicking on the “Get Token” button visible on the right of the interface generates the access token that can be seen in the “Access token” field in the middle.

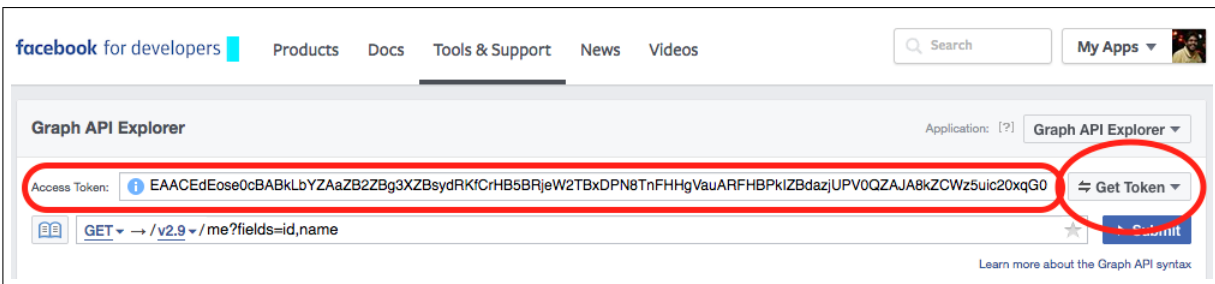


Figure 2.7: A snapshot of the Graph API explorer page offered by Facebook’s Developers Platform. Clicking on the “Get Token” button on the right generates an places an access token in the “Access token” field in the middle.

Once the token is obtained, any programming language can be used to construct an API query and retrieve data from the API. The code snippet below shows an example to query the Search API endpoint using Python.

```
import requests, urllib
base_url = 'https://graph.facebook.com/search?'
query = 'python'
token = 'XXXXXXXXXXXX'
params = urllib.urlencode({'q':query, 'access_token': token})
data = requests.get(base_url+params)
text = data.text
print text
```

Facebook provides a Search API which can be used to look for *public* posts containing a particular set of keywords.<sup>10</sup> This API endpoint is restricted to only public content, since getting access to non-public content requires authorization from the owner of the content. In order to collect public data specific to an event, this Search API needs to be supplied with relevant keywords. For example, keywords like *fifa*, *worldcup*, *fifaworldcup* can be used to collect public posts related to the FIFA World Cup. This is very similar to Twitter’s Search API<sup>11</sup>, which is commonly used to search for public tweets. However, a big drawback with Facebook is the absence of an API endpoint similar to Twitter’s Streaming API<sup>12</sup>, which can be used to collect event specific data in real-time. In addition, Facebook does not mention if the results returned by its Search API cover the entire public content generated on the network or only a portion of it. There is no way to verify the total number of posts generated on a topic, and the number of posts returned by the API. This can result in loss of an unknown amount of data. To overcome this issue to some extent, Facebook’s Search API needs to be queried at regular, short intervals of time. If the rate generation of public content during an event is too large, these intervals need to be tuned accordingly and shortened further.

As discussed in Section 2.3, event analysis on online social media has two broad components, viz. event detection, and data collection. Since event detection is a hard problem in itself, identifying events, and feeding them to a data collection framework requires human intervention. In order to automate this task, Twitter provides an API endpoint, which returns the *trending* topics that are being talked about, in a given locality, and at a given point in time.<sup>13</sup> Past research has shown an overlap of more than 85% between Twitter’s trending topics, and breaking / persistent news on Newswire [95]. Although these trending topics may not be an exact equivalent to events, they can be used to serve as input to the data collection process, thus eliminating the human in the loop. Facebook also launched a Trending feature for its users recently<sup>14</sup>, but its Trends API is not accessible for general public.<sup>15</sup> Again, this imposes a major implication, as the process of collecting Facebook data related to an event cannot be automated, and always requires human intervention, decreasing efficiency.

**Summary** There are multiple factors which make working with Facebook data a hard task. In this section, we discussed these factors and saw some of the limitations and challenges which exist in collection and analysis of public content on Facebook. In addition to the highly private nature

---

<sup>10</sup><https://developers.facebook.com/docs/graph-api/using-graph-api/v2.0#search>. The post search feature was deprecated on April 30, 2015. The Search API can still be used to search for pages, groups, users, events, etc. as of July 2017.

<sup>11</sup><https://dev.twitter.com/docs/api/1.1/get/search/tweets>

<sup>12</sup><https://dev.twitter.com/docs/api/1.1/post/statuses/filter>

<sup>13</sup><https://dev.twitter.com/docs/api/1.1/get/trends/place>

<sup>14</sup><http://techcrunch.com/2014/01/16/facebook-trending/>

<sup>15</sup><https://developers.facebook.com/docs/trends/v2.0>

of Facebook, we also saw some technical limitations, which hinder data collection on Facebook. However, these limitations and challenges do not eliminate the scope of working with Facebook data.

## 2.5 Discussion and Research Gaps

With this literature review, we attempted to highlight the existing gaps that exist in studying the Facebook network as one of the biggest channels for information dissemination during events. As evident from the existing literature, there is much need and scope of exploring the potential of the currently unmonitored public stream of Facebook content, and segregating poor quality content from useful information during events. Facebook’s content, especially during events, can prove to be a vital information channel, and hence needs utmost attention.

Building efficient, scalable solutions for big data services like Facebook, requires a representative sample of data for experimentation, and for drawing valid conclusions. However, as we saw in Section 2.1.3, getting a representative sample of the Facebook subgraph is a hard problem in itself. One of the major reasons for researchers being unable to get a convincing data sample is that Facebook’s fine-grained privacy settings make a majority of its content private, and publicly inaccessible. About 72% Facebook users set their posts to *private* [103]. This private nature of Facebook has been a major challenge in collecting and analyzing its content in the computer science research community.

Besides, existing techniques related to spread and mitigation of poor quality content on Facebook have not been studied comprehensively. Most of the techniques proposed for detecting malicious posts on Facebook lack comprehensive evaluation, which is essential to prove their worth and research contribution. There hardly exists any research in the computer science community which characterizes or analyzes malicious content on Facebook on a large scale. The only large scale study on Facebook [58] was on a dataset of 187 million wall messages which were collected from a random sample of 3.5 million users by crawling their Facebook walls in 2009. It would be interesting to study how malicious content identified from a random sample of Facebook differs from malicious content on Facebook during events. It is possible that the characteristics of malicious Facebook content vary across different events and differ from malicious Facebook content in general. It would be interesting to study if malicious content has evolved over time on Facebook.

We also saw some research attempts towards studying events from Facebook data (Section 2.3). However, Twitter has largely been the focus of researchers for studying events. We saw how Twitter was found to be a vital actor during sporting events, political campaigns, forest fires and even earthquakes. Content on other social networks has, however, not been given much attention in this respect. It is reasonable to assume that other social networks including Facebook also carry

event related content, which can be of importance to the population of Internet users where Twitter is not as widely used as some other social networks.

One of the major non-trivial tasks in identifying context-specific poor quality content in particular, is to be able to differentiate between poor quality content from benign content in the absence of any obvious or previously identified patterns. Moreover, past research has shown that it is much harder to identify an attack vector which contains contextual information as compared to a generic attack vector without context. Jagatic et al. demonstrated through an experiment that the number of victims who fell for a phishing attack containing contextual information was 4.5 times higher than the number of victims who fell for a regular phishing attack [81]. The presence of context makes it easier to gain the trust of the victim and make her fall for the attack. Also, since the attack vector contains context, it looks similar to genuine content, making it harder to identify even from a machine's point of view. Thus, features and methods used to identify generic attack vectors may not suffice to demarcate attack vectors with context. All these attributes make it non-trivial to identify context-specific poor quality content as compared to generic poor quality content.

## Chapter 3

# Towards Automatic Real Time Identification of Malicious Content on Facebook

This chapter is partly a reproduction of a paper published at the Annual Conference on Privacy, Security, and Trust (PST) 2015 [36] and a paper published at Social Network Analysis and Mining (SNAM) Journal 2017 [37].

### 3.1 Introduction

Social network activity rises considerably during events that make the news, like sports, natural calamities, etc. [148]. For example, the 2014 FIFA World Cup final inspired more than 618,000 tweets per minute, a new record for Twitter. Facebook also saw 350 million users generating over 3 billion posts, comments, and likes during the 32 days of the world cup.<sup>1</sup> This enormous magnitude of activity during sports and other news events makes OSNs lucrative venues for malicious entities to compromise system reputation and seek monetary gains. Facebook, being the most preferred OSN for users to get news [77], is potentially the most attractive platform for malicious entities to pollute and launch cyber-attacks. These attacks have become more sophisticated over the years, and are no longer limited to unsolicited bulk messages (spam and promotional campaigns), drive-by malware downloads, etc. Recently, cyber criminals exploited the context of various news events to spread hoaxes and misinformation, luring victims into scams, and phishing attacks on Facebook [105,172]. It has been claimed that Facebook spammers make \$200 million just by posting links [149]. Such activity not only degrades user experience but also violates Facebook's terms of

---

<sup>1</sup><http://edition.cnn.com/2014/07/14/tech/social-media/world-cup-social-media/>

service. In rare cases, hoaxes have reached the extent of claiming human lives.<sup>2</sup> Facebook has acknowledged spam and hoaxes as serious issues, and taken steps to reduce malicious content in users' newsfeed [116, 117].

Researchers have been studying malicious content in the form of spam and phishing for over two decades. However, the definition and scope of what should be labeled as "malicious content" on the Internet has constantly been evolving since the birth of the Internet. With respect to Online Social Networks, state-of-the-art techniques have become efficient in automatically detecting spam campaigns [58, 174], and phishing [4] without human involvement. Meanwhile, new classes of malicious content pertaining to appropriateness, authenticity, trustworthiness, and credibility of content have emerged in the recent past. Some researchers have studied these classes of malicious content on OSNs and shown their implications in the real world [19, 65, 69, 107]. All of these studies, however, resorted to human expertise to identify untrustworthy and inappropriate content and establish ground truth, due to the absence of efficient automated techniques to identify such content. We focus on a similar class of malicious content pertaining to trustworthiness and appropriateness in this work (in addition to traditional spam, and phishing), which currently requires human expertise to identify.

In this chapter, we address the problem of automatic real-time detection of malicious content generated during news-making events, that is currently evading Facebook's detection techniques [143]. To this end, we collect 4.4 million public posts generated by 3.3 million unique entities (users / pages) during 17 news-making events that took place between April 2013 and July 2014. We construct two ground truth datasets for malicious Facebook posts, using URL blacklists and human-annotation. Comparing the two datasets revealed various differences (and some similarities) amongst the malicious posts obtained using the two methodologies. Thus, we propose a two-fold scheme to identify malicious posts using two separate supervised learning models. We propose an extensive feature set consisting of 44 publicly available features to automatically distinguish malicious content from legitimate content in real-time. Unlike prior work [57, 58, 128], our technique does not rely on message similarity features which have been heavily used to detect spam campaigns in the past. In addition, we do not rely on the engagement level achieved by posts (*likes*, *comments*, etc.), since these attributes build up over time, and are unavailable at zero-hour.

Our experiments show that prior clustering based spam detection techniques are able to detect less than half the number of malicious posts as compared to our supervised learning model. We use our models to deploy Facebook Inspector (FbI), a REST API<sup>3</sup> based browser plug-in, that can be used to identify malicious content on Facebook in real-time. FbI is freely available for both Google Chrome and Mozilla Firefox browsers and has been downloaded over 2,500 times in the first nine months of its deployment. During this period, FbI received over 2.7 million requests and has

---

<sup>2</sup><http://news.discovery.com/human/psychology/social-media-ebola-hoax-causes-deaths-14100.htm>

<sup>3</sup>[http://multiosn.iiitd.edu.in/fbapi/endpoint/?version=2.0&fid=<post\\_id>](http://multiosn.iiitd.edu.in/fbapi/endpoint/?version=2.0&fid=<post_id>)

evaluated slightly over 0.97 million unique public Facebook posts. Using this data, we evaluated FbI in terms of response time and found that the response time for approximately 80% of all public posts analyzed by FbI was under 3 seconds. Our contributions are as follows:

- Characterization of malicious content generated on Facebook during news-making events. Our dataset of 4.4 million public posts is one of the largest datasets of public Facebook posts in literature.
- Development of an extensive feature set for identifying malicious content in real-time, excluding features like *likes*, *comments*, *shares*, etc. which are absent at post creation time.
- Conception of a two-fold filtering approach using models trained on separate ground truth datasets obtained through human-annotation and URL blacklists.
- Deployment of a publicly available end-user solution (Facebook Inspector) in the form of a REST API and a browser plug-in to identify malicious posts in real-time. We also evaluated FbI with real users and found it fast and useful in most cases.

The rest of the chapter is organized as follows. The methodology adopted for data collection and labeled dataset creation is described in Section 3.2. Sections 3.3 and 3.4 discuss the characterization and analysis of our datasets. The results of our automatic detection techniques are described in Section 3.5. Section 3.6 explains the limitations of our research in this chapter.

## 3.2 Methodology

There exists a wide range of malicious content on OSNs today. These include phishing, advertising campaigns, content originating from compromised profiles, artificial reputation gained through fake likes, etc. We do not intend to address all such attacks. We focus our research on identifying two kinds of malicious posts, a) posts containing a malicious URL, and b) posts which violate Facebook’s terms<sup>4</sup> or community standards<sup>5</sup> (containing hate speech, profanity, etc.) or are spam with respect to the event under consideration. We attempt to create automated means to detect such posts in real-time, without looking at the landing pages of the URLs (if any) and ignoring features which are absent at post creation time (*likes*, *comments*, etc.). We stress on not visiting the landing pages of URLs since this process induces time lag and increases the time taken by real-time systems to make a judgment on a post. Figure 3.1 represents the high-level flow diagram of the methodology we followed in this chapter.

---

<sup>4</sup><https://www.facebook.com/legal/terms>

<sup>5</sup><https://www.facebook.com/communitystandards>



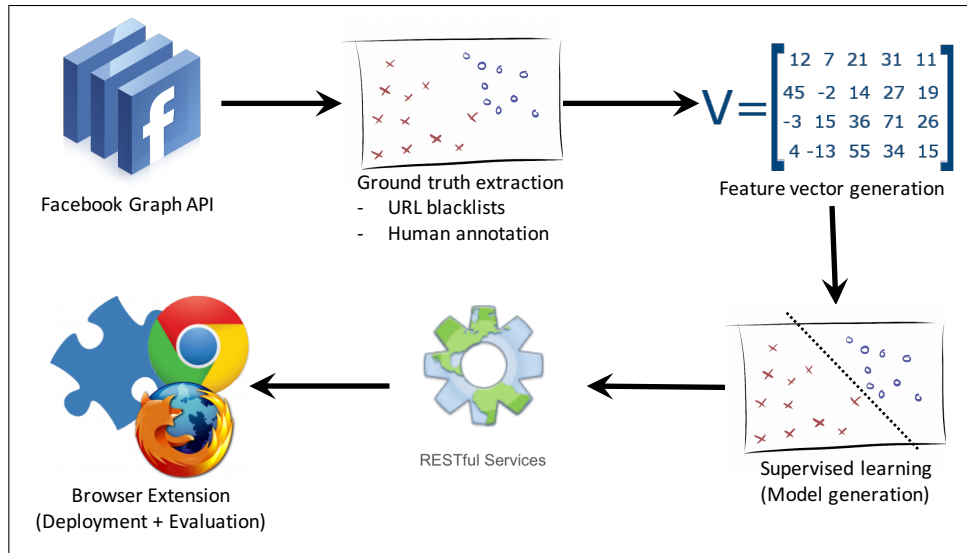


Figure 3.1: High level flow diagram capturing the methodology followed in this work.

### 3.2.1 Data collection

We collected data using Facebook’s Graph API Search endpoint [50] during 17 news-making events that took place between April 2013 and July 2014. We used event specific terms for each of the 17 events (see Table 3.1) to collect relevant public *posts*. All events we picked for our analysis, made headlines in international news. To maintain diversity, we selected events covering various domains like political, sport, natural hazards, terror strikes and entertainment news. For all 17 events, we started data collection from the time the event took place, and stopped about two weeks after the event ended.

Unlike other social networks like Twitter, Facebook does not provide an API endpoint to collect a continuous random sample of public posts in real-time. Thus, we used the search API to collect data. A drawback of the search method is that if a post is deleted or removed (either by the user herself, or by Facebook) before our data collection module queries the API, it would not appear in the search results. We repeated the search every 15 minutes to overcome this drawback as much as possible. In all, we collected over 4.4 million public Facebook posts generated by over 3.3 million unique users and pages. Table 3.2 shows the descriptive statistics of our final dataset.

### 3.2.2 Labeled dataset creation

Multiple techniques have been used in the past to obtain ground truth data for spam, fake profiles and other malicious content. These include looking up third party URL blacklists for phishing, malware, spam, and other kind of malicious URLs present in content, crowd-sourcing methods

Table 3.1: Event name, keywords used as search queries, number of posts, and description for the 17 events in our dataset of public Facebook posts.

Event ( <i>keywords</i> )	# Posts	Description
Missing Air Algeria Flight AH5017 ( <i>ah5017; air algerie</i> )	6,767	Air Algeria flight 5017 disappeared from radar 50 minutes after take off on July 24, 2014. Found crashed near Mali; no survivors.
Boston Marathon Blasts ( <i>prayforboston; marathon blasts; boston marathon</i> )	1,480,467	Two pressure cooker bombs exploded during the Boston Marathon at 2:49 pm EDT, April 15, 2013, killing 3 and injuring 264.
Cyclone Phailin ( <i>phailin; cyclonephailin</i> )	60,016	Phailin was the second-strongest tropical cyclone ever to make landfall in India on October 11, 2013.
FIFA World Cup 2014 ( <i>worldcup; fifaworldcup</i> )	67,406	20th edition of FIFA world cup, began on June 12, 2014. Germany beat Argentina in the final to win the tournament.
Unrest in Gaza ( <i>gaza</i> )	31,302	Israel launched Operation Protective Edge in the Hamas-ruled Gaza Strip on July 8, 2014.
Heartbleed bug in OpenSSL ( <i>heartbleed</i> )	8,362	Security bug in OpenSSL disclosed on April 1, 2014. About 17% of the world’s web servers found to be at risk.
IPL 2013 ( <i>ipl; ipl6; ipl2013</i> )	708,483	Edition 6 of IPL cricket tournament hosted in India, April-May 2013.
IPL 2014 ( <i>ipl; ipl7</i> )	59,126	Edition 7 of IPL cricket tournament jointly hosted by United Arab Emirates and India, April-May 2013.
Lee Rigby’s murder in Woolwich ( <i>woolwich; londonattack</i> )	86,083	British soldier Lee Rigby attacked and murdered by Michael Adebolajo and Michael Adebowale in Woolwich, London on May 22, 2013.
Malaysian Airlines Flight MH17 shot down ( <i>mh17</i> )	27,624	Malaysia Airlines Flight 17 crashed on 17 July 2014, presumed to have been shot down, killing all 298 on board.
Metro-North Train Derailment ( <i>bronx derailment; metro north derailment; metronorth</i> )	1,165	A Metro-North Railroad Hudson Line passenger train derailed near the Spuyten Duyvil station in the New York City borough of the Bronx on December 1, 2013. Four killed, 59 injured.
Washington Navy Yard Shootings ( <i>washington navy yard; navy yard shooting; NavyYardShooting</i> )	4,562	Lone gunman Aaron Alexis killed 12 and injured 3 in a mass shooting at the Naval Sea Systems Command (NAVSEA) headquarters inside the Washington Navy Yard in Washington, D.C. on Sept. 16, 2013.
Death of Nelson Mandela ( <i>nelson; mandela; nelsonmandela; madiba</i> )	1,319,745	Nelson Mandela, the first elected President of South Africa, died on December 5, 2013. He was 95.
Birth of the first Royal Baby ( <i>RoyalBabyWatch; kate middleton; royalbaby</i> )	90,096	Prince George of Cambridge, first son of Prince William, and Catherine (Kate Middleton), was born on July 22, 2013.
Typhoon Haiyan ( <i>haiyan; yolanda; typhoon philippines</i> )	486,325	Typhoon Haiyan (Yolanda), one of the strongest tropical cyclones ever recorded, devastated parts of Southeast Asia on Nov. 8, 2013.
T20 Cricket World Cup ( <i>wt20; wt2014</i> )	25,209	Fifth ICC World Twenty20 cricket competition, hosted in Bangladesh during March-April, 2014. Sri Lanka won the tournament.
Wimbledon Tennis 2014 ( <i>wimbledon</i> )	2,633	128th Wimbledon Tennis championship held between June 23, and July 6, 2014. Novak Djokovic from Serbia won the championship.

such as human annotations through services like Amazon’s Mechanical Turk <sup>6</sup>, CrowdFlower <sup>7</sup>, annotation / coding by topic experts, etc. However, most work in the past has employed only one of these techniques to obtain ground truth data for solving a problem. Although each of these techniques are widely accepted, they may not necessarily achieve completeness when used individually. URL blacklists, for example, are likely to miss out on self-promoting or advertising spam generated by Facebook applications since the *facebook.com* domain never appears on a URL blacklist. This kind of content, however, can easily be identified and marked by human annotators. On the other hand, using human annotation alone causes scalability issues and cannot be used to identify ground truth from large datasets containing millions of posts. This technique is often used on small samples drawn from large populations, making it susceptible to missing out on true positives which may not be captured in the sample under consideration. Using automated techniques like machine learning on such small datasets poses further complications like over-fitting etc., which need to be addressed separately.

To obtain ground truth data for malicious Facebook posts, we employed two methodologies, viz. URL blacklist lookup, and human annotation, and created two separate datasets of malicious posts. We now discuss the methodologies for each of these datasets in detail.

### **Dataset I: Using URL blacklists**

To create a labeled dataset of Facebook posts containing a malicious URL, we started by filtering out all posts containing one or more URLs. These URLs were added to the set of URLs present in the *link* field (if available) for each post. These extracted URLs were then visited using Python’s Requests package. <sup>8</sup> In case the Requests package failed, the URLs were visited using LongURL API. <sup>9</sup> Visiting the landing pages of the URLs helped us to eliminate invalid URLs and capture the correct final destination URLs corresponding to shortened URLs. After the extraction and validation process, we were left with a total of 480,407 unique URLs across 1,222,137 unique posts (see Table 3.2). Each URL was then subjected to six blacklist lookups, viz. Google Safebrowsing [61], SURBL [146], PhishTank [111], SpamHaus [142], VirusTotal [76], and Web of Trust [167] in October 2014. This methodology of identifying malicious content using URL blacklists has also been used in past research [28, 58].

For a given domain, the scan results returned by the VirusTotal API contain domain information from multiple services like TrendMicro, BitDefender, WebSense ThreatSeeker, etc. We marked a URL as malicious if one or more of these services categorized the domain of the URL as *spam*, *malicious*, or *phishing*. The Web of Trust (WOT) API returns a reputation score for a given domain.

---

<sup>6</sup><https://www.mturk.com/mturk/welcome>

<sup>7</sup><http://www.crowdfunder.com/>

<sup>8</sup><http://docs.python-requests.org/en/latest/>

<sup>9</sup><http://longurl.org/api>

Table 3.2: Descriptive statistics of complete dataset collected over April 2013 - July 2014.

Unique posts	4,465,371
Unique entities	3,373,953
- Unique users	2,983,707
- Unique pages	390,246
Unique URLs	480,407
Unique posts with URLs	1,222,137
Unique entities posting URLs	856,758
Unique posts with malicious URLs	11,217
Unique entities posting malicious URLs	7,962
Unique malicious URLs	4,622

Reputations are measured for domains in several *components*, for example, trustworthiness, and child safety. For each  $\{\text{domain}, \text{component}\}$  pair, the system computes two values: a *reputation* estimate and the *confidence* in the reputation. Together, these indicate the amount of trust in the domain in the given component. A *reputation* estimate of below 60 indicates *unsatisfactory*. The WOT browser add-on requires a confidence value of  $\geq 10$  before it presents a warning about a website. We tested the domain of each URL in our dataset for two components, viz. *Trustworthiness* and *Child Safety*. For our experiment, a URL was marked as malicious if both the aforementioned conditions were satisfied (Algorithm 1). In addition to reputations, the WOT rating system also computes categories for websites based on votes from users and third parties. We marked a URL as malicious if it fell under the *Negative* (including malware, scams etc.) or *Questionable* (including hate, incidental nudity etc.) category group (Table 3.3). Further, a URL was marked malicious if it was marked as malicious by SURBL, Google Safebrowsing, SpamHaus or PhishTank.

Table 3.3: Category labels and descriptions returned by the WOT API. Source: WOT API Wiki (<https://www.mywot.com/wiki/API>), last retrieved on May 31, 2016.

Category	Description
Negative	Malware, viruses, poor customer experience, phishing, scam, potentially illegal, adult content
Questionable	Misleading claims, unethical, privacy risks, suspicious, hate, discrimination, spam, potentially unwanted programs, ads, pop-ups, incidental nudity, gruesome / shocking

The reason for including WOT reputation scores in our labeled dataset of malicious posts was two-fold. Firstly, to study Facebook’s current techniques to counter malicious content. Facebook partnered with WOT to protect its users from malicious URLs [49]. Secondly, during news-making events, malicious entities tend to engage in spreading fake, untrustworthy and adult content to degrade user experience [67]. This kind of information, despite being malicious, is not captured

---

**Algorithm 1** Detecting malicious posts from WOT reputation scores

---

```
for all posts do
  for all URL domains do
    components = GetComponentFromWOT_API
    for all components do
      if reputation < 60 and confidence ≥ 10 then
        post = malicious
      end if
    end for
  end for
end for
```

---

by blacklists like Google Safebrowsing and SURBL, since they do not fall under the more obvious kinds of threats like malware and phishing. WOT scores helped us to identify and tag such content. In all, we found 4,622 unique malicious URLs across 11,217 unique Facebook posts (see Table 3.2). We refer to this dataset as **Dataset I** for the rest of the chapter.

### Dataset II: Using human annotation

We took help from human annotators to obtain another ground truth dataset regarding whether a post is malicious or not, in context of the event under consideration. Human annotation for understanding the ground truth is a well-established research methodology and has been widely used to categorize OSN data [5,11,19]. We recruited annotators through word-of-mouth publicity in multiple educational institutions. All our annotators were undergraduate computer science students between the age of 17 and 21, and were regular Facebook users. For the 17 selected events, we picked a random sample of 500 posts per event for annotation. We developed a web interface for the annotation task and assigned unique login credentials (username and password) to each annotator. We offered a monetary reward of approximately US \$4 to each annotator for annotating one event (500 posts). We asked the annotators to select one of three options for each post: *This is spam*, *This is not spam*, or *Can't say (skip)*. To help the annotators make a decision, we gave them the following definition of spam in general, and spam in terms of online social networks:

*“Any irrelevant or unsolicited messages sent over the Internet, typically to large numbers of users, for the purposes of advertising, phishing, spreading malware, etc. are categorized as spam. In terms of online social media, social spam is any content which is irrelevant / unrelated to the event under consideration, and / or aimed at spreading phishing, malware, advertisements, self promotion etc., including bulk messages, profanity, insults, hate speech, malicious links, fraudulent reviews, scams, fake information etc.”*

Figure 3.4 shows a screenshot of the annotation portal. Each post was annotated by three different annotators. A total of 25,500 judgments were made (17 events × 500 posts × 3 annotations) to establish the ground truth dataset.

For all our analysis, we only selected posts for which, all three annotators agreed upon the same label. This was done to ensure that our ground truth dataset is of best possible quality. After eliminating posts with partial agreement, we were left with a final dataset of 4,412 posts (571 spam and 3,841 not spam). We refer to this dataset as **Dataset II** for the rest of the chapter.

### 3.3 Analysis: Dataset I

We first present our findings about the efficiency of Facebook’s current techniques of malicious content detection and the differences between malicious and legitimate content on Facebook using Dataset I.

#### 3.3.1 Efficiency of Facebook’s current techniques

Facebook’s immune system uses multiple URL blacklists to detect malicious URLs in real-time and prevent them from entering the social graph [143]. Understandably, the inefficiency of blacklists to detect URLs at zero-hour limits the effectiveness of this technique [141]. We queried the Graph API in November 2014 to check if Facebook removed any of the 11,217 malicious posts<sup>10</sup> from our Dataset I after being posted. We found that only 3,921 out of the 11,217 (34.95%) malicious posts had been deleted. It was surprising to note that almost two thirds of all malicious posts (65.05%) which got past Facebook’s real-time detection filters remained undetected even after 4 (or more) months (July - November, 2014) from the date of post. Collectively, these posts had gathered *likes* from 52,169 unique users and *comments* from 8,784 unique users at the time we recollected the data. Using the URL endpoint of the Graph API<sup>11</sup>, we also found that the 4,622 unique URLs present in the 11,217 malicious posts had been shared on Facebook over 37 million times. Figure 3.2 shows one such malicious post from our dataset which went undetected by Facebook. The short URL in the post points to a scam website which asks users to *like* posts on Facebook to earn money.



Figure 3.2: One of the 7,296 malicious posts from our dataset which were not deleted by Facebook. We revisited this post after 11 months of being posted.

<sup>10</sup>We refer to a post as *malicious* if it contains a malicious URL.

<sup>11</sup><https://developers.facebook.com/docs/graph-api/reference/v2.2/url>

Above analysis suggests that a large portion of malicious content which goes undetected by Facebook's filters not only stays undetected, but thrives on users' *likes*, *comments* and *shares*. With 4.75 billion posts generated on Facebook every day [48], re-scanning all posts to check for malicious content can be computationally expensive. This demands for an alternative real-time detection technique which does not rely on blacklists to identify malicious content.

**WOT warning pages:** Facebook partnered with Web of Trust in 2011 to protect its users from malicious URLs [49]. According to this partnership, Facebook claims to show a warning page to the user whenever she clicks on a link which has been reported for spam, malware, phishing or any other kind of abuse on WOT (Figure 3.3). To verify the existence of this warning page, we manually visited a random sample of 1,000 posts containing a URL marked as malicious by WOT, and clicked on the URL. Surprisingly, the warning page did not appear even once. We also noticed that over 88% of all malicious URLs in our dataset (4,077 out of 4,622) were marked as malicious by WOT. Although we could not identify the reason behind the absence of the proposed warning pages, their absence indicated much scope and need for techniques to make Facebook users aware of the potential risks posed by URLs posted on the network.



Figure 3.3: Warning page supposed to be shown by Facebook whenever a user clicks on a link reported as abusive on WOT.

### 3.3.2 Key characteristics of posts containing malicious URLs

We analyzed Dataset I in three aspects – a) textual content and URLs, b) entities who post malicious content, and c) metadata associated with malicious content. We now look at all these three aspects individually.

### Spam Annotation Portal - Typhoon Haiyan

[Log out](#)

**About the event**

Typhoon Haiyan, known in the Philippines as Typhoon Yolanda, was one of the strongest tropical cyclones ever recorded, which devastated portions of Southeast Asia, particularly the Philippines, on November 8, 2013. It is the deadliest Philippine typhoon on record, killing at least 6,300 people in that country alone. Haiyan is also the strongest storm recorded at landfall, and unofficially the strongest typhoon ever recorded in terms of wind speed. As of January 2014, bodies were still being found.

After becoming a tropical storm and attaining the name Haiyan at 0000 UTC on November 4, the system began a period of rapid intensification that brought it to typhoon intensity by 1800 UTC on November 5. By November 6, the Joint Typhoon Warning Centre (JTWC) assessed the system as a Category 5-equivalent super typhoon on the Saffir-Simpson hurricane wind scale; the storm passed over the island of Kayangel in Palau shortly after attaining this strength.


---

**Note:** If you see a URL in the message, **please visit the URL** to help make a judgement.

---

**Message:** yolanda go away....

**Story:** Melissa Larazo Mendez shared I LOVE MY CRUSH : "> <3's photo.

**Picture:** 

**Link:** <http://www.facebook.com/photo.php?fbid=688756251149435&set=a.146452432046489.22214.123265261031873&type=i>

Posts  
left: 473

Figure 3.4: Snapshot of the web interface presented to the annotators. The interface also presented a short description of each of the 17 events.

## Textual content and URLs

We first looked at the most commonly appearing posts in Dataset I. Similar to past work [58], we found various *campaigns* promoting a particular entity or event. However, campaigns in our dataset were very different than those discussed in the past. Table 3.4 shows the top 10 campaigns in our dataset of malicious posts. We found that most of the campaigns in our dataset were event specific, and talked about celebrities and famous personalities who were part of the event. Although this seems fairly obvious because of our event based dataset, such campaigns reflect the attackers' preferences of using the context of an event to target OSN users. Attackers now prefer to exploit users' curiosity about news-making events in addition to hijacking trends and posting unrelated content (like promoting free iPhone, illegal drugs, cheap pills, free ringtones, etc.) using topic specific keywords to spread malicious content.

Investigating further, we found that the most common type of malicious posts (52.0%) in our dataset were the ones with URLs pointing to adult content and incidental nudity, and marked unsafe for children by WOT. The second most common type of malicious posts comprised of negative and questionable category URLs. These categories comprised of malware, phishing, scam, misleading claims or unethical, spam, hate, discrimination, potentially unwanted programs, etc., and accounted for 45.2% of all posts. Posts containing untrustworthy sources of information (38.22%)



Table 3.4: Top 10 most common posts in our dataset of malicious posts.

Post Summary	Count
Sexy Football Worldcup - Bodypainting	155
10 Things Nelson Mandela Said That Everyone Should Know	154
Was Bishop Desmond Tutu Frozen Out of Nelson Mandela’s Funeral?	105
Nude images of Kate Middleton	73
The Gospel Profoundly Affected Nelson Mandela’s Life After Prison	72
Promotion of Obamacare (Affordable Care Act) through Nelson Mandela’s death	67
Radical post about Nelson Mandela	54
Was Nelson Mandela a Christian?	41
R.I.P. Nelson Mandela: How he changed the world	36
Easy free cash	29

were the third most common type of malicious posts. Interestingly, only 325 malicious posts (2.9%) advertised a phishing URL. This is a drastic drop as compared to the observations made by Gao et al. in 2010, where authors found that over 70% of all malicious posts in their dataset advertised phishing [58]. We also found that 18.4% of the malicious posts in our dataset (2,064 posts out of 11,217) advertised one or more shortened URLs. Past literature has shown wide usage of shortened URLs to spread malicious content on microblogging platforms [7, 27]. Use of short URLs has significantly increased mostly due to restriction of message length on OSNs like Twitter. However, restriction on message length does not apply on Facebook. This implies that the primary reason behind usage of shortened URLs on Facebook is obfuscation of actual malicious URLs.

In addition to post categories, we also looked at the most common URL domains in our dataset. We observed that Facebook and YouTube constituted almost 60% of all legitimate URLs shared during the 17 events. The remaining legitimate URLs largely belonged to news websites (cnn.com, bbc.co.uk, etc.). On the contrary, malicious URLs were more evenly distributed across a mixture of news, blogs, sports, entertainment, etc. websites. Our dataset revealed that a large fraction of malicious content comprised of untrustworthy sources of information, which may have inappropriate implications in the real world, especially during events like elections, riots, etc. Most previous studies on detecting malicious content on online social networks have concentrated on identifying more obvious threats like malware and phishing [11, 62, 163]. There exists some work on studying trustworthiness of information on micro-blogging platforms like Twitter [19, 65]. However, to the best of our knowledge, no past work addresses the problem of identifying untrustworthy content on Facebook.

## Entities posting malicious content

Content on Facebook is generated by two types of entities – *users* and *pages*. Pages are public profiles specifically created for businesses, brands, celebrities, causes, and other organizations. Unlike users, pages gain “fans,” people who choose to *like* a page. In our dataset, we identified pages by the presence of *category* field in the response returned by Graph API search [50] during the initial data collection process. The *category* field is specific to pages; we used this field to differentiate between pages and user profiles. We found that pages were more active in posting malicious URLs as compared to legitimate URLs. Pages were observed to constitute 21% (1,676 out of 7,962) of all malicious entities, while only 10% of all legitimate URL posting entities were pages. A similar percentage of pages (12%) was found to constitute all legitimate entities in our dataset. We also found 43 verified pages and 1 verified user among entities who posted malicious content. The most common type of verified pages were radio station pages (12; e.g. 957thebeat, catcountry987), website pages (5; e.g. guyismdotcom, Biographile) and public figure pages (4; e.g. Unmuktchandofficial, danijohnsonlive). Combined together, the 43 verified pages had over 71 million *likes*.

It is important to note that most of the past attempts at studying malicious content on Facebook did not capture content posted by pages, and concentrated only on users [5, 58, 145]. Malicious content originating from pages in our dataset brings out a new dimension, which hasn’t been studied in the past. Facebook limits the number of *friends* a user can have, but there is no limit on the number of people who can *like* (subscribe to) a page. Content posted by a page can thus, have much larger audience than that of a user, making malicious content posted by pages potentially more widespread and dangerous than that posted by individual users. We found that in our dataset, pages posting malicious content had 123,255 *likes* on average (min. = 0, max. = 50,034,993), whereas for legitimate pages, the average number of *likes* per page was only 45,812 (min. = 0, max. = 502,938,006).

## Metadata

There are various types of metadata associated with a post, for example, application used to post, time of post, type of post (picture / video / link), location etc. Metadata is a rich source of information that can be used to differentiate between malicious and legitimate users.

Facebook allows users to post content through a variety of platforms (applications), viz. web (browsers), mobile devices (apps for Android, iOS etc.), cross platform and media sharing applications (HootSuite, Photos, Videos, etc.), and other third party applications. Figure 3.5 shows the distribution of the top 25 applications used to post content in Dataset I.<sup>12</sup> We observed that over 51% of all legitimate content was posted through mobile apps. This percentage dropped to below

---

<sup>12</sup>The top 25 applications were used to generate over 95% of content in all three categories we analysed.

15% for malicious content. Third party applications were used to generate 6% of all malicious content in our dataset as compared to only 0.11% of all legitimate content being generated by such applications. This behavior reflects that malicious entities make use of web and third party applications (possibly for automation) to spread malicious content, and can be an indicator of malicious activity. Legitimate entities, on the other hand, resort to standard mobile platforms to post.

Although Facebook has more web users than mobile users [46], our observations may be biased towards mobile users due to our event specific dataset. Past literature has shown high social network activity through mobile devices during such events [67].

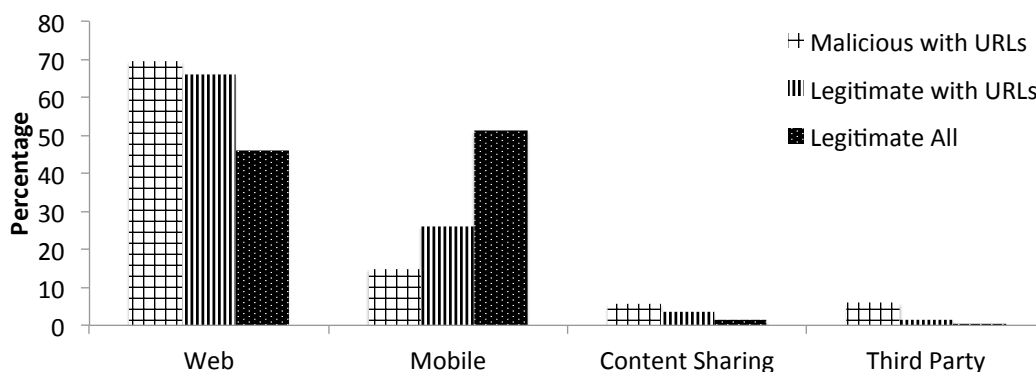


Figure 3.5: Sources of malicious content, legitimate content with URLs, and all legitimate content. Mobile platforms were preferred over web for posting legitimate content.

We also observed significant difference in the content *types* that constituted malicious and legitimate content. Over 50% of legitimate posts containing a URL were photos or videos whereas this percentage dropped to below 6% for malicious content. A large proportion of these photos and videos were uploaded on Facebook itself. This was one of the main reasons for facebook.com being the most common legitimate domain in our dataset. We used these, and some other features to train multiple machine learning algorithms for automatic detection of malicious content. The results of our experiments are presented in Section 3.5. We now look present our findings from our analysis of Dataset II.

### 3.4 Analysis: Dataset II

Similar to our findings from Dataset I, we characterized Dataset II and observed some key characteristics which separated human-annotated malicious posts from legitimate posts. We focus on a similar set of characteristics as Dataset I.

### 3.4.1 Textual content

Similar to Dataset I, we found three campaigns in Dataset II advertising different URLs from the same domain. Interestingly, the top two campaigns comprised of posts containing URLs pointing to Facebook applications. Table 3.5 shows the details of these campaigns we found in Dataset II. We observed that only 11% of the posts which were part of the Hyundai Nepal campaign (first row in Table 3.5) contained a text message, and none of the posts in the other two campaigns contained a text message. Further, unlike Dataset I, we did not find any campaigns based on textual similarity in Dataset II. The main reason for this finding can be attributed to the fact that more than half the posts in Dataset II did not have any text messages at all.

Table 3.5: Campaigns found in Dataset II. Two of the three campaigns existed across more than one event.

Campaign (URL)	App. name	Size in Dataset II	Total size	No. of events affected
<a href="https://www.facebook.com/hyundainepal/app_652950454797250">https://www.facebook.com/hyundainepal/app_652950454797250</a>	Hyundai Worldcup Prediction	32	5,112	1
<a href="https://apps.facebook.com/sskampot/?source=feed&amp;like_book=*.user_likes_*">https://apps.facebook.com/sskampot/?source=feed&amp;like_book=*.user_likes_*</a>	Smile	100	1,650	7
<a href="http://www.iforum.gr/?arthro=*">http://www.iforum.gr/?arthro=*</a>	iForum.gr	23	481	4

We noticed that 66.19% of malicious posts in Dataset II (378 out of 571) did not contain any text message, this percentage dropped to 14.16% (544 out of 3,841) for legitimate posts. For Dataset I, the percentage of posts containing text messages was considerably higher, and did not vary much across malicious (73.6%) and legitimate posts (88.6%). We also noticed that majority of legitimate posts in Dataset II comprised of short text messages containing “prayers” for victims (in case of crisis events), updates and personal opinions (in case of non-crisis events). On the other hand, spam posts were often long pieces text containing advertisements, promotional content, and propaganda. On average, text messages in spam posts in Dataset II were 6.1 times longer than legitimate posts. We observed a similar trend for Dataset I, where text messages in malicious posts were approximately 3 times longer than legitimate posts. These numbers indicate rudimentary differences between spam and legitimate posts generated during news-making events, and highlight that posts containing short text messages are indicators of legitimate content.

It was interesting to note that 250 out of the 571 malicious posts (43.78%) contained URLs pointing to Facebook apps, as compared to zero legitimate posts containing URLs pointing to Facebook apps in Dataset II. We found no malicious posts pointing to Facebook apps in Dataset I, since Dataset I was obtained using URL blacklists and *facebook.com* or *apps.facebook.com* domains never appeared on any URL blacklist. This characteristic highlights the inability of URL blacklists in detecting

malicious content, and hence a drawback of detection models built using ground truth data obtained from URL blacklists [28, 58]. This also signifies that Facebook posts containing URLs to Facebook apps are a strong indicator of malice during news-making events.

### 3.4.2 Entities posting malicious content

Observations from Dataset II revealed a uniform distribution of male users, female users, and pages across malicious and legitimate posts. We observed that approximately 22% of all content originated from pages across both malicious and legitimate posts. Similarly, the `male:female` gender ratio was found to be approximately 3:1 for both malicious and legitimate posts. These observations were contradictory from our observations in Dataset I where `male:female` ratio and `user:page` ratio were different across malicious and legitimate posts.

### 3.4.3 Metadata

Facebook posts have a separate *link* field (Figure 3.6), which captures URLs (if any) present in text messages. We noticed that 82.13% of all malicious posts (469 out of 571) in Dataset II contained a URL, whereas only 34.44% of legitimate posts (1,323 out of 3,841) contained a URL. This behavior was contrasting as compared to text messages (discussed previously), where majority of malicious posts in Dataset II *did not* contain any text. In all, we observed that 66.19% malicious posts in Dataset II were isolated URLs posted with no supporting textual content, whereas this percentage dropped to 13.66% for legitimate posts. This behavior highlights that posts containing isolated URLs are indicators of malicious content.

```
{
  "link": "http://bbc.in/2uKMSG3",
  "shares": {
    "count": 38
  },
  "application": {
    "category": "Business",
    "link": "http://socialflow.com/",
    "name": "SocialFlow",
    "namespace": "socialflowapp",
    "id": "128869720483178"
  },
  "created_time": "2017-07-22T08:32:26+0000",
  "description": " ",
  "message": "Boots UK had said it didn't want to \"incor",
  "id": "228735667216_10154923834512217"
}
```

Figure 3.6: The *link* field present in the post object returned by the Facebook Graph API. This field is absent when the post does not contain a link.

To see if URLs contained in posts in Dataset II were malicious, we looked up six URL blacklists, viz.

Web of Trust, Google Safebrowsing, Phishtank, SpamHaus, SURBL, and VirusTotal for each URL present in the 469 malicious posts containing a URL. We found that only three out of the 469 posts contained a malicious URL captured by at least one of the six URL blacklists we looked up. Prior approaches to detect spam and other types of malicious content on Facebook have heavily relied on URLs blacklisted by third parties, and spam keywords in posts to identify spam posts [58,128]. Our observations reveal that these approaches would fail in our case due to the absence of blacklisted URLs, and text content in posts generated during news-making events.

Figure 3.7 shows the distribution of all applications used to post content in Dataset II. We observed that 45.23% of all legitimate content was posted through mobile devices, followed by 35.95% content originating from the web (Figure 3.5). These percentages dropped significantly to 15.93% and 20.84% respectively in case of malicious posts. Over 51% of malicious posts originated from third party applications, which was a drastic increase from 6.74% in case of legitimate content. This behavior was contradictory to our observations from Dataset I, where majority of malicious posts originated from the web. However, the increased participation of third party applications in generating malicious posts was in line with our observations from Dataset I.

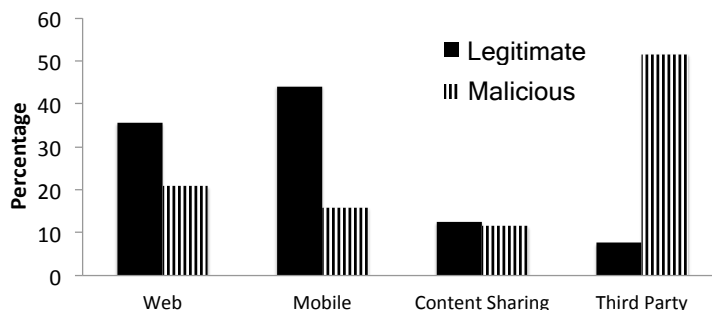


Figure 3.7: Distribution of platforms used to post malicious and legitimate content in our dataset. Mobile was the most preferred platform for posting legitimate content. Third party applications were most common in malicious posts.

### 3.5 Detecting malicious content automatically

Past efforts for automatic detection of spam and malicious content on Facebook largely focus on detecting campaigns [57, 58], and rely heavily on *message similarity* features to detect malicious content [128]. Researchers using this approach have reported consistent accuracies of over 80% using small feature sets comprising of 6-7 features. However, this approach is ineffective for detecting newly emerged malicious content since the aforementioned models require to have seen similar spam messages in the past. To overcome this inability, we propose an extensive set of 44 features (see Table 3.6) to detect malicious content, excluding features like message similarity, likes, comments,

shares etc., which are absent when new malicious posts surface. We group these 44 features into four categories based on their source; Entity, Text content, Metadata, and Link.

Table 3.6: Features used for machine learning experiments. We extracted features from four sources, viz. entity, content, metadata, and link.

Source	Features
Entity (9)	is a page / user, gender, page category, has username, username length, name length, no. of words in name, locale, likes on page
Text content (18)	Presence of !, ?, !!, ??, emoticons (smile, frown), no. of words, avg. word length, no. of sentences, avg. sentence length, no. of English dictionary words, no. of hashtags, hashtags per word, no. of characters, no. of URLs, no. of URLs per word, no. of uppercase characters, no. of words / no. of unique words
Metadata (10)	App, has <i>facebook.com</i> URL, has Facebook <i>app</i> URL, has Facebook <i>event</i> URL, has <i>message</i> , has <i>story</i> , has <i>link</i> , has <i>picture</i> , type, <i>link</i> length
Link (7)	has HTTP / HTTPS, hyphen count, parameters count, parameter length, no. of subdomains, path length

Using our observations from Section 3.3 and Section 3.4, we included features like *App* (platform used to post), post *type*, entity gender, page *category*, and entity class (entity is page or user), which we found to be differentiating the malicious and benign class. In addition, we tried to capture as much publicly available information corresponding to a post as possible while constructing the feature set, in order to make our model robust. All categorical features were one-hot encoded and converted to an array of binary features to make them compatible with the classification algorithms.

We performed experiments using multiple supervised learning techniques on both our datasets, including probabilistic classification techniques (Gaussian Naive Bayes), decision trees, ensemble methods (Random Forests), and Support Vector Machines. We now present our analysis on Dataset I and Dataset II separately. The same set of 44 features was used to train separate models on Dataset I and Dataset II. All experiments were performed using Python Scikit-learn library.<sup>13</sup>

### 3.5.1 Dataset I

We trained four supervised learning models using 11,217 unique malicious posts as the positive class and 11,217 unique legitimate posts, randomly drawn from the 1,210,920 unique legitimate posts in Dataset I (see Table 3.2) as the negative class. We performed 10-fold cross validation on the training set for evaluation. In addition, each experiment was performed ten times, each time using a random sample drawn from the negative class (legitimate posts), and keeping the positive class (malicious posts) constant. The results were averaged out across all ten observations. Standard deviation in accuracy across all ten observations was less than 1.5% for all experiments.

<sup>13</sup><http://scikit-learn.org>

Table 3.8 describes the results averaged across all ten experiments in detail. The Random Forest Classifier performed the best and achieved an average accuracy of 85.05%, with an average recall of 86.89%. Ensemble methods, especially the Random Forest algorithm, are known to combine multiple weak learners to come up with a strong learner, which have been known to outperform probability based and support vector based algorithms in practice. This phenomenon was evident in Dataset I, since we observed multiple weak differentiators (platform used to post, entities, etc.), and no strong differentiators between the two classes (poor quality and benign). This presence of multiple weak differentiators is the most likely reason for the Random Forest classifier performing better than the other techniques.

We also performed classification experiments using the four category features separately, and observed that link features performed the best, yielding an accuracy of 82.56% using the Random Forest Classifier. A combination of all four category features, however, outperformed the individual category scores, signifying that none of the category features individually could identify malicious posts as accurately as their combination.

To study the effect of the number of features on the model, we calculated 10-fold cross validation accuracy of the Random Forest Classifier using 1 through all 44 features, adding features one by one in decreasing order of their feature importance value. Feature importance was calculated using the *feature\_importances\_* attribute in Python Sklearn’s RandomForestClassifier method.<sup>14</sup> This attribute computes the Gini importance which measures the average gain of purity by splitting a given feature [16]. Gini importance is one of the most widely used metrics in literature for computing feature importance in tree based machine learning algorithms. We found that the accuracy peaked to 86.62% at the top 7 features (Figure 3.8(a)). All four classifiers achieved higher accuracy when trained on the top 7 features, as compared to accuracy when trained on all 44 features (see Table 3.8). Table 3.7 shows the normalized feature importance value and source of the top 10 features.

Presence of a Facebook.com URL was found to be the most important feature in our model (see Table 3.7). This was mainly because the facebook.com domain appeared in approximately half of all legitimate posts in Dataset I. However, facebook.com never appeared in malicious posts since it never appeared on any of the URL blacklists from where we obtained our ground truth for Dataset I (as discussed previously). We also found the “post type” feature to be the third most important feature, conforming to our initial observations from Section 3.3.2 where we found considerable difference in malicious and legitimate post *types*. Figure 3.8(b) shows the receiver operating characteristics (ROC) curve for the Random Forest classifier using all 44 features, where we achieved maximum area under curve (AUC) of 0.935.

---

<sup>14</sup><http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>



Table 3.7: Source and feature importance value (normalized) of the top 10 features.

Feature	Source	Importance
Presence of Facebook.com URL	Metadata	0.0974
Length of <i>link</i> field	Metadata	0.0783
Post type	Metadata	0.0708
Length of parameter(s) in URL	Link	0.0695
Length of URL path	Link	0.0694
Number of hyphens in URL	Link	0.0577
Number of sub-domains in URL	Link	0.0549
Length of name	Entity	0.0363
Length of username	Entity	0.0351
Average sentence length	Text	0.0328

Table 3.8: Ten-fold cross validation accuracies for four classifiers over six different feature sets.

Feature Set	Metric	Entity	Text	Metadata	Link	All	Top 7
Naive Bayesian	Accuracy	54.79	52.41	71.60	69.25	56.15	74.72
	Precision	76.53	70.61	64.30	73.89	78.62	68.05
	Recall	13.82	8.29	97.23	59.57	16.91	93.93
	ROC_AUC	0.552	0.636	0.784	0.766	0.566	0.799
Decision Tree	Accuracy	63.02	64.78	80.56	82.34	84.67	86.17
	Precision	65.00	69.84	77.40	77.57	84.20	83.81
	Recall	56.45	52.00	86.36	91.03	85.47	89.65
	ROC_AUC	0.663	0.645	0.872	0.885	0.847	0.883
Random Forest	Accuracy	63.47	66.25	80.67	82.56	85.05	<b>86.62</b>
	Precision	64.94	73.85	76.86	77.48	83.76	<b>83.97</b>
	Recall	58.25	50.70	87.45	91.66	86.89	<b>90.56</b>
	ROC_AUC	0.696	0.705	0.879	0.901	0.935	<b>0.935</b>
SVM (rbf)	Accuracy	61.77	64.89	78.75	81.45	75.89	83.66
	Precision	74.80	67.74	72.42	75.36	69.46	78.28
	Recall	35.53	56.86	92.93	93.49	92.45	93.20
	ROC_AUC	0.665	0.693	0.843	0.856	0.860	0.894

### General model versus event-specific model

The focus of this work is to address malicious posts generated on Facebook *during news-making events*. It can be argued that malicious Facebook posts *in general* may not be different from malicious posts appearing *during news-making events* and thus, a general model trained on a random sample of Facebook posts may suffice to mitigate all malicious posts irrespective of whether they are generated during an event or not. However, from our initial observations, we hypothesize that this approach would be incompetent since malicious posts *in general* are different from malicious

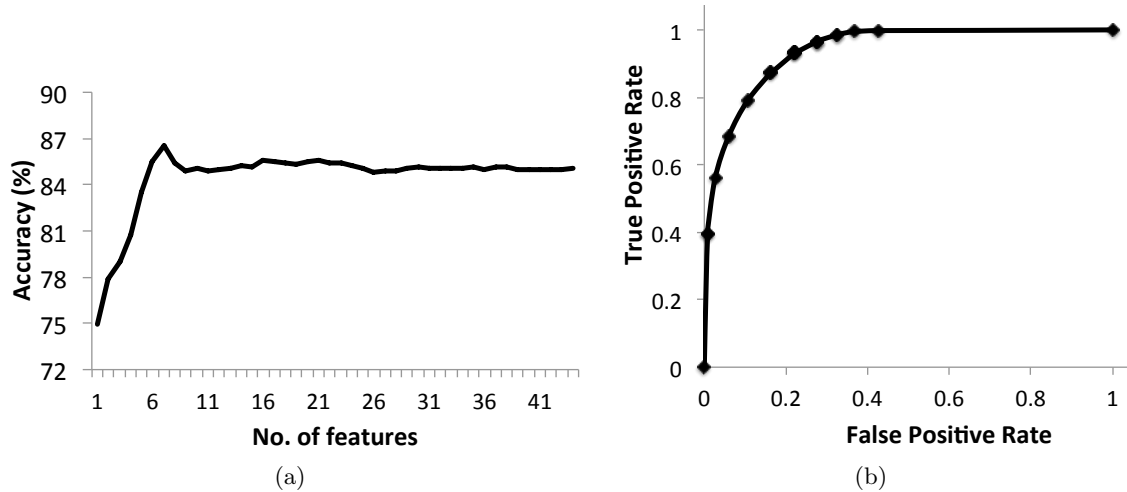


Figure 3.8: (a) Accuracy values of the Random Forest Classifier trained on Dataset I for 1 through 44 features. Accuracy peaked to 86.62% at top 7 features. (b) Receiver operating characteristic (ROC) curve for the Random Forest Classifier trained on Dataset I using all 44 features. The Area Under Curve (AUC) was 0.935.

posts generated *during news-making events*. To confirm our hypothesis, we collected a separate, non event-specific dataset of Facebook posts across April 2013 – March, 2015. This dataset was collected by querying the Graph API using *http* as a search keyword and consisted of 267,517 posts. We used the same methodology to find malicious posts from this dataset, as we did for Dataset I (Section 3.2.2) and obtained 5,446 posts containing malicious URLs. We refer to this dataset as **Dataset III**. We compared these 5,446 malicious posts from Dataset III with the 11,217 malicious posts from Dataset I and found some drastic differences.

**Textual content and URLs:** We found multiple spam campaigns in Dataset III. Majority of these campaigns advertised adult content, followed by malicious app downloads, online money making scams, and politics. Unlike Dataset I, none of the campaigns in Dataset III involved any celebrities or events. Figure 3.9 shows tag clouds of the most frequently occurring terms in the *message* field of posts in Dataset I and Dataset III. As evident from the tag clouds, there was minimal overlap between malicious content posted during news-making events and malicious content posted in general. We also observed that a large proportion of malicious posts in Dataset III (38.49%) comprised of a *facebook.com* URL. Although the *facebook.com* domain never appeared in a URL blacklist, a large proportion of malicious posts in Dataset III comprised of a *facebook.com* URL in addition to a blacklisted URL in the same post. This behavior affected the “no. of URLs per word” feature in our feature set. This feature was thus found to be the fourth most important feature during our supervised learning experiments.



(a) Most frequently occurring terms in Dataset I.



(b) Most frequently occurring terms in Dataset III.

Figure 3.9: Tag clouds of top 75 most frequently occurring terms in Dataset I and Dataset III. The size of the word is proportional to its frequency. Stop words and URLs have been removed.

**Entities posting malicious content:** Observations from Dataset I revealed that about 21% of all posts containing a malicious URL originated from Facebook pages. We observed that this percentage rose sharply to 65.86% in Dataset III.<sup>15</sup> This rise increased the importance of the *pageLikes* feature, which turned out to be the third most important feature in our supervised learning model trained on Dataset III. In contrast, this feature was amongst the least important features in Dataset I.

**Metadata:** We observed that less than 1% (53 out of 5,446) malicious posts in Dataset III were generated from third party applications. This was a considerable drop as compared to Dataset I, where 6% of all malicious posts were generated from third party applications. Malicious posts originating from the web also increased sharply from about 70% in Dataset I to 87.49% in Dataset III. Additionally, 36.83% of all malicious posts in Dataset III were photos. This percentage dropped to a bare 2.51% in case of Dataset I.

The above differences provided initial proof that malicious posts *in general* are fairly different from malicious posts generated *during news-making events*. To consolidate and quantify support for this initial proof, we performed further experiments using supervised learning. We picked a data sample comprising of the 5,446 malicious posts and a random sample of 5,446 legitimate posts drawn from the aforementioned non event-specific dataset (Dataset III) of 267,517 posts containing URLs. Using this data sample, we trained a Random Forest Classifier with the same feature set as described in Table 3.6. A 10-fold cross validation on this model yielded an accuracy of 90.39%. However, when this model was tested on Dataset I (event-specific dataset comprising

<sup>15</sup>Overall, only 24.74% of all posts in Dataset III originated from pages.

11,217 malicious and 11,217 legitimate posts), the accuracy plummeted to 55.64%. This drop in performance disproved our null hypothesis and confirmed that a model trained to detect malicious posts *in general* is not capable of detecting event-specific malicious posts efficiently. Malicious posts *generated during news-making events* need to be addressed separately, which we attempt to achieve in this work.

### Performance over time

To check the effectiveness of our technique over time, we trained 3 models (M1, M2, and M3) using all 44 features by splitting our dataset into 3 equal-sized subsets across time (M1 trained on D1 = April - July, 2013; M2 trained on D2 = August - December, 2013; M3 trained on D3 = January - July, 2014). We also collected test data about the Ebola outbreak in Africa during August - October, 2014, consisting of 3,248 malicious and 3,248 randomly picked legitimate posts (D4). Each model was evaluated a) using 10-fold cross validation, and b) by testing on all data subsets from future time intervals. For example, M1 was tested on D2, D3, and D4; M2 was tested on D3, and D4, etc. Figure 3.10 represents the performance of all models over time. True positive rates obtained from 10-fold cross validation on all models were consistently high, and varied between 88.6% and 94.3%. However, we noticed a gradual overall decrease in the true positive rates of all models over time (except M1). Model M1 (trained on April - July, 2013 data) showed a considerable rise in performance when tested on D4 (Ebola dataset).

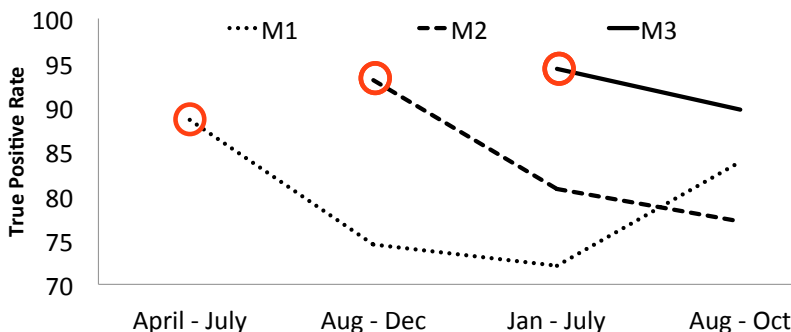


Figure 3.10: True positive rates of all models generated from Dataset I over time. Encircled points indicate true positive rates on 10-fold cross validation. M1 reached the lowest true positive rate of 72.1% over Jan - July, 2014 test data.

Drop in true positive rates possibly indicate attackers' changing strategies over time. However, to maintain high true positive rates, the model can easily be retrained by collecting data using the methodology described in Section 3.2.2.

## Performance comparison with previous models

Gao et al. [57] used a clustering approach to identify spam campaigns and used features from these clusters to train a supervised machine learning model. In contrast, our technique treats each post separately and also takes individual malicious posts (which are not part of a campaign) into consideration. Since Gao et al.’s system was designed to detect spam messages which were part of a campaign, authors could not estimate the amount of malicious posts that this approach missed (false negatives). To this end, we applied the same clustering technique and threshold values used by Gao et al. [58] on our dataset to get an estimate of the false negatives of their approach. Since our entire dataset was already labeled (as opposed to Gao et al.’s dataset), we did not apply clustering on our entire dataset to find malicious posts. Instead, we applied clustering only on malicious posts in our dataset, and compared how many of those clusters met the *distributed* and *bursty* threshold values previously used ( $>5$  users per cluster, and  $<90$  minutes median time between consecutive posts respectively). Applying clustering on our 11,217 malicious posts yielded a total of 4,306 clusters. Out of these, only 183 clusters (containing 4,294 posts) met the *distributed* and *bursty* thresholds, yielding a high false negative rate of 61.7%. These results indicate that existing clustering techniques may not be the best approach to detect malicious posts which are not part of a campaign. Our supervised learning models eliminate dependency on message and time proximity, and makes our technique more robust and efficient at detecting malicious posts irrespective of whether they are part of a campaign or not.

Note that the dataset used by Gao et al. was collected by crawling the Facebook network (which is no longer permitted by Facebook) and captured features like users’ social degree and interaction history. Due to the unavailability of Gao et al.’s dataset, and absence of social degree and interaction history features in our dataset, we were not able to do an ideal comparison of our detection techniques. Also, while Gao et al.’s system was designed to be deployed at the OSN service provider side, our approach uses completely public features and is deployed at the client side. We believe that the performance of our model will increase if it is deployed at the OSN service provider side and supplied with more user and post metadata which is not available publicly.

We were unable to compare our model accuracy results with other previous work due to two major reasons; a) absence of features like *likes*, *comments*, *message similarity* etc. at zero-hour (used by Rahman et al.), and b) public unavailability of features like number of friends, message sent, friend choice, active friends, page likes etc. (used by Stringhini et al. [145] and Ahmed et al. [5]). We believe that the addition of features like *likes*, *comments*, and *shares* to our feature set is likely to increase the accuracy of our model. However, as discussed earlier, using these features hampers the real-time aspect of our detection technique.

## Crisis versus Non-crisis events

Extensive research has been conducted to highlight the use and importance of OSNs during crisis events [2, 107, 118, 133, 138]. This prompted us to analyze crisis and non-crisis events in our dataset separately, and perform experiments to see if our feature set could be used more effectively for identifying malicious content during crisis events specifically. Table 3.9 shows the events in our dataset which we marked as *crisis*<sup>16</sup> and non-crisis.

Table 3.9: Crisis and non-crisis events in our dataset. We used the Oxford Dictionary definition of *crisis* to mark events as crisis or non-crisis.

Crisis (9)	Air Algeria missing plane, Boston Blasts, Cyclone Phailin, Unrest in Gaza, Malaysian Airlines Flight MH17 crash, Metro North Train Derailment, Typhoon Haiyan, London Terror Attacks, Washington Navyyard shooting
Non-crisis (8)	FIFA Worldcup 2014, Heartbleed bug in OpenSSL, IPL 2013, IPL 2014, Death of Nelson Mandela, Birth of the Royal baby, T20 cricket world cup, Wimbledon 2014

We split Dataset I into two components – comprising of posts related to a) crisis, and b) non-crisis events. We used Random Forests to train models for each of these components separately using the same 44 features, and performed 10-fold cross validation for evaluation. The experiment produced an accuracy of 83.79% (ROC area under curve = 0.921) for crisis events, and an accuracy of 85.93% (ROC area under curve = 0.947) for non-crisis events. These results indicated that although marginally, it was harder to identify malicious posts during crisis events as compared to non-crisis events.

### 3.5.2 Dataset II

Similar to our machine learning analysis for Dataset I, we performed extensive analysis on Dataset II using multiple supervised learning techniques. For Dataset II, we generated our training set using 571 malicious posts and 571 legitimate posts randomly drawn from the 3,841 legitimate posts we obtained during the annotation process. Similar to Dataset I, we performed 10-fold cross validation on the training set for evaluation. In addition, each experiment was performed ten times, each time using a random sample drawn from the legitimate class, and keeping the malicious class constant. The results were averaged out across all ten observations. Standard deviation in accuracy across all ten observations was less than 4% for all experiments. Table 3.10 shows the average performance we achieved for all our experiments.

Similar to Dataset I, The Random Forest Classifier performed slightly better than the other clas-

<sup>16</sup>We marked events as crisis based on the following definition of *crisis* from the Oxford English Dictionary: "A time of intense difficulty or danger"

Table 3.10: Ten-fold accuracy scores averaged across ten experiments for each classifier. Although SVM classifier gave best accuracy, Random Forest Classifier had better recall and ROC AUC.

Feature Set	Metric	Entity	Text	Metadata	Link	All
Naive Bayesian	Accuracy	51.67	51.60	72.45	77.58	67.63
	Precision	50.74	52.57	98.64	85.39	71.52
	Recall	80.28	8.28	45.81	66.71	76.46
	ROC_AUC	0.530	0.738	0.803	0.835	0.728
Decision Tree	Accuracy	51.66	73.16	79.01	81.04	76.17
	Precision	51.51	70.36	86.85	88.41	75.67
	Recall	48.62	80.52	68.60	71.47	76.88
	ROC_AUC	0.525	0.678	0.785	0.833	0.763
Random Forest	Accuracy	52.86	76.56	79.87	81.49	<b>80.56</b>
	Precision	53.20	77.27	86.99	89.05	<b>87.59</b>
	Recall	51.38	74.92	70.19	72.56	<b>70.76</b>
	ROC_AUC	0.537	0.806	0.832	0.856	<b>0.873</b>
SVM (rbf)	Accuracy	53.16	76.52	78.18	80.37	73.79
	Precision	52.48	81.50	83.57	89.29	88.49
	Recall	63.89	68.84	70.42	69.20	54.95
	ROC_AUC	0.544	0.780	0.836	0.845	0.810

sifiers and achieved an accuracy of 80.56% with an area under the ROC curve (ROC AUC) of 0.873. Similar to our observations from Dataset I, we found multiple weak differentiators in the two classes present in Dataset II as well. These weak differentiators were the most likely reason why the Random Forest classifier outperformed the other classifiers in this case too. Although the SVM classifier yielded a higher precision than Random Forest Classifier, a better recall and higher ROC AUC value prompted us to use the Random Forest Classifier for further experiments. Link based features performed better than a combination of all features for Dataset II. This observation was unlike our observation from Dataset I, where a combination of all features performed better than Entity, Text, Metadata or Link features individually.

We studied the effect of the number of features on the model by adding features one by one and observing the 10-fold cross validation accuracy of the Random Forest Classifier trained on Dataset II. The accuracy remained fairly consistent throughout, and peaked at all 44 features. Table 3.11 shows the top 10 features in Dataset II. Interestingly, even though Link features performed the best, no Link features appeared in the top 10 features. This implies that while individual Link features might be weak, a combination of these features was the strongest combination for distinguishing malicious posts from legitimate ones. The top 10 features also conformed to some of our observations from Section 3.4, where we found “message length” (captured by no. of characters in text message, no. of sentences, etc.) and “application used to post” to be discriminating features. Further, as compared to Dataset I, we observed a heavy shift from Metadata and Link features towards Text

features among the most important features in Dataset II.

Table 3.11: Source and feature importance value (normalized) of the top 10 features in Dataset II.

Feature	Source	Importance
Average word length	Text	0.0771
No. of characters in text message	Text	0.0693
Length of <i>link</i> field	Metadata	0.0640
Average sentence length	Text	0.0620
No. of sentences	Text	0.0456
Application used to post	Metadata	0.0452
Length of name	Entity	0.0437
Length of username	Entity	0.0419
No. of upper case characters in text message	Text	0.0371
No. of URLs per word in text message	Text	0.0363

### Crisis versus Non-crisis events

Similar to the experiments we performed in Section 3.5.1, we split Dataset II into crisis and non-crisis event related posts and repeated the experiments for this dataset. A 10-fold cross validation on a model trained using Random Forests and the same feature set resulted in an accuracy and ROC AUC value of 76.34% and 0.841 respectively for crisis events. In case of non-crisis events, these values increased to 84.06% and 0.922 respectively. These results were in line with the results obtained during experiments performed on Dataset I, and reflected that it was harder to detect malicious posts during crisis events as compared to non-crisis events.

### 3.5.3 Dataset I versus Dataset II

Observations from our two ground truth datasets revealed a series of differences in the characteristics of malicious content present in both these datasets. To quantify these differences, we performed cross validation of the models obtained from Dataset I and Dataset II separately. To this end, we tested the performance of the Random Forest Classifier model obtained from Dataset I, on Dataset II. Similarly, we tested the performance of the Random Forest Classifier model obtained from Dataset II, on Dataset I. We performed a total of 100 experiments (using the 10 equally balanced randomly drawn training datasets from Dataset I, tested on the 10 equally balanced randomly drawn testing datasets from Dataset II, and vice versa) and averaged out the results of all experiments. Performance of both the models dropped significantly when tested across each other. The Random Forest Classifier model trained on Dataset I achieved an accuracy of 39.42% when tested on Dataset II. The Random Forest Classifier model trained on Dataset II was only



43.03% accurate when tested on Dataset I. These observations confirmed the intrinsic differences in the two types of malicious content we found in the two ground truth datasets, and highlighted that a single model is not enough to solve the problem of real-time malicious content detection on Facebook during news-making events.

We thus propose a two-fold approach wherein each post goes through two separate supervised learning models to decide whether it is malicious or not. The post is marked as *malicious* if either of the two models classify the post as malicious. Since the two models are independent of each other, both models can work in parallel, thus minimizing time delay in the final decision and maintaining real-time efficiency.

### 3.6 Discussion, Limitations, and Future work

The dataset of Facebook posts we collected is specific to news-making events. We would like to emphasize that the aim of our research is to identify malicious posts generated during news-making events. Several efforts have aimed to detect malicious content on Facebook in the past [58, 128]. However, none of them specifically addressed content specific to news-making events. Our results highlighted that malicious content generated during news-making events differs from malicious content in general. We also showed that models for detecting malicious content in general do not perform well on event-specific data (Section 3.5.1).

We do not claim that our dataset is representative of the entire Facebook population. Facebook does not provide any information about what fraction of public posts is returned by Graph API search. However, to the best of our knowledge, our dataset of 4.4 million public posts and 3.3 million users is the biggest dataset in literature, collected purely using Facebook APIs. The only bigger dataset of Facebook posts used in research in the past was collected by Gao et al. in 2010 [58]. This dataset was collected by scraping Facebook users' walls in 2010, and consists of 187 million wall posts from 3.5 million users. However, this dataset is not available publicly. There exist some other anonymized datasets consisting of Facebook's network structure [20, 106, 112, 154, 161], and Facebook usernames [137], but we were not able to find any free datasets of user-generated Facebook content (posts) in the public domain.

We understand that the WOT ratings that we used to create Dataset I are obtained through crowd sourcing, and may suffer biases. However, WOT states that in order to keep ratings more reliable, the system tracks each user's rating behavior before deciding how much it trusts the user. In addition, the meritocratic nature of WOT makes it far more difficult for spammers to abuse.

Preliminary experiments with crisis and non-crisis event related malicious posts separately revealed that it was harder to identify malicious posts during crisis events as compared to non-crisis events. In future, we would like to explore these results more, and attempt to better understand the reasons

behind these results.

As a logical next step, we used our findings to devise a framework for automatic identification of poor-quality posts on Facebook in real time. This framework is called Facebook Inspector (FbI) and we discuss it in detail in the next chapter.

## Chapter 4

# Facebook Inspector: Implementation and Evaluation

This chapter is partly a reproduction of a paper published at Social Network Analysis and Mining (SNAM) Journal 2017 [37].

To provide an easy-to-use, real-world solution for the problem of detecting malicious content on Facebook, we deployed Facebook Inspector in the form of a REST API (Representational State Transfer Application Programming Interface) and a browser plug-in. In this chapter, we outline the implementation, deployment, and evaluation details of Facebook Inspector, an automatic real-time framework and browser plug-in for identifying malicious posts on Facebook. We utilized both models we obtained (from Dataset I and Dataset II described in Chapter 3) in the finally deployed version on Facebook Inspector.

### 4.1 Implementation

The implementation of Facebook Inspector includes a back-end (pre-trained models) and a front-end (browser plug-in) which communicate over HTTP RESTful APIs. Figure 4.1 shows the basic architecture and flow of the system.

#### 4.1.1 Back-end

The flow of data and information in Facebook Inspector is as follows: A user opens `https://www.facebook.com` in a browser with Facebook Inspector installed and enabled, and logs into her Facebook account. Once all the posts have loaded on the page, the browser plug-in parses the

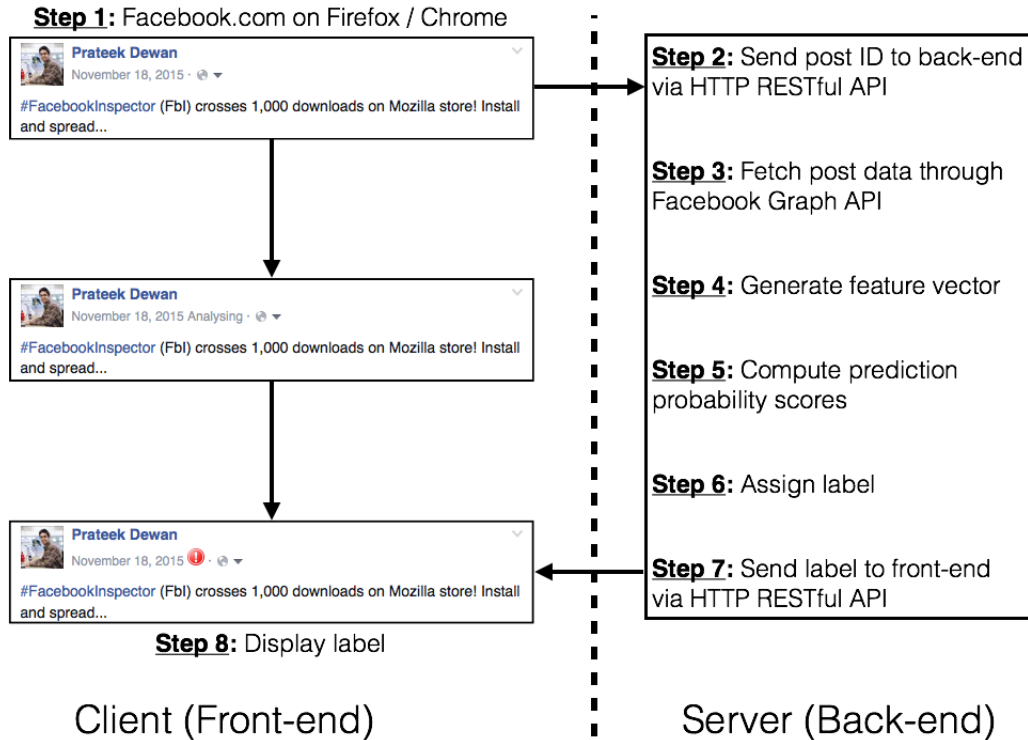


Figure 4.1: Architecture / flow diagram for Facebook Inspector. For each post, the plug-in sends the post ID to the back-end via a HTTP request to the API. While the post is being processed at the back end, the plug-in displays the text “Analysing” next to the post on the user’s Facebook page. Once processed, this text disappears, and a “red alert marker” is displayed next to the post, if the post is deemed malicious.

page and extracts the post IDs of all posts displayed on the page. These post IDs are then sent by the plug-in to our back-end server using REST protocol. This communication takes place in parallel, i.e. communication for each post is independent of the others. We do not scrape the post or user information from the raw HTML source of the webpage; we merely pass the post IDs to our back-end server. This is done for two reasons: a) the webpage does not provide enough information about the post as required by our model. We make authenticated requests to the Graph API, which provides more information about the post than what is available on the webpage; and b) to maintain user privacy. While querying the Graph API, we are only able to access *public* posts. This restricts the system to get access to any private content, which users do not intend to share publicly.

Once the back-end server receives a post ID, it makes an authenticated request to Facebook’s Graph API to fetch data about the post. This data is used to generate a feature vector for the post, and two prediction probability scores are computed for the post, corresponding to the two pre-trained models obtained from Dataset I and Dataset II (Section 3.5). We use the *predict\_proba* method in Python Sklearn’s RandomForestClassifier module (since the Random Forest Classifier

achieved highest accuracy (Table 3.8, Table 3.10)) to predict the class probability for the post instead of obtaining a class label directly. This is done to provide flexibility to external third-party applications using only the API, and to enable them to choose their own thresholds (depending on the use case) while making a decision on whether a post is malicious or not. The system thus returns two scores corresponding to each of the two models separately, where each score (between 0.0 and 1.0) captures the *probability* of the post being malicious. A score greater than 0.5 signifies that the model deems the post as malicious. In addition to probability scores, the system also computes a *label* (Malicious / Benign) and a *confidence* level (high / low) in this label, for the front-end (browser plug-in). Table 4.1 shows the *label* and *confidence* level corresponding to the classification probability scores produced by Model I (trained on Dataset I) and Model 2 (trained on Dataset II). The final result (probability scores, label, and confidence level) is output by the back-end in JSON format along with the original Facebook post ID, and sent to the front-end.

Table 4.1: Label and Confidence level corresponding to probability scores produced by Model I and Model II.

Model I score	Model II score	Label	Confidence
Above 0.7	Above 0.7	Malicious	High
Mean score above 0.5		Malicious	Low
Below 0.3	Below 0.3	Benign	High
Mean score below 0.5		Benign	Low

Facebook Inspector’s REST API is publicly accessible and can be queried by sending a HTTP GET request containing a Facebook post ID (*fid*) and *version* parameter (example: <http://multiosn.iiitd.edu.in/fbapi/endpoint/?version=2.0&fid=369318403277259>). We used Python’s webpy framework to implement the API. <sup>1</sup> All feature extraction and computation scripts were written in *Python*, with *MongoDB* as the database. The hardware used for the implementation was a mid-range server with running Ubuntu (server edition) with 64 Gigabytes of RAM, Intel Xeon CPU E5-2640 0 @ 2.50GHz with 8 physical cores, and hyperthreading enabled. Due to Facebook’s API limitations, our API currently works only for public posts which are accessible through Facebook’s Graph API. <sup>2</sup>

Open access to the Facebook Inspector API ensures reproducibility and enables researchers to test our models on new data, and build upon our results. The API also enables other researchers to easily compare their results with ours, which has been otherwise non-trivial in the past.

<sup>1</sup><http://webpy.org/>

<sup>2</sup><https://developers.facebook.com/docs/graph-api/reference/v2.5/post>

### 4.1.2 Front-end

Google Chrome and Mozilla Firefox are two of the most famous browsers, and account for over 87% of the total market share.<sup>3</sup> These two browsers were thus, the ideal target for the first version of our light-weight browser plug-in. In order to minimize computation load on the web browser, all heavy computations were offloaded to the back-end server. The task of the plug-in was therefore, limited to making a GET request (to our back-end server), and injecting a small image indicator next to *malicious* posts. As a result, the browser plug-in had a minimalistic memory and CPU footprint. This design ensured that the system does not result in any performance bottleneck on the client's web browser.

Facebook Inspector displays a small indicator image next to the post, if the post is adjudged as malicious (as shown in Figure 4.2). Hovering over the indicator image shows the confidence level in the judgement. Users seeking more details about how the judgement was made, can click on the image, which takes them to a web page<sup>4</sup> with details about the entire framework. For posts that are adjudged as benign, there is no indication. In addition, since the plug-in analyzes only public posts, no indicator image is present for posts that are not public.



Figure 4.2: Sample Facebook post marked as malicious by Facebook Inspector. The small red image next to the time of post indicates that the post may be malicious. Hovering over the red image gives additional details to the user.

<sup>3</sup>As of May, 2016. Source: [http://www.w3schools.com/browsers/browsers\\_stats.asp](http://www.w3schools.com/browsers/browsers_stats.asp)

<sup>4</sup><http://precog.iiitd.edu.in/osm.html#fbi>

Table 4.2: Summary statistics for the usage of Facebook Inspector.

Date of first release of Facebook Inspector	Aug. 23, 2015
Total incoming requests	2,765,211
Total posts processed	974,426
Total unique browsers (user-agent strings)	602
Total downloads	2,500+
Total daily active users	150+
Posts marked as malicious	188,650
Posts marked as benign	785,776
Average processing time per post	2.635 seconds

Facebook Inspector is currently available for download publicly on Google Chrome store <sup>5</sup> and Mozilla store. <sup>6</sup> As of May 2016, Facebook Inspector had been downloaded over 2,500 times, and had over 150 daily active users on average. This information helped us evaluate Facebook Inspector in terms of response time, performance, and usability. We now present the findings of our evaluation.

## 4.2 Evaluation

During the first nine months of its deployment (August 2015 - May 2016), Facebook Inspector received over 2.76 million incoming requests. Out of these requests, the system processed and evaluated 0.97 million unique public posts. The remaining requests made were for private posts that were not publicly accessible via Facebook Graph API. Table 4.2 shows the summary statistics of the usage of Facebook Inspector since its date of first release.

### 4.2.1 Response Time

We analyzed the response time of the browser plug-in, measured as the time elapsed between the moment our back-end server received a request, and the moment the server returned a response for this request. This technique eliminated unwanted network delays induced because of variable Internet speeds across different Internet Service Providers and browser versions. Figure 4.3 shows the CDF of response times for approximately 374,000 post analyzed by the latest version of the Facebook Inspector framework. We did not capture the response time information for posts during the initial few weeks of the deployment, until the framework stabilized and some initial bugs were fixed.

<sup>5</sup><https://chrome.google.com/webstore/detail/facebook-inspector/jlhjfkmlldnokgkhhbghbnmiejokohmlfc>

<sup>6</sup><https://addons.mozilla.org/en-US/firefox/addon/fbi-facebook-inspector/>

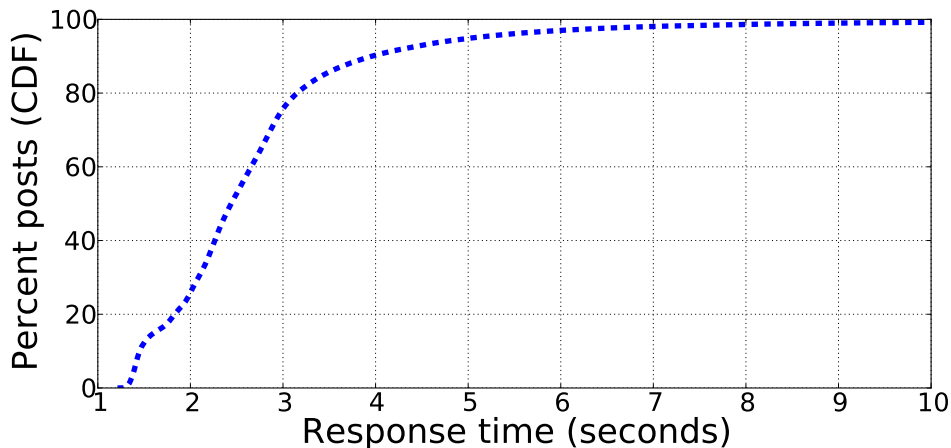


Figure 4.3: CDF of the response time (in seconds) of Facebook Inspector. The response time was 3 seconds or lesser for approximately 80% of all posts. For over 99.3% of the posts, the response time was under 10 seconds.

The average response time for analyzing a post was 2.635 seconds, with a lowest response time of 1.212 seconds. As visible from the figure, the response time was under 3 seconds for approximately 80% of all posts, and under 10 seconds for 99.3% of all the posts. We further broke down the total response time into the various steps (fetching data from the Graph API, feature vector creation, classification, etc.) that comprise the complete processing sequence of a post. This analysis helped us to better understand the entire process in terms of time consumed, and identify potential bottlenecks and scope for further improvement (if any) in terms of response time. We now discuss the results of this analysis.

## 4.2.2 Performance

We recorded and analyzed the end-to-end response time of Facebook Inspector as a combination of the time consumed by five main components of the framework; a) classification (model I), b) classification (model II), c) Graph API call (post data), d) Graph API call (user data), and e) URL resolution. All other processing steps like database entries, variable assignments, etc. accounted for less than 0.8% of the total time. Table 4.3 shows the mean and median time period (in seconds) of these components.

As evident from Table 4.3, URL resolution was the most time consuming step, followed by GET requests made to the Facebook Graph API for fetching user and post data. Classification on pre-trained models took approximately 1/10th of a second each, and was least time consuming. Although a total mean response time of 2.635 seconds makes Facebook Inspector faster than some of the existing real-time malicious content detection systems [66], we observed further scope of



Table 4.3: Breakdown of time consumed by the main components of Facebook Inspector.

Step	Mean time (seconds)	Median time (seconds)
Classification – Model I	0.108	0.106
Classification – Model II	0.106	0.106
GET request to Graph API – post data	0.755	0.633
GET request to Graph API – user data	0.636	0.592
URL resolution	1.005	0.811

improvement by introducing parallel processing for the Graph API calls and classification steps. Parallely executing Graph API calls and classification steps further improves the response time by 0.742 seconds on average, reducing the total response time to 1.893 seconds.

### 4.2.3 Usability

The usability of any computer system is an important aspect of its overall evaluation. Low usability not only hampers user experience, but can also drive users away from the system. It is therefore desirable for researchers to test the usability of any new system they propose, and work towards having an easy-to-use system for end users.

To assess the overall usability and usefulness of Facebook Inspector, we conducted an online survey among its users. The survey contained the 10 standard questions of the *System Usability Scale* (SUS) [17]. We recruited participants through word-of-mouth, mailing lists, and Facebook groups to participate in the usability study. The participants were offered a small cash prize as an incentive for the time they spent for the study. Participants were asked to download and install the Facebook Inspector plug-in, use it, and fill the survey. A total of 53 participants completed the study. The entire process took approximately 12-15 minutes per participant.

We obtained an overall SUS score of 81.36 for Facebook Inspector, which translates to grade A on the SUS scale, indicating that Facebook Inspector had higher perceived usability than over 90% of all products tested using the *SUS*.<sup>7</sup> Almost all participants (98.1%) found Facebook Inspector easy to use (agree / strongly agree), and none of the participants (0%) found the system to be unnecessarily complex. Also, 67.9% of the participants said that they would like to use this system frequently (agree / strongly agree). However in terms of consistency, only 66% of the participants found the system to be consistent (agree / strongly agree).

In addition to the usability survey, we also collected qualitative feedback for the plug-in. We asked a group of 36 undergraduate students studying computer science to download and use the plug-in, and report their findings in terms of the usefulness of the plug-in. Most of the student participants

<sup>7</sup><https://measuringu.com/sus.php>

found the plug-in useful and easy to use. Upon being asked about their initial reaction on the plug-in, one of the participants said, *“Saves your time spent on spam links and hence enhances user experience.”* Another participant spoke about the utility of the plug-in, and how she found it to be a means to mitigate spam; *“A very interesting tool to curb the problem of users clicking random posts and spreading spam-like behavior.”* Some participants mentioned that they would have liked the plug-in to analyze private posts as well. Participants also commented on the response time of the tool. One participant mentioned, *“[Facebook Inspector] Has good response time. Definitely useful tool but false positive rate can be reduced.”* The number of false positives generated by the plug-in was a recurring concern among some other participants too. One participant remarked, *“The false positives were high and the product can be made more useful by incorporating user feedback which will further lead to decrease in false positives.”* This finding was in line with the consistency score we obtained for the plug-in during the usability evaluation. Participants expressed that some of the posts in their timeline were marked as malicious even though they were published by a *verified* page. For instance, one of the participants said, *“Pages like Humans of New York and The Logical Indian are very reputed, verified and legit accounts who post useful and interesting stuff. A lot of their posts were marked as malicious.”* This happened because Facebook Inspector does not mark a post as benign just because its publisher is *verified* by Facebook, but looks at a combination of features to decide whether a post is malicious or not (refer Table 3.6 for the list of features). The 2013 Wall Street crash based on a hoax tweet by a verified Associated Press account is a fitting example for why it is not sufficient to trust content based on its source alone.<sup>8</sup> Moreover, the framework is trained to work on post related to news-making events. General posts not relating to an event are expected to produce unpredictable results in some cases.

Some participants also expressed the need for a separate label (probably green in color) for indicating benign posts. One participant stated that the plug-in could be of particular help for minors; *“[Facebook Inspector] Can be useful for minors and people who lack the judgement to decide how the post is.”* While most participants found some use for the plug-in, one participant differed in opinion, and stated, *“Generally, my feed consists of friends activities and articles of my interest, usually from verified sources. So, it really didnt help in that case.”*

### 4.3 Discussion, Limitations and Future Work

We propose and evaluate Facebook Inspector (FbI), a freely available real-world solution for automatic real-time detection of malicious posts on Facebook. Facebook Inspector is a completely client-side system, fully functional using only publicly available data. Our measurement study revealed substantial presence of malicious content which evades Facebook’s existing immune system. Based on empirical findings, Facebook Inspector works on pre-trained models capturing character-

---

<sup>8</sup><http://www.bbc.com/news/world-us-canada-21508660>

istic differences between malicious and legitimate posts, and computes class labels using a two-fold approach based on two diverse ground truth datasets. Facebook Inspector distinguishes from existing malicious content detection techniques in that it does not rely on message similarity features, which have been largely used in the past to detect campaigns. Facebook Inspector detects malicious posts in real-time without depending on any engagement metrics associated with a post (*likes*, *comments*, or *shares*). Performance and user evaluation suggests that Facebook Inspector is not only fast, but also easy to understand and use for its end users.

The current architecture of Facebook Inspector is restricted to public Facebook posts only. In future, we would like to add support for analysing private posts, such that users can opt to grant access to their private content to Facebook Inspector for analysis. This would not only expand the scope of Facebook Inspector, but would also help users to identify which users in their network share poor quality content, and be extra careful while consuming content coming from such users in the future. This technique can thus be further extended to associate a “content quality quotient” with users based on the content they generate.

## Chapter 5

# Hiding in Plain Sight: The Anatomy of Malicious *Pages* on Facebook

This chapter is partly a reproduction of a short paper published at the IEEE/ACM Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2016 [33] and a book chapter accepted for publication at Springer Lecture Notes in Social Networks 2017 [34].

### 5.1 Introduction

With the advent of OSNs and Web 2.0, the scope of what is deemed as “malicious” on the Internet has evolved. Facebook, for example, has established community standards to safeguard users against nudity, hate speech, etc. [51], and considers any pages, groups or events that confuse, mislead, surprise or defraud people, as abusive [47]. In a recent study, we discovered the presence of a similar set of malicious Facebook *pages* accounting for over 30% of malicious posts in our dataset, and have not been studied in detail in the past [36]. Security experts and news sources have also acknowledged the presence of malicious *pages* on Facebook, set up intentionally to spread fraudulent claims, scams, and other types of malicious content. A group of scammers, for example, set up a fake British Airways page, offering free flights to customers for a year (Figure 5.1). The *page* asked users to *share* a photo, *like* the *page* and leave a *comment* to claim their free flights.<sup>1</sup> In another similar incident, an international football player’s name was used as bait to set up a Facebook page, and users were asked to sign a fake petition.<sup>2</sup>

---

<sup>1</sup><https://grahamcluley.com/2015/09/british-airways-isnt-giving-away-free-flights-year-facebook-scam/>

<sup>2</sup><http://www.marca.com/2014/07/18/en/football/barcelona/1405709402.html>

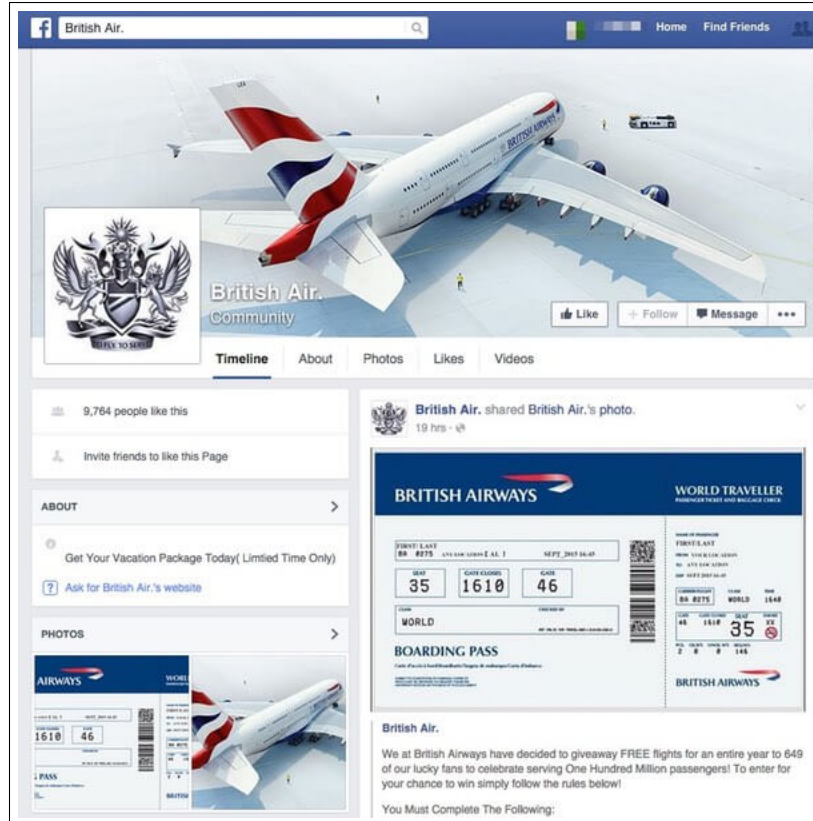


Figure 5.1: Fake British Air Facebook page offering free flights for a year in return for liking, commenting on, and sharing their post.

In addition to scams and fake information, researchers have also identified and studied the spread of rumors (which are a class of untrustworthy information) on Facebook [56]. In case of events like earthquakes, rumors on OSNs have been observed to contribute to chaos and insecurity in the local population [107]. Facebook also faced criticism because of presence of fake news and polarized politics on the platform during the 2016 presidential elections in the USA [63]. Such instances highlight a new and emerging class of malicious content, which is much harder to identify using automated means, and hasn't been widely explored in literature. It is thus, crucial to identify and control the spread of untrustworthy and fake information, and minimize adverse real-world impact.

One of the distinct characteristics that makes Facebook *pages* important to study is their reach and audience. Unlike a typical Facebook user profile, Facebook *pages* can have millions of subscribers (followers) which translates to a much bigger audience for the content generated by *pages*. Scammers can use news-making events to gather large audience for their pages by exploiting the content of events (for example, support communities for people affected by a terror attack or a natural calamity), and bombard users with unwanted, malicious promotions, and potentially dodgy links

that could lead to a malware infection or users being phished. It has been claimed that Facebook pages spam is a \$200 million business.<sup>3</sup>

In this chapter, we identify and characterize a set of 627 Facebook *pages* that published one or more malicious URLs in their most recent 100 posts. We focus our analysis on *pages* which spread untrustworthy information, hate speech, nudity, misleading claims, etc., that are deemed as malicious by the community standards [51] and “Page Spam” definitions [47] established by Facebook. We use our labeled dataset to train and evaluate multiple supervised learning models to automate the process of identifying malicious Facebook *pages*. We extract a total of 96 features from *page* information, and posts published by the *pages*. Further, we train and evaluate supervised learning models using a bag-of-words obtained using the textual content published by these *pages*. Our broad findings are as follows:

- **Politically polarized malicious entities:** We identified and manually verified the presence of numerous politically polarized entities, which dominated our dataset of malicious *pages*, and published URLs from untrustworthy web domains.
- **Malicious *pages* were more active:** We found that malicious *pages* were more active (in terms of posting) than benign *pages*; the number of malicious *pages* that were active daily was over three times the number of benign *pages* that were active daily.
- **Malicious *pages* showed collusive behavior:** We found presence of collusive behavior within malicious *pages* in our dataset; malicious *pages* engaged in promoting (*liking*, *commenting on*, and *sharing*) each others’ content.
- **Malicious and benign *pages* had similar temporal behavior:** We performed a longitudinal study over a period of one year, by capturing daily snapshots of malicious and benign *pages* in our dataset, and found minimal statistically significant different between the two types of *pages* in terms of temporal behavior.
- **Artificial neural networks outperformed all other algorithms:** Artificial neural networks trained on a bag-of-words outperformed all other supervised learning algorithms for automatic detection of malicious *pages*, achieving an area under the ROC curve value of 0.9. Grid search experiments help improve the performance further, attaining a maximum ROC AUC value of 0.931.

The rest of the chapter is structured as follows. Section 5.2 gives the background and scope of our research, and explains the data collection process. Characterization and analysis of malicious *pages* make up Section 5.3. We present the results of our supervised learning experiments in Section 5.4.

---

<sup>3</sup><http://mashable.com/2013/08/29/facebook-fan-pages-spam-200-million-business/>

Section 5.5 discusses the implications and limitations of our results. We also conclude and discuss the future directions of our work in this Section.

## 5.2 Scope and data collection

Facebook (unlike Twitter, Instagram, etc.) poses a restriction on the number of connections a user can have (max. 5,000 friends), and provides *pages* to enable large following for celebrities, groups, businesses, etc. A Facebook *page* can have multiple administrators managing the *page* under the same name, without the audience knowing. This allows *pages* to have a higher degree of interaction with its audience and keeping it more active as compared to a single user profile. Facebook *pages* are an important and integral part of the Facebook ecosystem, that offer a free platform for promotion of businesses, brands and organizations.<sup>4</sup> From an attacker’s perspective, Facebook *pages* are potentially lucrative tools to gather large audiences and target all of them at once. Our past research has shown greater participation of Facebook *pages* in posting malicious URLs as compared to posting benign URLs [36]. Such inflated malicious activity and reach of Facebook *pages* make them a vital aspect to study in detail.

### 5.2.1 Scope

The definition and scope of what should be labeled as “malicious content” on the Internet has been constantly evolving since the birth of the Internet. Researchers have been studying malicious content in the form of spam and phishing for over two decades. With respect to Online Social Networks, state-of-the-art techniques have become efficient in automatically detecting spam campaigns [58, 174], and phishing [4] without human involvement. However, new classes of malicious content pertaining to appropriateness, authenticity, trustworthiness, and credibility of content have emerged in the recent past. Some researchers have studied these classes of malicious content on OSNs and shown their implications in the real world [19, 65, 69, 107]. All of these studies, however, resorted to human expertise to identify untrustworthy and inappropriate content and establish ground truth, due to the absence of efficient automated techniques to identify such content. We aim to study a similar class of malicious content pertaining to trustworthiness and appropriateness in this work, which currently requires human expertise to identify. In particular, we look at Facebook *pages* that generate content deemed as malicious by Facebook’s community standards and definitions of “Page Spam”. Facebook defines “Page Spam” as *pages* that *confuse, mislead, surprise or defraud people* [47]. Additionally, *pages* that are misleading, deceptive, or otherwise misrepresent the prize or any other aspect of promotion are considered as “Page Spam”. Facebook has also established community standards to protect users against nudity, hate speech, violence

---

<sup>4</sup><https://www.facebook.com/help/174987089221178>

and graphic content, fraud, sexual violence etc. [51].

### 5.2.2 Establishing ground truth

Given the complex definition of malicious content for the scope of our study, there exist no accurate detectors for establishing ground truth. Detectors such as URL blacklists (Google Safebrowsing, PhishTank, SURBL, SpamHaus, etc.) used for identifying malicious content in the past, are restricted to identifying classical threats like phishing, malware, etc. In order to obtain ground truth for the malicious content we aim to study, we resorted to a crowd sourced approach. Crowdsourcing techniques have been shown to perform well for establishing ground truth for complex and subjective aspects of OSN content such as credibility [19, 65]. For our study, we used the Web of Trust (WOT) API [167]. WOT leverages crowdsourcing to collect ratings and reviews from millions of users who rate and comment about websites, based on their personal experiences. This crowdsourced, community based mechanism enables WOT to protect users against threats that only the human eye can spot such as scams, unreliable web stores, misleading websites, nudity, and questionable content, which largely overlaps with Facebook definitions of spam. To the best of our knowledge, WOT is one of the only services which covers the broader definition of malicious content that is required for our study.

**Are WOT ratings biased?** We understand that WOT ratings are obtained through crowd sourcing, and may seem to suffer from biases. However, WOT states that in order to keep ratings more reliable, the system tracks each user’s rating behavior before deciding how much it trusts the user. In addition, the meritocratic nature of WOT makes it far more difficult for spammers to abuse. This approach is similar to other crowdsourcing services like Amazon’s Mechanical Turk<sup>5</sup> and CrowdFlower<sup>6</sup>, which have been widely used in OSN research in the past (as discussed above). To further increase the confidence in the ratings, we used conservative thresholds for confidence values associated with the reputation scores. We discuss these thresholds in more detail in Section 5.2.3.

### 5.2.3 Dataset

We collected an initial dataset of 4.4 million public posts published by 390,246 unique *pages* and 2,983,707 unique users on Facebook between April 2013 and July 2014, using Facebook’s post search API. These posts were collected by using event related search keywords belonging to 17 real-world events that took place in the aforementioned time frame (e.g. Boston Marathon Blasts, Death of Nelson Mandela, FIFA World Cup, Birth of the first Royal Baby, Gaza unrest, etc.). We queried the WOT API for domain reputations of all URLs present in the 4.4 million posts, and identified 10,341 posts containing one or more malicious URLs (we will discuss the exact definition of “malicious”

---

<sup>5</sup><http://mturk.com/>

<sup>6</sup><http://crowdfower.com/>



Table 5.1: Category labels and descriptions returned by WOT API. Source: WOT API Wiki (<https://www.mywot.com/wiki/API>).

Category	Description
Negative	Malware, viruses, poor customer experience, phishing, scam, potentially illegal, adult content
Questionable	Misleading claims, unethical, privacy risks, suspicious, hate, discrimination, spam, potentially unwanted programs, ads, pop-ups, incidental nudity, gruesome / shocking

later in the section). These 10,341 posts containing malicious URLs originated from 1,557 *pages* and 5,868 users. The complete details of this dataset can be found in our prior work [36].

To capture a more recent view of the 1,557 *pages* posting malicious URLs, we re-queried the Graph API and collected their *page* information in August 2015. We also collected 100 most recent posts (or all posts, whichever was smaller) published by these *pages* using the */page-id/posts* endpoint of the Graph API <sup>7</sup> along with all *likes*, *comments*, and *shares* on these posts. We then looked up the WOT API for all URL domains present in the most recent 100 posts, and found that 627 *pages* published one or more malicious URLs. This exercise of rescanning the 1,557 *pages* eliminated those *pages* which had not shown malicious activity in the recent past (recent 100 posts), and could be deemed as non-malicious for our study. For the rest of the chapter, we use the remaining 627 *pages* as our dataset of malicious *pages*.

According to its documentation <sup>8</sup>, the WOT API returns a reputation score for a given domain. Reputations are measured for domains in several *components*, for example, trustworthiness. For each {domain, component} pair, the system computes two values: a *reputation* estimate and the *confidence* in the reputation. Together, these indicate the amount of trust in the domain in the given component. A *reputation* estimate of below 60 indicates *unsatisfactory*. The WOT browser add-on requires a confidence value of  $\geq 10$  before it presents a warning about a website. We tested the domain of each URL in our dataset for *Trustworthiness* and *Child Safety* components. For our analysis, a URL was marked as malicious if both the aforementioned conditions were satisfied (*reputation* $<60$ ; *confidence* $\geq 10$ ). In addition to reputations, the WOT rating system also computes categories for websites based on votes from users and third parties. We marked a URL as malicious if it fell under the *Negative* or *Questionable* category (Table 5.1). We used the same approach in the past to develop techniques for automatic identification of individual malicious posts on Facebook [36].

We also drew an equal random sample of 1,557 *pages* from the benign *pages* in our dataset of 4.4 million posts, which had not posted any malicious URLs during our initial data collection phase

<sup>7</sup><https://developers.facebook.com/docs/graph-api/reference/page/feed>

<sup>8</sup><https://www.mywot.com/wiki/API>

(April 2013 - July 2014). Similar to our approach for identifying malicious *pages*, we re-queried the Graph API and collected the *page* information along with the most recent 100 posts (including their *likes*, *comments*, and *shares*) published by these *pages*. We found 1,278 *pages* which published no malicious URLs in their most recent 100 posts. These 1,278 *pages* made up our dataset of benign *pages*. Table 5.2 shows the descriptive statistics of all Facebook *pages* in our dataset.

Table 5.2: Descriptive statistics of our dataset of Facebook *pages*. Numbers in the parentheses indicate verified *pages*.

Category	Malicious	Benign
No. of <i>pages</i>	627 (31)	1,278 (49)
Recent 100 posts	60,306	120,184
Recent 100 posts with URLs	55,233	92,980
Likes (recent 100 posts) <sup>9</sup>	3,447,669	31,680,263
Comments (recent 100 posts)	354,502	1,245,959
Shares (recent 100 posts)	507,964	1,012,151

Figure 5.2 shows the distribution of the number of posts published, and number of page *likes* gathered by malicious and benign pages in our dataset. We observed that malicious pages published significantly more posts than benign pages ( $\mu_{malicious} = 15,091.14$  posts per page,  $\mu_{benign} = 3,072.74$  posts per page; Mann Whitney U statistic = 230474.0,  $p\text{-value} < 0.01$ ). However, benign pages gathered more *likes* than malicious pages ( $\mu_{malicious} = 64,330.59$  *likes* per page,  $\mu_{benign} = 112,250.36$  *likes* per page; Mann Whitney U statistic = 314109.5,  $p\text{-value} < 0.01$ ).

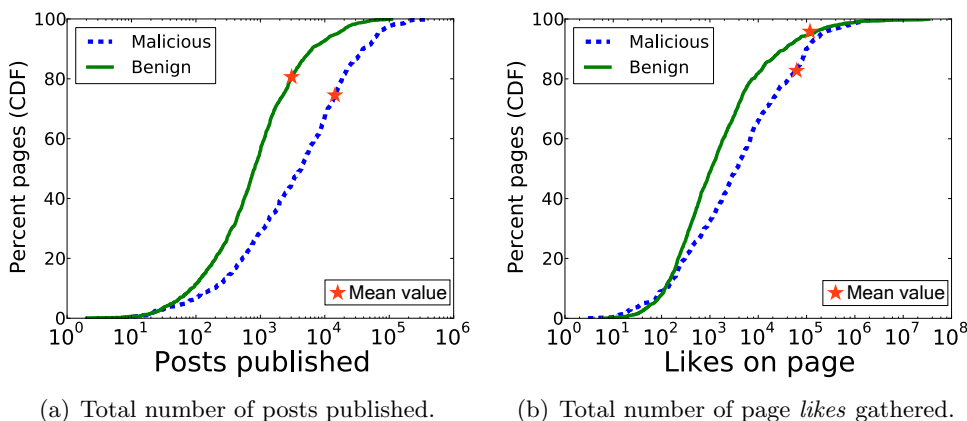


Figure 5.2: Distribution of the total number of posts published, and number of page *likes* gathered by pages in our dataset. Malicious pages published more posts than benign pages.

Table 5.3 provides a detailed description of the number of posts and *pages* along with their WOT

<sup>9</sup>Due to API rate limitations, we had to restrict our data collection to 50,000 *likes* per post. We had 2 malicious and 291 benign posts which exceeded this limit.

categories in our dataset of 627 malicious *pages*. We found a total of 20,999 posts which contained one or more malicious URLs. These posts engaged a total of 675,162 unique users who *liked*, *commented*, or *shared* these posts. Interestingly, we found that spam and phishing (two of the most common types of malicious content studied in literature) were least common in our dataset. Child unsafe content was the most common, followed by untrustworthy content.

Table 5.3: Number of malicious posts and *pages* in each category in our dataset. Number of *pages* posting phishing and spam URLs was the lowest.

WOT Response	No. of <i>pages</i>	No. of posts
Child unsafe	387	10,891
Untrustworthy	317	8,057
Questionable	312	8,859
Negative	266	5,863
Adult content	162	3,290
Spam	124	4,985
Phishing	39	495
Total	627	20,999

**Is the dataset size significant?** We understand that our sample size of 627 malicious *pages* is not a large dataset as compared to some of the other studies done on OSNs in the past. However, gathering Facebook data is a challenging task now. To the best of our knowledge, our dataset of 4.4 million public Facebook posts (from which we identified 627 malicious *pages*) is one of the biggest samples of Facebook data studied in literature. The only dataset of Facebook posts larger than ours was collected by Gao et al. [58]. This dataset was gathered in 2009 by performing large scale crawls on 8 regional Facebook networks over 3 months. Authors gathered 187 million posts which originated from roughly 3.5 million users (almost equal to the 3.3 million users + *pages* in our dataset). In contrast, we gathered all our data through authenticated requests made to the Graph API over a much larger time span of 16 months. All other studies on Facebook data have used much smaller datasets [5, 128, 145].

### 5.3 Malicious pages on Facebook

To understand the differences (and similarities) between malicious and benign *pages*, we studied both the spatial and temporal behavior of these *pages*. We present our findings in detail in this section.

### 5.3.1 Spatial behavior

Most OSNs can be divided into three basic components that make up the social network; *the entity* (user / page), *the content* it posts, and its *network* (friends / followers / subscribers). We study all these three components separately.

#### Entities

We performed term-frequency analysis on unigrams, bigrams, and trigrams obtained from *page* names in our dataset to identify the most prominent entities generating malicious and benign content. Table 5.4 lists the top 30 unigrams appearing in page names in our dataset. Manual analysis revealed dominant presence of politically polarized entities and religious groups with keywords like *american, british, english, league, patriot, defense, etc.* in malicious *pages*. Bigram and trigram analysis confirmed wide presence of entities like *British National Party (BNP), The Tea Party, English Defense League, American Defense League, American conservatives, Geert Wilders supporters,* etc. We also found some malicious *pages* dedicated to pop bands (One Direction), radio channels (Kiss FM), *pages* using *anonymous* in their names, etc. We manually inspected all the aforementioned *pages* and validated that the *page* names were aligned with the content they published, and were not misleading.

Table 5.4: Word frequency of the top 30 terms appearing in page names in our dataset. We found substantial presence of politically polarized entities among malicious pages.

Malicious page names				Benign page names			
Keyword	#	Keyword	#	Keyword	#	Keyword	#
news	11	group	5	church	20	county	8
league	11	one	5	center	15	one	8
defense	10	world	5	llc	14	services	8
online	8	videos	5	love	14	fans	8
american	8	national	5	photography	14	south	8
party	8	cricket	5	inc.	13	national	8
english	8	new	4	news	12	life	7
free	7	network	4	united	12	get	7
media	7	bnp	4	school	11	arts	7
truth	7	division	4	team	11	confessions	7
british	6	says	4	community	10	world	7
direction	6	club	4	club	9	health	7
edl	6	tea	4	st.	9	fire	7
forum	5	patriot	4	page	9	dr.	6
radio	5	united	4	cricket	9	beauty	6

Facebook has been shown to play a significant role in the political context, especially during elec-

tions [18, 63, 162]. Researchers have found that knowledge gained by youngsters from Facebook about electoral candidates influenced their evaluation of the candidates [42]. Such impactful role of Facebook on the users prompted us to study and understand the sentiment and emotion of content generated by politically polarized entities in our dataset.

Using the bigram and trigram analysis, we divided pages belonging to politically polarized entities into four broad groups based on page names to help us study them better. These groups were i) America (9 pages), containing pages with “america” or “american” in their page name, ii) British National Party (7 pages), containing pages mentioning British National Party or BNP in their page name, iii) Conservative (6 pages), containing pages with the term “conservative” in the page name, and iv) Defence League (11 pages), containing pages using the phrase “defence league” in the page name. We manually verified each page to ensure that they fit in the group they were assigned. To maintain anonymity, we do not reveal the exact page names. We performed linguistic analysis on the content published by these 4 categories of pages separately using LIWC2007 [124]. LIWC is a text analysis software to assess emotional, cognitive, and structural components of text samples using a psychometrically validated internal dictionary. It determines the rate at which certain cognitions and emotions (for example, personal concerns like religion, death, and positive or negative emotions) are present in the text. LIWC has been widely used in the past to study social media content related to politics [144, 155, 156].

We focused our analysis on 12 dimensions in order to profile the sentiment of content published by these groups of pages: Positive emotion, negative emotion, anxiety, anger, sadness, money, religion, death, sexuality, past orientation, future orientation, and swear words. We chose these 12 dimensions since they seemed most relevant among over 25 available dimensions, in the context of content published online. Figure 5.3 shows the results of our analysis. We found high degree of anger in content from all categories. We also observed that only one category of pages (British National Party) had more positive emotions than negative emotions. The Defence League pages had much higher negative emotions as compared to positive emotions, followed by America pages. Conservative pages were almost equal in terms of positive and negative emotions. These findings contradicted prior results where researchers found that positive emotions outweighed negative emotions by 2 to 1 for profiles of all German political candidates [156]. However, it is important to note that the content studied in [156] was general content, and not a poor quality subset of the content as in our case. We also found substantial presence of content related to religion. These observations are indicative of the kind of influence that politically polarized pages in our dataset can have on their audience.

Benign *page* names were found to represent a variety of categories and interests like *photography*, *school*, *love*, *news*, *confessions*, etc. Bigram and trigram analysis revealed presence of a set of *methodist church pages*. We also found some overlap between malicious and benign *page* names, for example, *One Direction fan pages*, and radio channel *pages*. Unlike malicious *pages*, we did not

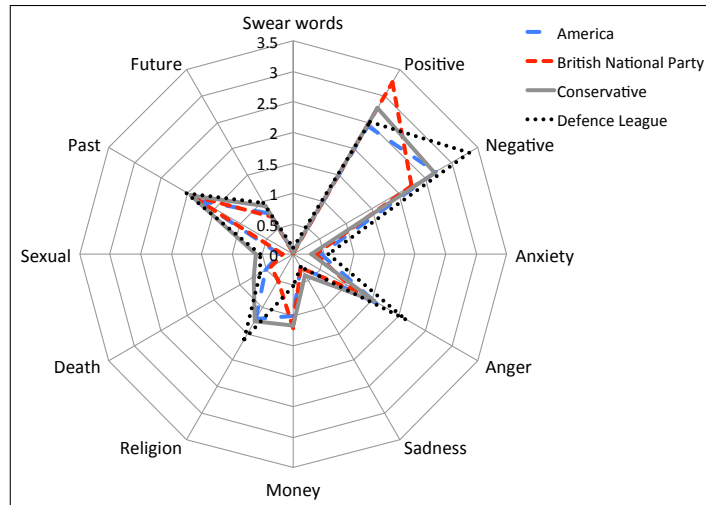


Figure 5.3: Linguistic analysis of content produced by politically polarized groups of pages in our dataset. We found high presence of negative emotion, anger, and religion related content.

find any fixed category dominating in benign *pages*.

From the above findings, it is evident that politically polarized entities that exist in the real world, also have a strong online presence. These results can be used to identify such entities on other social networks as well, and control (if not eliminate) the spread of polar political views online.

## Content

Scanning the most recent 100 posts (Section 5.2.3) revealed that almost half of the *pages* (49.28%) in our dataset published 10 or less posts containing a malicious URL. Overall, the median number of domains shared by these *pages* was 24.5. On the contrary, the median number of domains shared by the other half of the *pages* posting more than 10 posts containing a malicious URL (50.72%) was 5. Figure 5.4 shows the distribution of the number of malicious posts versus the total number of domains shared by all malicious *pages* in our dataset. We found a weak declining trend in the number of domains as the number of malicious posts increased ( $r = -0.223$ ,  $p\text{-value} < 0.01$ ).

This declining trend (and negative correlation) indicated that *pages* posting a large number of malicious URLs tend to do so from a small subset of domains. In fact, 84 *pages* in our dataset (13.39%) shared URLs from only 1 domain. Out of these 84 pages, 53 *pages* (8.45%) published more than 90 posts containing a malicious URL, gathering *likes* and *comments* from 55,171 and 31,390 distinct users respectively. Most certainly, such *pages* exist for the sole purpose of promoting a single (malicious) domain, and are successful in engaging thousands of Facebook users. This sort of activity closely resembles social spam campaigns, which have been studied by multiple researchers

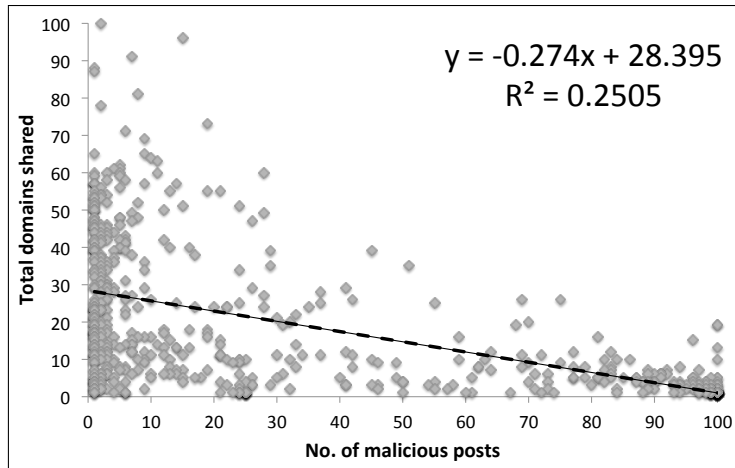


Figure 5.4: Number of malicious posts versus all domains published by all 627 *pages* in our dataset. We observed a weak declining trend in the total number of domains when the number of malicious posts published by a *page* increased. Three outliers (sharing 1,257, 202 and 140 domains) have been removed in this graph.

in the past [58, 174]. However, since most past research has focused on more obvious threats like unsolicited and targeted spam, advertising, and bulk messaging, other types of malicious content concerned with trustworthiness and child safety has largely remained unaddressed.

Note that there also exist multiple legitimate *pages* on Facebook dedicated to promote a single domain, for example, FIFA World Cup *page* (exclusively posting `fifa.com` URLs), BBC News *page* (exclusively posting `bbc.com` URLs), etc. We found 118 *pages* in our benign dataset (9.23%) which were dedicated to promote a particular domain. Such behavior cannot therefore be associated exclusively with malicious activity. Malicious *pages* seem to take advantage of this fact and continue their activity, hiding in plain sight. However, the vocabulary used in the content published by these *pages* can be used to differentiate between the malicious and benign classes using a bag-of-words. We explore this possibility, and report our findings in Section 5.4.2.

**Top domains:** Table 5.5 lists the 10 most frequently occurring domains in our dataset of malicious pages, along with their WOT classification, Facebook audience, and Alexa world ranking.<sup>10</sup> For each domain, we calculated the number of posts the domain appeared in, the sum of *likes*, *comments*, and *shares* on all these posts, the number of pages the domain appeared in, and the sum of *likes* on all these pages. It was interesting to observe that 3 out of the top 10 domains were very famous, and were ranked within the top 3,000 domains worldwide on the Alexa ranking. Two of these domains were reported for being unsafe for children and spreading adult content. Although the Internet does not restrict the creation and promotion of adult and child unsafe content, most OSNs including Facebook have established community standards which restrict the display of adult and explicit

<sup>10</sup><http://www.alexa.com/>

content [51]. All of the other domains had low Alexa ranking worldwide. Only 3 of the top 10 domains were marked as spam, and none of the domains in the top 10 were reported for phishing or malware, signifying that untrustworthy and child unsafe content is much more prominent on Facebook than traditional forms of malicious content like spam and phishing.

Table 5.5: Top 10 malicious domains in our dataset with their Web of Trust classification, Facebook audience, and Alexa world rank.

Domain	WOT class, categories	Posts	Likes — comments — shares	Pages	Page likes	Alexa rank
ammboi.com	Untrustworthy, suspicious, spam, privacy risks	456	666 — 61 — 195	5	109,012	352,191
ridichegratis.com	Untrustworthy	424	428 — 14 — 252	21	2,650,802	-
blesk.cz	Child unsafe, adult content	402	3,674 — 2,103 — 1,494	8	864,554	2,924
says.com	Child unsafe	386	387 — 15 — 62	5	97,784	27,684
ghanafilla.net	Untrustworthy, scam, spam, suspicious	296	192 — 8 — 6	3	54,246	1,360,634
9cric.com	Child unsafe	281	1,189 — 121 — 177	13	193,348	923,243
perezhillton.com	Child unsafe, adult content	274	26,088 — 3,516 — 1,128	8	1,701,834	2,192
nairaland.com	Untrustworthy, misleading claims or unethical	201	238 — 89 — 31	3	116,131	1,329
pulsd.com	Untrustworthy, child unsafe	199	2 — 0 — 0	2	19,020	247,480
970wfla.com	Spam	194	700 — 448 — 280	2	22,486	277,467

The number of posts and pages associated with each of the top 10 domains revealed that there existed multiple Facebook pages dedicated to promoting all of these domains. We observed that all of the top 10 domains appeared in 2 or more pages, and two of the domains appeared in over 10 pages (ridichegratis.com in 21 pages; 9cric.com in 13 pages). At least 4 of the top 10 domains (ammboi.com, ghanafilla.net, pulsd.com, and 970wfla.com) had two or more Facebook pages each (3 for ghanafilla.net, 5 for ammboi.com), heavily promoting their respective domains (over 90 out of the 100 posts containing the domain, per page). Pages set up for these domains also had a substantial audience, with 6 out of the 10 domains collectively having over 100,000 *likes* on their pages. Two of the top 10 domains had over 1 million *likes* (collectively) on pages promoting them. The collective number of *likes*, *comments*, and *shares* on posts was however, considerably low as compared to collective *likes* on the pages. Only 3 out of the top 10 domains managed 1,000 or more *likes* on the posts associated with them. This indicated that while malicious domains in our dataset were successful in gathering a substantial audience in the form of page *likes*, they were



not as successful in engaging the audience with their content. We also observed that 2 of the 3 domains with high Alexa rank (blesk.cz and perezhilton.com) also had high number of page *likes* and high number of *likes*, *comments*, and *shares* on posts. This signified that domains which were popular (more visited) on the Internet were also more famous on Facebook. This observation can be utilized to better control the spread of poor quality content on Facebook by monitoring and focusing on popular untrustworthy websites on the Internet in general, more than websites that are not too popular.

## Network

Past research has shown that decentralized networks are prone to *sybil attacks*, wherein malicious entities tend to collude together and attempt to infiltrate the legitimate part of the network [41]. Such attacks have also been studied in context of OSNs [169]. To investigate if such phenomenon exists for Facebook *pages* too, we analyzed the *like*, *comment*, and *share* networks for both malicious and benign *pages* in our dataset. Facebook does not provide an API endpoint to gather the list of users who have *liked* (subscribed to) a page. However, it is possible to collect the list of users who have *liked*, *commented* on, or *shared* posts published by a page. As described in Section 5.2.3, we collected all *likes*, *comments*, and *shares* on the most recent 100 posts of all *pages* in our dataset, and analyzed the inter and intra-*page* networks. In particular, we analyzed networks consisting of *pages* and users *liking*, *commenting on*, or *sharing* posts from two or more *pages* in our dataset (malicious and benign separately) (inter-*page* networks), and networks of *pages liking*, *commenting on*, or *sharing* posts from *pages* within our dataset (malicious and benign separately) (intra-*page* networks). To keep the network size comparable, we averaged out the results for 10 random samples of 627 benign *pages* each (same size as malicious *pages* dataset) drawn from the full 1,278 benign *pages* dataset.

Table 5.6 shows the details of the network analysis. We found that the Inter-*page likes* network for benign *pages* (83,799 nodes) was much larger and stronger (avg. weighted degree: 41.695) than the Inter-*page likes* network for malicious *pages* (21,947 nodes, avg. weighted degree: 24.273), indicating that a larger number of users *liked* two or more benign *pages* as compared to the number of users who *liked* two or more malicious *pages* in our dataset. More interestingly, we found stronger ties (avg. weighted degree for Intra-*page* networks) within malicious *pages* in all aspects (*likes*, *comments*, and *shares*) as compared to benign *pages*, indicating collusion and sybil behavior within malicious *pages*. We also found a much larger number of communities in all Inter-*page* networks for benign *pages* as compared to Inter-*page* networks for malicious *pages*, indicating a larger and more diverse audience for benign *pages* as compared to malicious *pages*.

Stronger ties within malicious *pages* prompted us to further investigate the communities we detected from Intra-*page likes*, *comments*, and *shares* networks. Figure 5.6 shows the network graphs of the

detected communities. We observed that post *sharing* was the most prominent intra-*page* activity, followed by *liking* and *commenting*. The network graphs also revealed a distinct community of six Facebook *pages* completely connected to each other in terms of *likes* (Figure 5.6(a)) and *shares* (Figure 5.6(c)). Five out of these six *pages* also formed a community in the intra-page comments graph (Fig 5.6(b)). We manually inspected and observed that all *pages* in this community belonged to adult stars and promoted pornographic content. This behavior closely resembled a sybil network, and indicated that all these *pages* were controlled by / belong to the same real-world entity (person or organization). We also found multiple two-*page* communities involving politically polarized *pages*, where one *page* heavily engaged in *liking*, *commenting on*, and *sharing* the other page’s content.

## Metadata

Analyzing the metadata of posts in our dataset revealed some significant differences in the type of content published by malicious and benign pages. Figure 5.5 shows the distribution of the content type of posts published by all pages in our dataset. We observed that more than half of the content published by benign pages were photos and videos (50.16%). This percentage went down to 32.42% for malicious pages. The metadata also revealed that over half of the posts published by malicious pages were links (54.69%), where as less than a quarter of all posts published by benign pages were links (24.45%). These numbers indicate that malicious pages are inclined towards posting links, and directing user traffic to external websites. On the other hand, benign pages tend to post more pictures, which can be consumed by users without leaving the OSN. In addition to content types, we looked at the status types of posts and found that benign pages published almost double the amount of content through mobile devices (23.80%) as compared to malicious pages (12.33%).

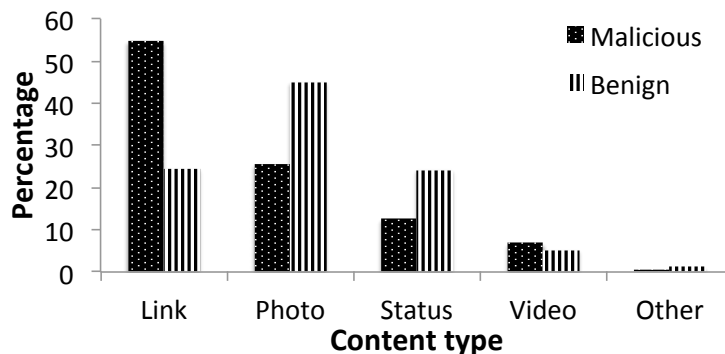


Figure 5.5: Types of content published by malicious and benign pages in our dataset. Malicious pages published more links, while benign pages published more pictures.

All pages on Facebook have a *category* associated with them, for example *Community*, *Company*, *Personal Website*, etc. This category is assigned to the page by the page administrator(s) at the

time of page creation, according to the person / organization represented by the page, and content that the page generates. To see if any subset of categories was more popular among a particular class of pages (malicious or benign), we compared category ranks and found strong correlation between category ranks across malicious and benign pages (Spearman’s  $\rho = 0.67$ ,  $p\text{-value} < 0.01$ ). This indicated that the distribution of malicious and benign pages across various categories was fairly similar, and that categories more popular among malicious pages were also more popular among benign pages. We also compared the page *likes* and page mentions (*talking\_about\_count* field) of malicious and benign pages, and did not find any significant differences.

Table 5.6: Network analysis of *likes*, *comments* and *shares* networks within and between *pages* in our dataset. We observed that malicious *pages* had stronger intra-network ties as compared to benign *pages*.

Network type	Total nodes	Total edges	Avg. weighted degree	Density	No. of communities
Malicious (All 627 <i>pages</i> )					
Inter- <i>page likes</i> network	21,947	103,683	24.273	0	18
Inter- <i>page comments</i> network	3,901	13,957	11.255	0.001	19
Inter- <i>page shares</i> network	14,318	67,513	15.796	0	14
Intra- <i>page likes</i> network	27	35	8.333	0.05	9
Intra- <i>page comments</i> network	9	9	1.667	0.125	3
Intra- <i>page shares</i> network	68	65	6.309	0.014	21
Benign (Results averaged across 10 random samples of 627 benign <i>pages</i> each)					
Inter- <i>page likes</i> network	83,799	390,854	41.695	0	3070
Inter- <i>page comments</i> network	2,958	7,722	8.919	0.001	142
Inter- <i>page shares</i> network	3,406	10,234	9.920	0.001	30
Intra- <i>page likes</i> network	4.3	3.6	0.408	0.075	0.7
Intra- <i>page comments</i> network	0	0	0	0	0
Intra- <i>page shares</i> network	7.8	6.9	1.168	0.072	1.1

These observations indicated that apart from the type and source of published content, there were no significant differences in the meta information between malicious and benign pages in our dataset. Metrics like popularity (*likes*) and user mentions (*talking\_about\_count*) associated with OSN entities can be used to identify spammers, since they capture the notion of influence of entities in the network [22]. However, similarities in such metrics across malicious and benign pages can aid malicious pages to continue operating regularly and go undetected for long periods of time, hiding in plain sight.

### 5.3.2 Temporal behavior

We explored the temporal posting activity of all *pages* in our dataset to determine how active the *pages* were, in terms of publishing posts. We also monitored the status of these pages daily, for a period of over one year to observe any changes in the *pages*’ behavior and attributes over time.

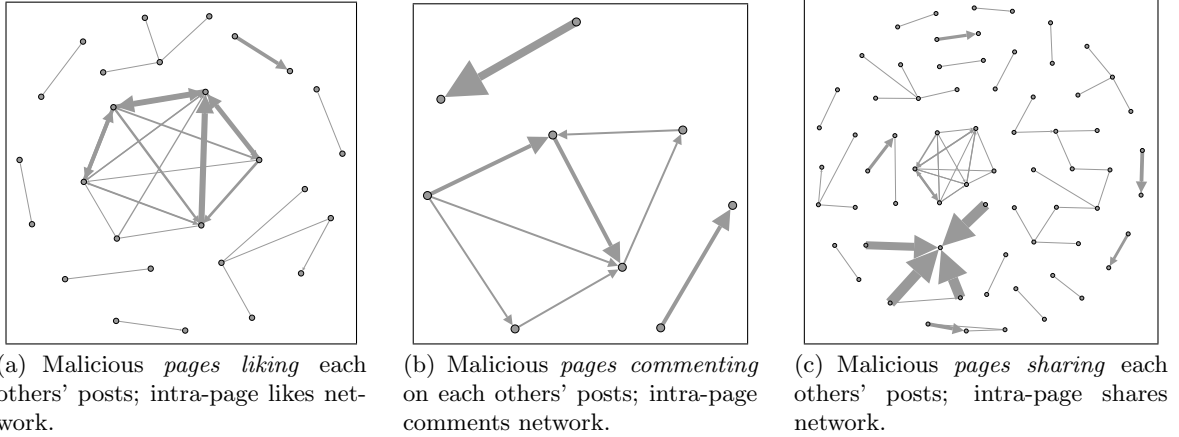


Figure 5.6: Network graphs capturing intra-*page* activity of malicious *pages* in our dataset. We found multiple two-node communities and a few bigger communities.

## Posting Activity

To be able to quantitatively compare the activity of malicious and benign *pages*, we calculated a *daily activity ratio* for each page, defined by the ratio of number of days a *page* was active (published one or more posts) versus the total number of days between the first and hundredth post by the page.

$$\text{daily activity ratio} = \frac{\text{no. of days active}}{\text{no. of days between first and last post}}$$

Figure 5.7(b) shows the *daily activity ratio* plots of all malicious and benign *pages* in our dataset. We observed that 27.43% of all malicious *pages* were active daily as compared to only 8.60% daily active benign *pages*. On average, malicious *pages* were 1.4 times more active daily as compared to benign *pages* in our dataset. We also calculated activity ratio in terms on number of hours and number of weeks, and observed similar results. All activity ratio values were compared using Mann-Whitney U test and the differences were found to be statistically significant ( $p\text{-value} < 0.01$  for all experiments) [104]. These difference confirmed that malicious *pages* in our dataset were more active as compared to benign *pages*, and published more frequently.

## Attributes over time

We studied the temporal behavior of all *pages* in our dataset over the period of an year, between October 2015 and October 2016. During this period, we captured daily snapshots of the *page* information for all the *pages* through the Graph API. The aim of this study was to observe the change in attributes of malicious *pages* over time, and to identify if these changes were significantly

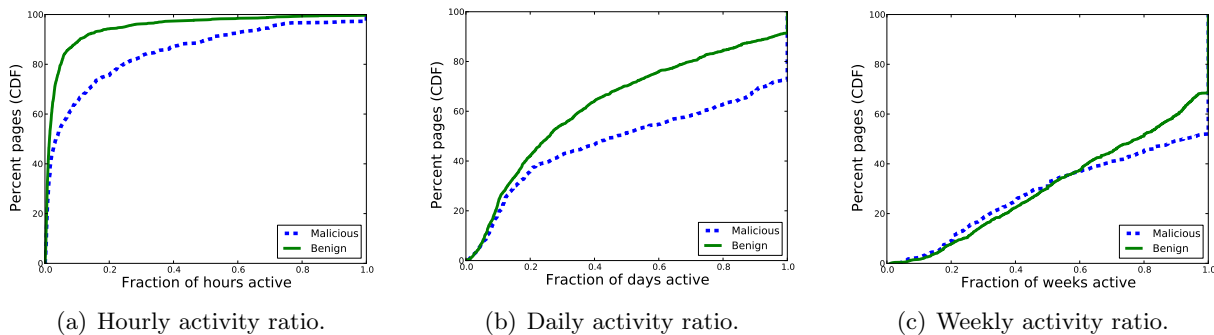


Figure 5.7: Daily, hourly, and weekly temporal activity of *pages* in our dataset. We found that malicious *pages* were more active than benign *pages*.

different from the changes in attributes for benign *pages*. In particular, we studied changes in two types of attributes over time; a) popularity, and b) description.

**Popularity** To study the change in popularity over time, we computed a *gain factor* corresponding to the change in the number of *likes* on all *pages* in our dataset as follows:

$$gainFactor_P = \frac{likesOnDayLast_P - likesOnDayOne_P}{likesOnDayOne_P} \times 100$$

where  $likesOnDayLast_P$  = no. of *likes* on *page P* on the last day (October 15, 2016), and  $likesOnDayOne_P$  = no. of *likes* on *page P* on the first day (October 16, 2015) of the study. A positive value of the *gain factor* for a *page P* indicates an increase in the number of *likes*, while a negative value depicts a drop in the number of *likes* for *P* over the span of the one year time frame under consideration.

Figure 5.8 shows the *gain factor* between malicious and benign *pages* for all *pages* in our dataset. We observed that a larger proportion of malicious *pages* (28.54%) lost *likes* as compared to benign *pages* (20.26%). Contrarily, while computing the average *gain* over all pages, we found that malicious *pages* had a larger *gain factor* (32.52%) as compared to benign pages (24.03%). This difference, however, was statistically insignificant ( $p\text{-value} > 0.1$ ). Prior statistics show that the average growth rate for a Facebook *page* is 0.64% per week (approximately 33.28% per year) [87]. Interestingly, this number is much closer to malicious *pages* in our dataset. However, given the statistical insignificance of our results, we cannot conclude that the growth rate of malicious *pages* is more similar to an average Facebook *page* as compared to benign pages in our dataset.

We investigated the change in popularity over time in more detail, by computing the rate of change of the number of *likes* per day for all *pages* in our dataset, to see if there was a statistically significant

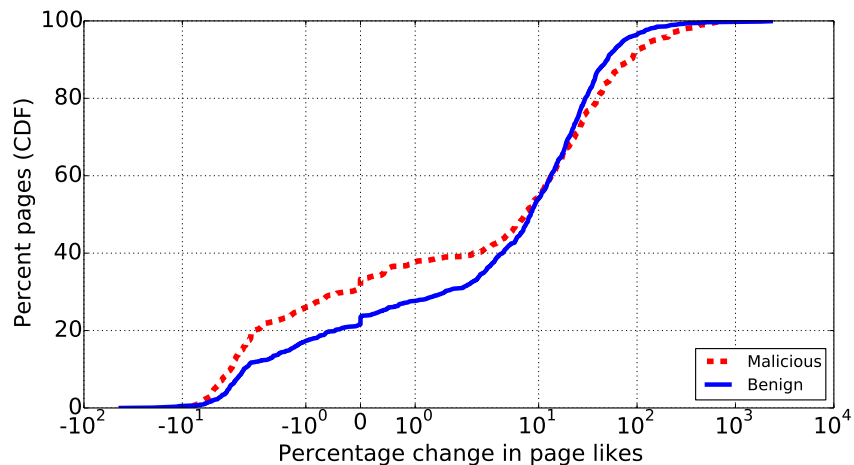


Figure 5.8: Percentage change in *page likes* (*gain factor*) over one year for all *pages* in our dataset. We observed that a larger proportion of malicious *pages* lost likes over time as compared to benign *pages*.

difference between malicious and benign *pages* with respect to this metric. We modelled the growth rate of *likes* on a *page* as a linear function over time and studied the distribution of the gradient for *page likes* across malicious and benign classes. This technique has been used in the past to study popularity on OSNs over time [31]. We observed low values for standard error of the estimated gradient, and significant *p-values* for both classes, signifying good fit (see Table 5.7).

Table 5.7: Mean values for standard error of estimated gradient and correlation *p-values* for linear model. We obtained low error rates and *p-values* signifying a good fit.

Metric	Class	Malicious	Benign
Standard error of estimated gradient	$\mu_{err}$	0.8273	0.5583
	$\sigma_{err}$	4.6282	4.3492
p-value for correlation	$\mu_p$	0.016	0.009
	$\sigma_p$	0.121	0.082

Figure 5.9 shows the distribution of the gradients ( $\tan^{-1} m$ , where  $m = \frac{y-c}{x}$ ;  $y$  = no. of likes,  $x$  = days,  $c$  = intercept) we obtained for malicious and benign classes. We observed that gradients for malicious pages were more evenly distributed as compared to benign pages. The median gradient value for the malicious class (7.85) was lower than the median gradient value for the benign class (9.11), but the difference in the two distributions was statistically insignificant (*p-value*=0.38).

**Description** Each Facebook *page* has multiple attributes that make up its description, for example, *username*, *description*, *general.info*, *personal.info*, *category*, *location*, *phone\_number*, *mission*, *bio*, etc. While some attributes (like *username*, *category*) are available for all *pages*, the presence

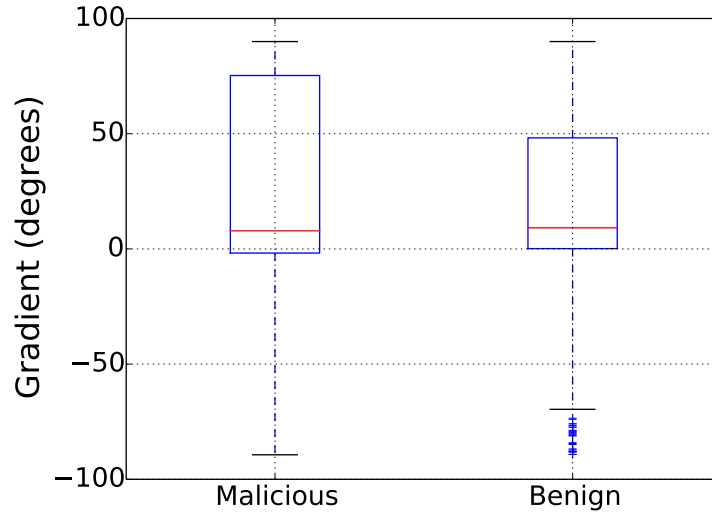


Figure 5.9: Distribution of the popularity gradients (in degrees) for malicious and benign pages in our dataset. Although the distributions look different, we did not find the difference to be statistically significant.

of other attributes (like *general\_info*, *mission*, *bio*, etc.) is dependent on the category of the *page*. We examined changes in all such attributes (wherever available) for both, malicious and benign *pages* in our dataset. Figure 5.10 shows the top 20 attributes in which we observed at least one change during the one year period of our study. We came across a total of 44 attributes that were changed once or more.<sup>11</sup> The remaining 24 attributes were changed by less than 1% of all pages in our dataset.

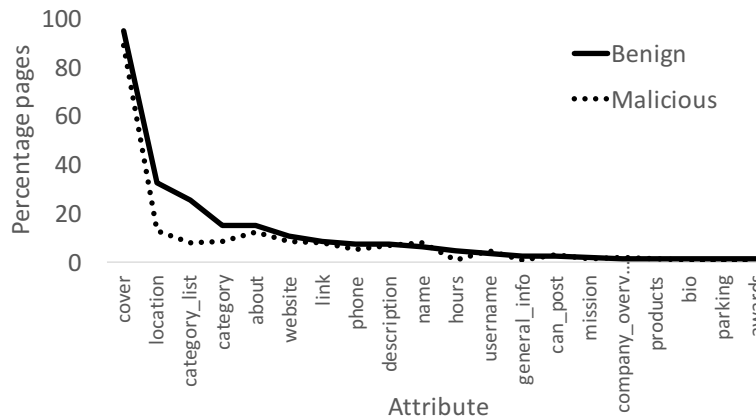


Figure 5.10: Top 20 attributes across malicious and benign pages that were changed at least once during one year of observation. We identified a total of 44 such attributes, but all remaining attributes underwent one or more changes in less than 1% pages.

<sup>11</sup>Exact description for each of these attributes can be found at <https://developers.facebook.com/docs/graph-api/reference/page/>

We observed a strong correlation between malicious and benign *pages* in terms of the proportion of *pages* changing each attribute, for all 44 attributes ( $r = 0.967$ ,  $p - value < 0.01$ ). This correlation depicted that an attribute that was changed by a large proportion of benign *pages*, was changed by a large percentage of malicious *pages* too, and vice versa. For example, the *cover* attribute (cover picture) was changed by 94.44% of benign *pages* and 88.51% of malicious *pages*, while *name* was changed by 6.25% of benign *pages* and 7.49% of malicious *pages*.

We further investigated each of the top 20 attributes individually to see if the changing behavior of any of these attributes could help distinguish between malicious and benign *pages*. We applied the Kolmogorov-Smirnov (KS) 2 sample test to compare the distributions of the number of times each of these attributes were changed by malicious and benign *pages* in our dataset, and found that 19 out of the 20 attributes were not informative ( $p - value > 0.1$ ). The only statistically significant distribution corresponded to changing behavior of the *category\_list* attribute ( $p - value < 0.05$ ).

The above results corroborate with our previous findings, suggesting minimal presence of a significant difference between malicious and benign *pages*, even in terms of popularity and attribute change behavior over time. These findings further suggest that distinguishing between malicious and benign Facebook *pages* based on spatial characteristics, temporal behavior, and other information associated with these *pages* is a hard and challenging problem. Using all the aforementioned insights and observations obtained from our dataset, we construct a diverse and robust feature set, and attempt to automate the task of identifying malicious *pages* from benign *pages* using supervised learning, as described in the next section.

## 5.4 Automatic detection of malicious pages

Past research has shown that URL blacklists and reputation services are ineffective initially, and take time to update [141]. Moreover, lack of blacklists and reputation services for malicious content other than phishing, and malware demand the need for an automated solution to analyze and detect malicious Facebook *pages*. To fulfil this need, we trained multiple supervised learning algorithms on our dataset of malicious *pages* in an attempt to create an effective model for automatic detection of malicious Facebook *pages*, independent of third party reputation services.

**Classification algorithms:** We experimented with a variety of classification algorithms – naive bayesian, logistic regression, decision trees, random forests, and artificial neural networks. We used balanced training and test sets containing equal numbers of positive and negative examples (627 malicious *pages*, and 627 benign *pages*), so random guessing results in an accuracy, as well as an area under the receiver operating characteristic (ROC) curve (AUC) of 50%. Although our actual dataset is highly unbalanced, we use a balanced dataset for our experiments in order to obtain a model for better classification of new data, as opposed to a model that would represent our dataset



better.

**Feature set:** We extracted a total of 96 features, 55 features from *page* information, and 41 from the posts published by the *pages* in our dataset, to train and evaluate the aforementioned algorithms. Table 5.9 shows a list of all these features, along with their category and feature type.

In addition, we trained and evaluated bag-of-words models obtained using the textual content present in the posts published by these *pages*. A bag-of-words was chosen owing to differences we found in the textual content generated by the pages, especially politically polarized content that we observed in Section 5.3.1. We used the most recent 100 posts published by Facebook *pages* in our dataset for calculating post features and building our bag-of-words. We did not find any explicit distinctive features in our dataset to separate the malicious class from benign, thus making effective automation a hard goal to achieve. We thus tried to build an extensive feature set to capture as much characteristics as possible.

#### 5.4.1 Supervised learning with *page* and post features

Table 5.8 shows the accuracy and ROC AUC values for various classification algorithms that we applied on the *page* and post level features. We considered post features extracted from the most recent 100 posts generated by the pages. All trained models were evaluated using 10-fold cross validation. A combination of post and *page* level features performed the best, signifying that both the characteristics, and posting behavior of *pages* need to be recorded for efficient automatic detection of malicious *pages*. The Logistic Regression classifier achieved highest accuracy of 76.71% with an area under the ROC curve of 0.846.

A possible reason for the Logistic Regression classifier performing the best in this case could be the presence of a large number of linearly separable numeric features in our dataset. However, Logistic Regression did not drastically outperform other techniques like Random Forest. Given that the overall performance of all classification algorithms was well under 80% in terms of accuracy, we believe that differentiating between poor quality and benign Facebook pages using this set of features is a hard problem to solve.

We performed further experiments by varying the number of most recent posts we considered for generating post features. Figure 5.11 shows the ROC AUC values achieved by the Logistic Regression classifier with varying post history. We started the experiment by considering 20 most recent posts for post features, and observed an overall increasing trend in performance as we increased the number of most recent posts to 100. We did not go beyond 100 for the post history because our ground truth dataset for malicious and benign pages was derived based on this limit.

The classifier achieved a maximum ROC AUC value of 0.85 (and an accuracy of 77.67%) with a post history size of 80 posts using a combination of page and post features. Performance remained

Table 5.8: Classification accuracy and ROC AUC values for automatically detecting malicious Facebook pages. Logistic Regression classifier performed the best.

Classifier	Feature set	Acc. (%)	ROC AUC
Naive Bayesian	<i>Page</i> features	63.95	0.685
	Post features	69.61	0.753
	<i>Page</i> + post features	70.81	0.776
Logistic Regression	<i>Page</i> features	67.38	0.745
	Post features	76.55	0.825
	<i>Page</i> + post features	<b>76.71</b>	<b>0.846</b>
Decision Trees	<i>Page</i> features	65.55	0.668
	Post features	71.37	0.720
	<i>Page</i> + post features	70.81	0.758
Random Forest	<i>Page</i> features	67.86	0.750
	Post features	74.95	0.829
	<i>Page</i> + post features	75.27	0.837

unchanged with respect to *page* features, since change in the size of post history does not affect *page* features (and is thus not reported in the figure).

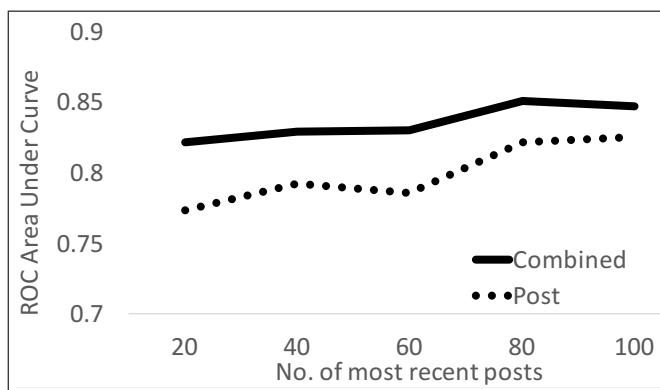


Figure 5.11: ROC area under curve values for Logistic Regression classifier corresponding to different sizes of post history. We observe an overall increase in performance as we increase the number of most recent posts used for computing post features.

## 5.4.2 Supervised learning with bag-of-words

In addition to *page* and post level features, we used a bag-of-words model to automatically identify malicious Facebook *pages*. We collected textual content from three sources (wherever present), viz. status message in the post, name and description of the link present in the post (if any).<sup>12</sup> We

<sup>12</sup><https://developers.facebook.com/docs/graph-api/reference/v2.6/post>

Table 5.9: Page and post level features used for training supervised learning models.

Category	Feature type	Feature
Page (55)	Boolean (19)	Affiliation, birthday, can post, cover picture, current location, working hours, description present, location, city, street, state, zip, country, latitude, longitude, personal interests, phone number, public transit, website field
	Numeric (34)	Average sentence length for description, average word length for description, parking capacity, category list length, check-ins, no. of email IDs in description, fraction of HTTP URLs in description, description length, fraction of URLs shortened, fraction of URLs active, likes, page name length, no. of subdomains in URLs, path length of URLs, no. of redirects in URLs, no. of parameters in URLs, [no. of !, no. of ?, no. of alphabets, no. of emoticons, no. of English stop words, no. of English words, no. of lower case characters, no. of upper case characters, no. of newline characters, no. of words, no. of unique words, no. of sentences, no. of total characters, no. of digits, no. of URLs] in description, description repetition factor, talking-about count, were-here count
	Nominal (2)	Category, description language
Posts (41)	Numeric (41)	Daily activity ratio, audience engaged, [average no. of upper case characters, average length, average word length, no. of English words, no. of English stop words] for description, message, and name fields, no. of posts containing the field [description, message, name], no. of comments, no. of likes, no. of shares, no. of posts with status_type [added_photos, added_video, created_event, created_note, mobile_status_update, published_story, shared_story, wall post], no. of posts with type [event, link, music, note, offer, photo, video, status], total no. of URLs, total no. of unique domains

performed experiments by calculating term frequencies of unigrams, bigrams, and trigrams, and limited our vocabulary size to the top 10,000 features.

A bag-of-words with 10,000 features produced a sparse feature vector. This sparse data prompted us to explore more state-of-the-art artificial neural network based learning techniques for fast and effective classification. We chose Sparsenn for this task.<sup>13</sup> Sparsenn is a C implementation of artificial neural networks based on stochastic gradient descent, designed for learning neural networks from high dimensional sparse data. The ability of Sparsenn to train a neural network quickly and efficiently with sparse data made it appropriate for our use case. Table 5.10 presents the results of our experiments.

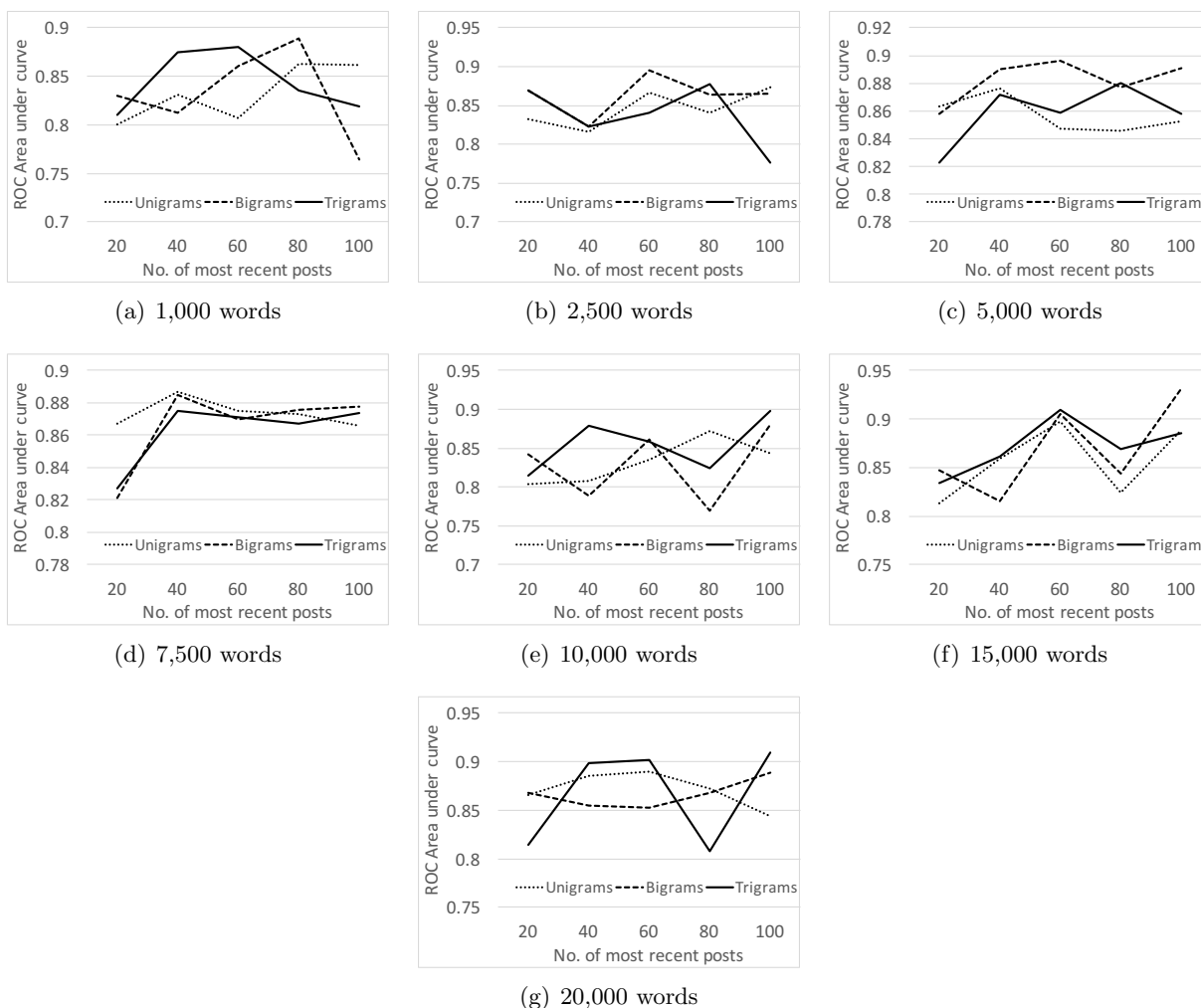


Figure 5.12: ROC AUC values obtained by neural networks trained on a bag of words for different sizes of bag of words.

<sup>13</sup><http://lowrank.net/nikos//sparsenn/>

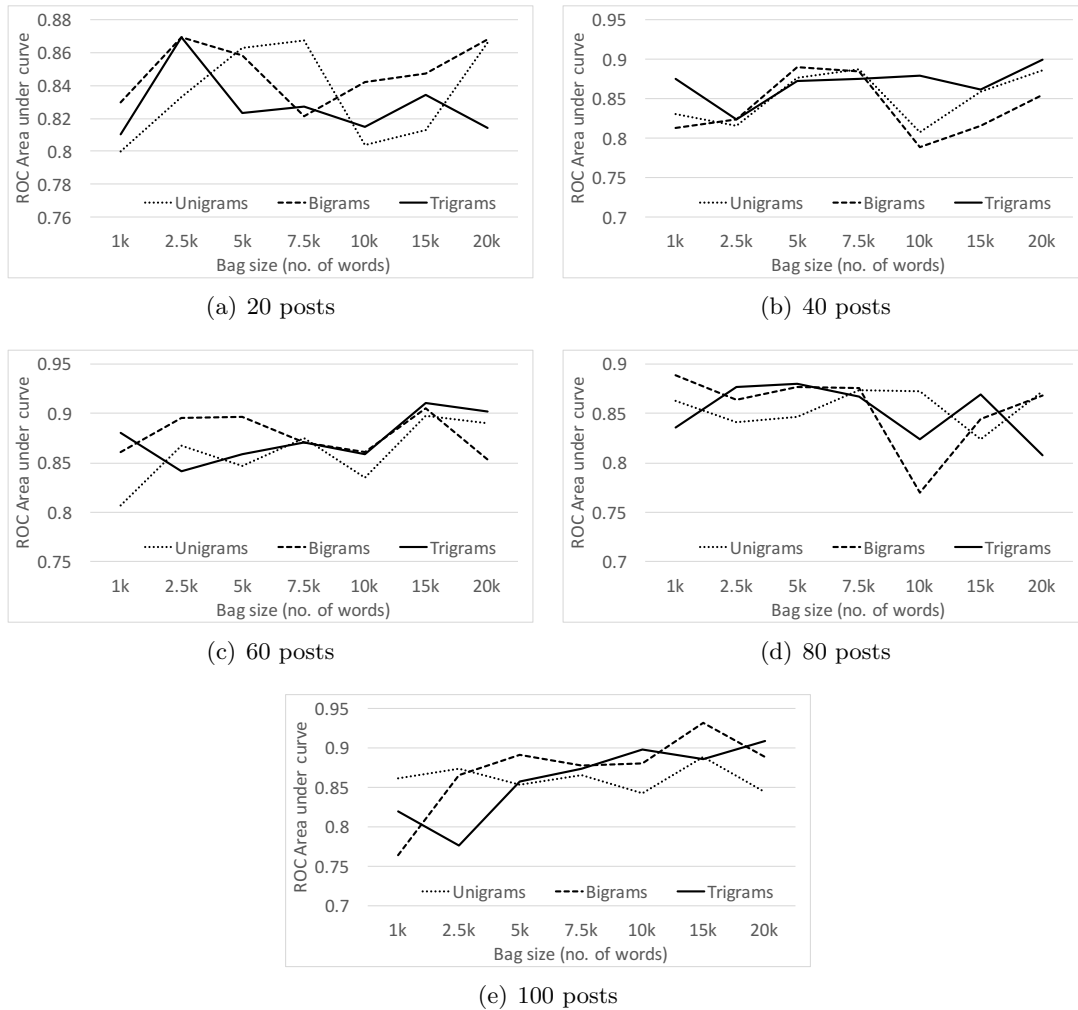


Figure 5.13: ROC AUC values obtained by neural networks trained on a bag of words for different sizes of post history.

Neural networks (hidden units = 64, and learning rate = 0.07, determined experimentally) on trigrams performed the best, achieving an accuracy of 84.13% with an area under the ROC curve of 0.9 (Figure 5.14). This meant that artificial neural networks trained on the top 10,000 trigrams outperformed all the other learning models including our previous models trained on *page* and post level features (discussed in Section 5.4.1).

We extended our experiments by performing a grid search over post history (number of most recent posts) and bag of words size. Using default values for hidden units (16) and learning rate (0.05), we varied the size of the bag of words from 1,000 through 20,000, and post history from 20 through 100 most recent posts published by the page. All these experiments were performed using unigrams, bigrams and trigrams. Figure 5.12 and Figure 5.13 show the varying values of area under the ROC

Table 5.10: Classification accuracy and ROC AUC values for automatically detecting malicious Facebook pages using bag-of-words. Artificial neural networks performed the best.

Classifier	Feature set	Acc. (%)	ROC AUC
Naive Bayesian	Unigrams	68.27	0.682
	Bigrams	69.06	0.690
	Trigrams	69.77	0.697
Logistic Regression	Unigrams	74.18	0.795
	Bigrams	74.34	0.791
	Trigrams	73.93	0.789
Decision Trees	Unigrams	68.12	0.678
	Bigrams	67.05	0.678
	Trigrams	66.63	0.672
Random Forest	Unigrams	72.26	0.794
	Bigrams	71.80	0.802
	Trigrams	72.18	0.794
Neural Networks	Unigrams	81.74	0.862
	Bigrams	84.12	0.872
	Trigrams	<b>84.13</b>	<b>0.900</b>

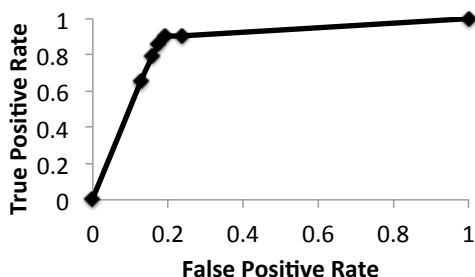


Figure 5.14: ROC curve for Neural Networks trained on trigrams.

curve for different sizes of post history and bag of words respectively. We achieved a maximum ROC AUC value of 0.931 using bigrams with a bag size of 15,000 words and post history size of 100.

## 5.5 Future work

**Politically polarized entities:** Our analysis revealed the presence of some politically polarized entities in our dataset of malicious *pages*. We verified the actual presence of all such entities in our dataset through manual verification. We do not intend to implicate such entities from our findings. Entities involved in politics tend to be followed by masses with similar orientation, and is a global phenomenon in the real world. It is likely that such activity exists on online platforms other than

Facebook too. We do not propose to debar such activity. However, we believe that extremely polar content should be moderated both online and offline, in order to maintain stability among the masses. An easy way to moderate such entities can be to display nudges or warning messages to users before they subscribe to such *pages* on any online platform [164].

**Beyond *pages*:** *Pages* on Facebook have a lot in common with Facebook groups and events. Groups and events can also be used to target large audiences at once. Moreover, Facebook has a common definition of “Page Spam” for *pages*, groups and events, and explicitly states that *Pages, groups or events that confuse, mislead, surprise or defraud people on Facebook are considered abusive*. Our analysis and results can thus be easily extended to study malicious groups and events as well.

**Automatic identification of malicious *pages*:** Our findings shed some light on subtle differences (like temporal behavior, content type, etc.) between malicious and benign *pages*, which we used to train various supervised learning algorithms to automatically differentiate between malicious and benign *pages*. These findings, however, are based on a limited history (100 posts) of *page* activity. Although it is possible to collect and analyze the entire history for all *pages*, doing so would be time consuming and computationally expensive. Moreover, *pages* can change behavior over time; malicious *pages* may stop spreading malicious content, while benign *pages* may start engaging in posting malicious content over time. To accommodate such changes in behavior, we recommend a self-adaptive model which relies on the most recent activity by the page. The history (number of posts) to consider can be decided experimentally. Such a model would be accommodative of the changing behavior of *pages* over time.

## 5.6 Conclusion

In this chapter, we identified and characterized Facebook *pages* posting malicious URLs. We looked beyond traditional types of malicious content like unsolicited bulk messages, spam, phishing, malware, etc., and studied a broader section of content that is deemed as malicious by community standards and Page Spam definitions established by Facebook. We focused on Facebook *pages* because of their public nature, vast audience, and inflated malicious activity [36]. Our observations revealed presence of politically polarized entities among malicious *pages*. We also found a substantial number of malicious *pages* dedicated to promote content from a single malicious domain. Further, we observed that malicious *pages* were more active than benign *pages* in terms of hourly, daily, and weekly activity. Network analysis revealed presence of collusive behavior among malicious *pages* that engaged heavily in promoting each others’ content. We applied multiple machine learning algorithms on our dataset to automate the detection of malicious pages, using *page* and post level features, and bag-of-words. Our experiments showed that artificial neural networks trained on

bag-of-words work best in detecting malicious *pages* automatically. We believe that our findings will enable researchers to better understand the landscape of malicious Facebook *pages* that have been hiding in plain sight and promoting malicious content seemingly unperturbed.



## Chapter 6

# Towards Understanding Crisis Events On Online Social Networks Through Pictures

This chapter is joint work with Anshuman Suri (undergraduate student at IIIT-Delhi), Aditi Mithal (undergraduate student at IIIT-Delhi), and Varun Bharadhwaj (undergraduate student at NIT-Trichy) and is an extension of a paper accepted for publication at the IEEE/ACM Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2017 [38].

### 6.1 Introduction

The last decade has witnessed a revolution in communication technology with the introduction of Online Social Networks like Facebook and Twitter. Especially during crisis or other news-making events such as earthquakes, bomb blasts, and terror attacks, people switch to social media to share updates, experiences and stay up to date [55, 118]. In the last few years, various researchers have analyzed textual content to identify and analyze sentiment, themes, and popular topics of discussion among the masses on OSNs during crisis events [2, 64, 79, 133, 150]. Sentiment and topic analysis of textual content on OSNs is widely used by researchers to understand the event, gauge the pulse of citizens, and draw inferences. However, all OSN content is not necessarily in textual format. Our past research has reported that high percentage of posts generated during real-world events contain only images and no text at all.<sup>1</sup> This points to the fact that the methodology adopted for most aforementioned research misses out on such sections of content, which do not contain text.

---

<sup>1</sup><http://precog.iiitd.edu.in/blog/2016/08/imagesononlinesocialmedia/>

Moreover, even if a post contains text along with an image, the text may not be representative of the topic or sentiment depicted by the accompanying image. Consider the post in Figure 6.1 for instance. While sentiment analysis on text would reveal positive sentiment for this post, the sentiment associated with the image is contrasting.



Figure 6.1: Example of a Facebook post where sentiment associated with the post text is in contrast with the sentiment associated with the text embedded in the image.

Considering the human brain’s affinity towards visual content [121], it is likely that the pulse and sentiment of user-generated content as perceived by researchers through text, differs from the true sentiment, since most past research does not consider images to draw inferences.

In this chapter, we attempt to answer two research questions, a) what are the popular themes and sentiment among images that are posted on OSNs during a crisis event? and b) how, if at all, are visual themes and sentiment different from their textual counterparts? To this end, we study a large dataset of over 57,000 images posted on Facebook during the terrorist attack in Paris in November 2015. We employ state-of-the-art image analysis techniques and construct a novel 3-tier pipeline for large-scale mining and measurement of the themes and sentiment of images posted on OSNs. Results of our measurement study reveal sizeable differences in prominent themes and sentiment drawn from images and text. We observed that textual content embedded in images, as well as text contained in posts, depicted negative sentiment. On the other hand, images, were found to inspire positive sentiment in general. Upon manual inspection, we observed that this contrasting behavior was largely due to the popularity of images offering support and solidarity

to the victims of the attacks. We extracted visual themes from images and found that two of the top 10 themes among images were related to instances of misinformation and were not prominent in textual content. Further, textual content extracted from images revealed multiple (potentially sensitive) topics associated with “refugees”, “passports”, etc. which were popular in image text, but not in post text.

These findings indicate the presence of useful information in the form of images posted on OSNs during crisis events, which haven’t been widely explored in the literature. Such information can be of particular interest to researchers who currently resort to text for mining, analyzing, and understanding popular topics, sentiment, sensitive information, misinformation, etc. from OSNs during events. To the best of our knowledge, this is the first large-scale study to understand visual themes and sentiment on social networks during crisis events. Further, the resulting 3-tier pipeline we employed for our analysis scales to a generalizable model that can be applied to understand any similar crisis event. We have made this pipeline-as-a-service publicly available as a RESTful API at <http://labs.precog.iiitd.edu.in/resources/Helix/>.

**About the event:** A series of coordinated terrorist attacks took place in Paris on November 13, 2015, at 21:20 Central European Time. Suicide bombers and gunmen attacked a stadium in Saint-Denis, Paris. This attack was followed by mass shootings, and a suicide bombing, at cafes and restaurants. Gunmen carried out another mass shooting and took hostages at a concert in the Bataclan theater, leading to a stand-off with police. The attackers were shot or blew themselves up when police raided the theater. A total of 130 people were killed, and 368 others were injured. News about the event spread instantly on all OSN platforms including Facebook. Hundreds of users posted live pictures of the event, and thousands posted messages offering condolence and support. <sup>2</sup>

## 6.2 Related Work

There exists literature in the space of studying images during crisis events on a small scale on OSNs, as well as studying crisis events on OSNs. Our work contributes towards enhancing the work done in both these areas.

### 6.2.1 Images on OSNs during crisis events

Multiple researchers have studied images posted on OSNs to analyze crisis events. Gupta et al. attempted to identify and characterize the spread of fake images on Twitter during Hurricane Sandy in 2013. Although the paper was focused on the identification of fake images, the methodology adopted by the authors relied on user and tweet features rather than image features. Authors

---

<sup>2</sup><http://www.bbc.com/news/world-europe-34818994>

manually identified a set of fake images from news articles and blogs and used the URLs of these fake images to expand their dataset of tweets containing fake images. This dataset was used to extract user and tweet level features to automatically identify tweets containing fake image URLs from tweets containing real image URLs [68]. Vis et al. conducted an exploratory analysis of images shared on Twitter during the 2011 UK riots. Similar to Gupta et al.'s approach, authors manually classified images into 14 categories for characterization [160]. More similar work includes empirical analysis of Twitter images during the 2012 Israeli-Hamas conflict, where authors examined images shared by two Twitter accounts over a 2-month time frame. A total of 243 images were captured and studied manually to discover prominent themes and frames, human characters, etc. present in the images [139]. Kharroub et al. studied 581 Twitter images from the 2011 Egyptian revolution and found more images depicting crowds and protest activity as compared to images depicting violent content. In addition to most prominent visual themes, authors of this work tried to find whether user information helps in predicting image retweets, and whether image themes vary across different phases of the event [88].

As evident from prior research, images play a crucial role in measuring public sentiment during crisis and mass emergency events like terror attacks, and in cases of detecting online radicalization. All aforementioned research however, is restricted to small scale, because of the manual effort involved in measurement and analysis. The use of images for analyzing events on a large-scale remains largely unexplored. We attempt to overcome this restriction by exploring automated methods to extract meaningful information from images posted on Facebook during the Paris Attacks in 2015.

### 6.2.2 Crisis event related studies on OSNs

Numerous researchers have looked at textual content to study crisis events on OSNs. Hughes et al. studied the use of the Twitter social network during four emergency events, and compared how this behavior was different from general Twitter use. Authors found that Twitter messages sent during these types of events contained more displays of information broadcasting and brokerage as compared to general Twitter messages. Textual features like *replies*, URLs, and the presence of certain keywords were used to draw these findings [79]. Gupta et al. presented a study to identify and characterize communities from a set of users who post messages on Twitter during three major crisis events that took place in 2011. Authors used textual content similarity in addition to link (network) and location similarity to identify clusters of users similar users [64]. Rudra et al. proposed a novel framework to assign tweets posted during mass emergency events into different situational classes, and then summarize those tweets. Similar to Hughes et al.'s approach of using textual features, authors of this work also extracted features like numerals, nouns, locations, verbs, etc. present in tweet text to identify and extract event summary [133]. Thelwall et al. studied sentiment of English tweets during a month long period and found that popular events

were normally associated with increases in negative sentiment strength. Authors completely relied on tweet text to extract sentiment strength and draw inferences [150].

All aforementioned research used textual content to study events on OSNs and draw inferences, thereby missing out on a large section of content pertaining to images. As discussed previously, researchers have looked at images on OSNs using manual techniques, and reported interesting findings. The aforementioned research highlights the need for automated large-scale techniques to study and mine images to extract sentiment, themes, and other similar useful information that can be used by researchers to better understand the users’ reactions with respect to crisis events on OSNs and draw more accurate inferences.

## 6.3 Methodology

### 6.3.1 Data collection

We collected public posts about the Paris attacks using Facebook’s Graph API Search endpoint [50] between November 14, and November 25, 2015. Although the post search feature was deprecated in April 2015, we were able to access this feature through the OAuth token generated by Facebook’s mobile app for iOS. This technique has been used in the past for gathering data from Facebook [70]. Table 6.1 provides the detailed description of our dataset.

Table 6.1: Descriptive statistics of our dataset we collected from Facebook during the Paris Attacks in 2015.

Keywords used	#ParisAttacks, #PrayForParis
Unique posts	131,548
Unique users	106,275
Posts with images	75,277
Unique users posting images	67,570
Total images extracted	57,748
Unique images extracted	15,123

The Graph API returns posts in JSON format (JavaScript Object Notation), and each post has a *type* associated with it. We filtered out all posts of *type* “*photo*” and re-queried the Graph API in February 2016 to obtain the actual images in these posts. Upon re-querying the Graph API, we noticed that some of the posts had been deleted from Facebook, and were no longer accessible. Also, posts with entire photo albums were also categorized as *type* “*photo*”, and images inside these albums were not directly accessible via the API. Eventually, we were able to collect 57,748 images in total. We identified duplicates using the difference hash (dHash) image hashing algorithm<sup>3</sup> and

<sup>3</sup><http://www.hackerfactor.com/blog/?/archives/529-Kind-of-Like-That.html>

obtained a total of 15,123 unique images (Table 6.1). This dataset of 15,123 unique images has been anonymized and is available for research purposes.<sup>4</sup>

**Ethical considerations:** The goal of this research was to study and understand the event under consideration (Paris Attacks, 2015) through the visual content posted on Facebook. We did not collect any private user data. We were not able to find any official documentation from Facebook about the feature we used for our data collection, at the time of writing this chapter. However, to the best of our knowledge, this technique of collecting data does not violate any of Facebook’s terms.<sup>5</sup> This data was strictly used for research purposes only. We respect the privacy of Facebook users, and our work does not disclose any personally identifiable information about any individual or group whose data was part of our dataset.

### 6.3.2 Image characterization

Understanding and interpreting images is a complex task. As previously discussed, past research on studying the role of images on OSNs has largely relied on manual methods to perform measurement studies on images [68, 130, 160]. This methodology is time-consuming and not scalable for bigger datasets containing more than a few hundred images. With millions of images generated on OSNs every day,<sup>6</sup> manually looking at images is a futile way to understand visual content and draw any meaningful conclusions in a timely manner.

To overcome this drawback, we attempt to use automated methods to characterize and study images in our dataset. We utilize state-of-the-art object detection techniques coupled with minimal human effort, domain transfer deep learning, and optical character recognition techniques, and propose a 3-tier pipeline to extract human understandable descriptors from images quickly, and on a large-scale. Unlike previously used low-level image descriptors like SIFT [101] and SURF [9], our pipeline generates high-level human understandable descriptors that associate abstract level concepts (themes) and sentiment with images.

This pipeline is almost entirely automated, and significantly reduces the amount of human involvement required for understanding news-making events through images on a large-scale. Further, this technique is the basis for a generalizable method that can be applied to any similar event. Figure 6.2 shows the architecture of our proposed pipeline.

**Tier 1: Visual Themes** We use TensorFlow implementation of Google’s Inception-v3 model [147] for image classification. Inception-v3 is a deep convolutional neural network (CNN) model trained

---

<sup>4</sup><https://goo.gl/jKgqJA>

<sup>5</sup>Facebook Terms of Service, retrieved on December 22, 2016. <https://www.facebook.com/terms.php>

<sup>6</sup><http://www.businessinsider.in/Facebook-Users-Are-Uploading-350-Million-New-Photos-Each-Day/articleshow/22709734.cms>

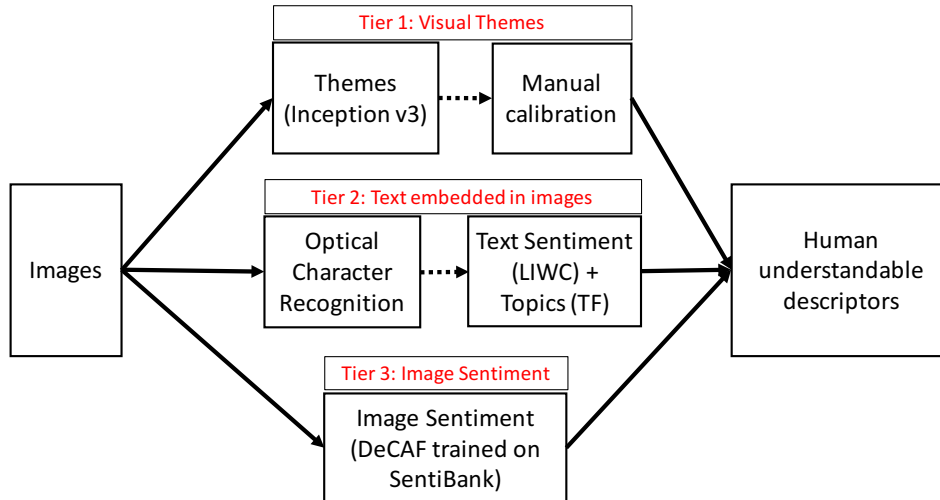


Figure 6.2: Architecture of our 3-tier pipeline used to extract human understandable descriptors from images.

for the ImageNet Large Visual Recognition Challenge using the data from 2012 and tries to classify images into 1,000 classes [134]. This model reports a top-1 error rate of 17.2%, signifying that the most probable label predicted by this model is correct in 82.8% cases in the test dataset. However, it is important to note that the generated label comes from a fixed set of 1,000 labels, which may not be large enough for characterizing the wide variety of images we come across on social networks in practice. To establish the accuracy of this model on our dataset, we opted for manual verification. We recruited human annotators through word-of-mouth publicity in an educational institute. All annotators were undergraduate computer science students and active Facebook users between the age of 18 and 21.

We got a random sample <sup>7</sup> of 2,545 unique images annotated by two or more annotators. Each annotator was independently shown one image at a time. The job of each annotator was to mark whether they *agreed* with the label generated by the Inception-v3 model for the given image, or not. Using majority voting, we found that the model achieved an accuracy of 38.87% on our data sample. Given that there are 1,000 possible output labels, this accuracy value is much better than random guessing; assuming equal class sizes, the probability of a random guess being correct is 0.1%. However, through a small manual exercise, we were able to boost this accuracy significantly.

We hypothesized that images in our dataset with the same labels are highly likely to be similar, regardless of the labels associated with them being correct or not. This is because of the fact that CNNs model the vision system in animals, and are likely to group similar images together. We tested this hypothesis for the images associated with the top 20 most frequently occurring labels in our dataset, and found it to hold true. For example, the “Peace for Paris” symbol created by

<sup>7</sup>Random sample was generated by using the “sample” function in Python’s *random* library.

French graphic designer Jean Jullien, was labeled “bolo tie” by the model (Figure 6.3). Using this observation, we renamed 8 out of the top 20 labels to better suit the images under each of these labels. Table 6.2 shows the 8 out of the 20 most common labels that were generated by the Inception-v3 model for our dataset, and the labels we replaced them with. This exercise of renaming labels boosted the accuracy of the model to 51.34% on our random sample. We used these modified labels for our analysis. This step of manually calibrating the model output is the only human effort required in our entire pipeline.

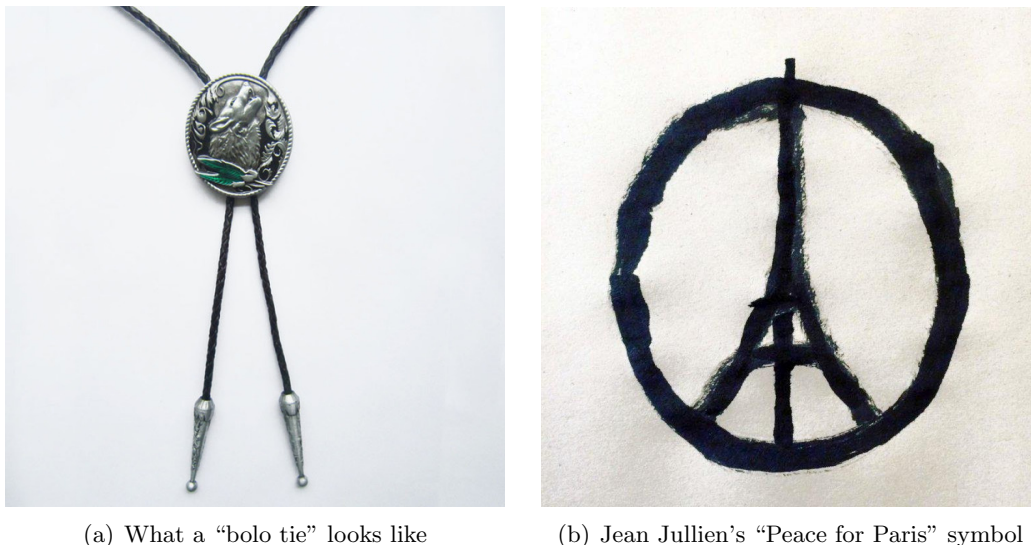


Figure 6.3: Visual similarity between a bolo tie and the famous ‘Peace for Paris’ symbol. This similarity (in addition to others) is captured by the Inception-v3 model, and can be exploited to increase accuracy.

Table 6.2: List of labels generated by the Inception-v3 model, and the labels they are renamed with. This renaming process boosted the accuracy of the model on our dataset by almost 13%.

Label generated by Inception-v3	Label they were replaced with	No. of occurrences in our dataset of unique images
book jacket, dust cover, dust jacket, dust wrapper	Poster	1,024
bolo tie, bolo, bola tie, bola	Jean Jullien’s “Peace for Paris” symbol	350
church, church building	Eiffel tower	258
obelisk	Eiffel tower	210
envelope	Poster	204
stupa, tope	Eiffel tower	173
drilling platform, offshore rig	Eiffel tower	137
radio telescope, radio reflector	Eiffel tower	68



**Tier 2: Text embedded in images** Past studies on analyzing topics, events, sentiment, etc. on OSNs have been largely limited to using textual content generated by users to draw inferences [64, 79,133,150]. This technique, however, misses out on a large section of textual information embedded in images, making it prone to missing out on being able to capture the complete picture. Figure 6.4 shows an example of textual content embedded in an image in our dataset.

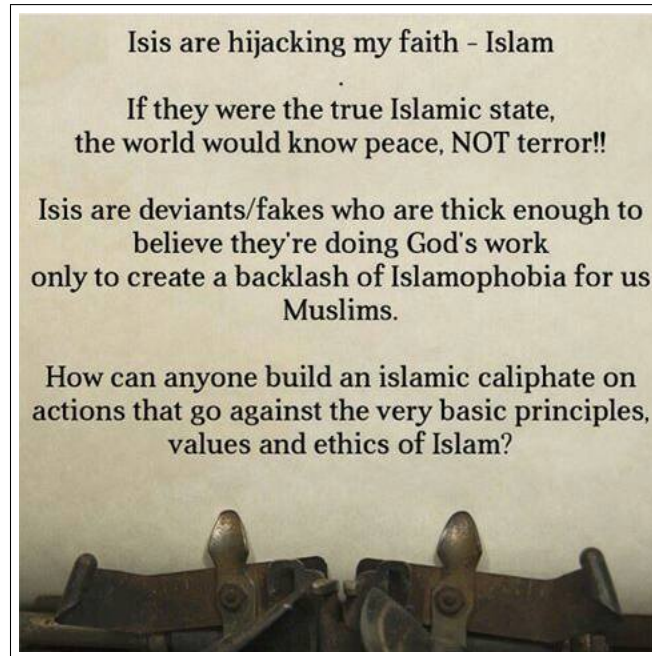


Figure 6.4: Example of text embedded in an image posted on Facebook during the Paris Attacks in 2015. We found thousands of images containing text in our dataset.

We tested and evaluated two optical character recognition (OCR) libraries – PyTesseract<sup>8</sup> (Python wrapper for Tesseract OCR<sup>9</sup>), and OCRopy<sup>10</sup> – to extract text from the images in our dataset. To compare the performance of the two libraries, we first established a ground truth dataset by manually extracting text from a random sample of 1,000 images from our dataset. We then used five string similarity metrics (Jaro-Winkler distance, Jaccard index, Cosine similarity, Hamming distance, and Levenshtein distance) to compare the results produced by PyTesseract and OCRopy with the ground truth, separately. Scores from all five metrics indicated that PyTesseract performed better on our dataset than OCRopy (p-value < 0.01 for all five metrics), so we used PyTesseract for our final analysis. In all, we were able to extract text from a total of 31,869 images in our dataset.

<sup>8</sup><https://pypi.python.org/pypi/pytesseract/>

<sup>9</sup><https://github.com/tesseract-ocr/tesseract>

<sup>10</sup><https://pypi.python.org/pypi/ocropy>

**Tier 3: Image sentiment** Sentiment derived from textual content generated by users on OSNs has been widely used by researchers in various contexts [12, 144, 150, 156]. However, few attempts have been made to understand the sentiment associated with images posted on OSNs [168, 170, 171]. Studies suggest that the human brain is hardwired to recognize and make sense of visual information more efficiently [121]. Thus, it is likely that sentiment extracted from textual content alone may not be representative of the overall sentiment associated with a theme or event. To this end, we attempt to extract sentiment from images using domain transfer deep learning.

The Inception-v3 model can be retrained to perform other visual recognition tasks using features extracted by the model during the training phase. This concept is known as domain transfer learning [120], and is available in the form of an open-source implementation, called Deep Convolutional Activation Feature (DeCAF).<sup>11</sup> DeCAF is a state-of-the-art deep CNN architecture for transfer learning based on a supervised pre-training phase [40]. We use this open-source implementation to retrain the Inception-v3 model on the SentiBank dataset to identify image sentiment. The SentiBank database comprises a total of half million Flickr images, extracted by querying the network using Adjective-Noun Pairs (ANPs) [13]. Since noun queries such as “dog”, “baby”, or “house” do not portray a well-defined emotion, these queries were prefixed with adjectives to form ANPs like “happy dog”, “adorable baby”, “abandoned house” etc., which associate these nouns with a strong emotion. We manually segregated these ANPs (and therefore, the images associated with them) into *positive* and *negative* classes for binary sentiment classification, and skipped the ANPs which did not fit clearly into a *positive* or *negative* sentiment. This exercise left us with a total of 305,100 positive sentiment images and 133,108 negative sentiment images. We performed a 10-fold random subsampling to balance the classes and obtain an unbiased model. For each fold, we split the dataset into three parts in an 80:10:10 ratio for training, validation, and testing respectively, and achieved a maximum accuracy of 69.8%.

Our 3-tier pipeline for image descriptor extraction is publicly available as a RESTful API and can be used to obtain descriptors for the images present in our dataset (API URL omitted to maintain anonymity).

## 6.4 Analysis and Results

Themes and sentiment are two of the most widely studied aspects of OSN content during crisis events in literature [26, 107]. We therefore focus on these two aspects of the images in our dataset and present our findings.

---

<sup>11</sup>[https://www.tensorflow.org/versions/r0.8/how\\_tos/image\\_retraining/index.html](https://www.tensorflow.org/versions/r0.8/how_tos/image_retraining/index.html)

### 6.4.1 Top visual themes featured misinformative images

Using the Tier 1 in our pipeline, we assigned a label to all 57 thousand images we collected (Table 6.1). Table 6.3 shows a list of the most commonly occurring image labels, along with their description according to our dataset. We manually browsed through images corresponding to each of the top 20 labels and found that the most common types of images comprised of posters, banners, screenshots of Facebook posts, Twitter tweets, etc. Cartoons and animated posters resembling a comic book were also very popular. More examples include the Pray for Paris peace symbol by French artist Jean Jullien (label: Bola Tie), images of candles and lamps offering support to the victims of the attacks (label: Candle waxlight), and images of the Eiffel Tower under varying lights, angles, and from various distances, that became very popular (labels: Obelisk, Crane, etc.).

Table 6.3: Top 20 most common image labels in our dataset. Majority of the images comprised of posters, banners, art work, etc. offering support for victims.

Label	Count	Description
Website	12,416	Images of posts, tweets, banners, etc.
Book jacket	5,383	Posters, banners, etc.
Comic book	3,803	Cartoons, animated posters
Fountain	1,264	Fountains at various locations
Envelope	1,248	Posters, banners, etc.
Suit (clothing)	1,246	People wearing suit-like clothes
Stage	1,135	Stages during public speeches, mass gathering events, etc.
Candle waxlight	1,021	Lit candles and lamps offering support to victims
Malinois	995	Police dog who died during the attack
Scoreboard	971	Images of sports stadium
Microphone	906	Individuals addressing the masses, reporters, etc. using microphones
Menu	868	Images containing well formatted text
Bola Tie	781	Peace for Paris symbol originally created by Jean Jullien
Bell cot	745	Various buildings
Jersey, T-shirt	743	People wearing t-shirts
Crane	677	Images of Eiffel Tower during twilight
Memorial Tablet	633	Variety of posters, hand written messages on boards, etc.
Church	629	Dark and grey scale images of Eiffel Tower
Palace	586	Large buildings, including Eiffel Tower from a distance
Obelisk	547	Eiffel Tower

We were also able to identify some peculiar topics and themes which were popular on the network during the event. The “Malinois” label appearing in the top 20 (see Table 6.3) corresponded to the breed of the police dog that died during the attacks, and evidently became very popular. However, the cause of death of the dog was incorrectly quoted in multiple such images. Figure 6.5(a) shows one such picture of the police dog and states that the dog was killed when a suicide bomber

detonated her explosive vest. However the real cause of the dog’s death, as later clarified by French police, was multiple gunshot wounds caused by the French police forces’ “Brenneke” bullets.<sup>12</sup> We collected all such images quoting misinformation in our dataset and found that these images had gathered over 1.1 million likes, 321 thousand shares, and 38 thousand comments.

Similarly, one of the blasts during the attacks took place outside a football stadium, whose pictures quoted incorrect information and became viral. These pictures were captured using the “Scoreboard” label; manual verification of images marked with this label revealed that most of these images captured the sports stadium. Figure 6.5(b) shows one such picture of the stadium and states that a Muslim security guard named Zouhier stopped a suicide bomber from entering the Stade de France stadium, thus saving hundreds of innocent lives of people inside the stadium. BBC later confirmed that it was not him who turned away the bomber. Instead, Zouheir was stationed elsewhere in the stadium, and related what he heard from colleagues who were closer to the bomb blast.<sup>13</sup> All instances of this misinformative image in our dataset garnered over 21 thousand likes, 11 thousand shares, and 450 comments.

This technique of automatic identification of themes and topics from images on a large-scale can be especially helpful to identify popular instances of misinformation spread through images. Slight modifications to convolutional neural network based labeling models like Inception-v3, can aid in identifying potentially harmful and sensitive content such as guns, blood, etc. in images, and help monitor the flow of such images, and react in a timely manner, if needed.

#### 6.4.2 Text embedded in images featured sensitive topics and reflected negative sentiment

Applying optical character recognition (OCR) on images in our dataset revealed 31,869 images (55% of all images in our dataset) which contained text embedded in them.

**Prominent topics:** Table 6.4 shows a mutually exclusive set of the 20 most frequently occurring relevant words in the text we extracted from images and posts. We picked 500 most commonly occurring words in images that were not present in post text, and vice versa, to identify prominent themes among image and post text independently. We noticed that the most commonly occurring words among image and post text had less than 45% overlap, highlighting that popular words among image text were considerably different from those in post text.

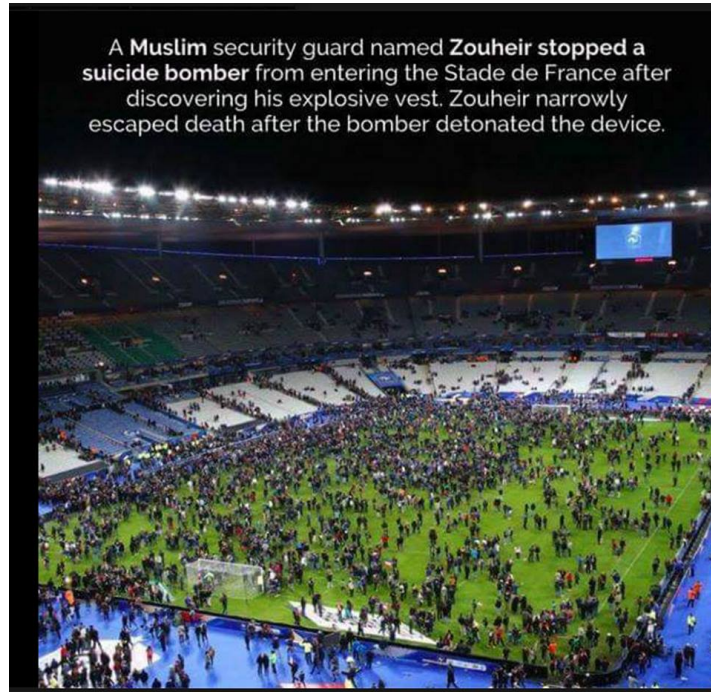
Text extracted from posts was dominated by words like “*prayers*”, “*prayfortheworld*”, “*life*”, “*support*”, “*god*” etc., depicting support and solidarity for the victims. Text extracted from images however, revealed some potentially sensitive topics like *refugees*, *passports*, etc. which were not

<sup>12</sup><http://www.dailymail.co.uk/news/article-3446511/Confirmed-Diesel-hero-police-dog-Paris-attacks-shot-dead-wounded-innocent-neighbours-reckless-shooting.html>

<sup>13</sup><http://www.bbc.com/news/blogs-trending-34845882>



(a) Diesel the dog who was allegedly killed by terrorists



(b) Zouheir, the security guard who was claimed to have stopped a terrorist from entering the stadium

Figure 6.5: Rumors spread on Facebook in the form of images during the Paris Attacks in 2015. We used CNN based image summarization techniques to identify image themes and discovered that some of the most popular image themes were associated with rumors.

amongst the most talked about topics in post text. The terms “*refugees*” and “*texas*” in Table 6.4 captures the scanned image of a letter from the Governor of Texas to the president of the United States, which went viral. The letter stated that Texas would not accept any refugees from Syria in the wake of the attacks. Other similar examples of images containing text comprised of quotes from famous personalities including The Dalai Lama, Stacy Washington, Barrack Obama, Martin Luther King Jr., etc. We also uncovered a popular conspiracy theory surrounding the Syrian “passports” that were found by French police near the bodies of terrorists who carried out the attacks, and were allegedly used to establish the identity of the attackers as Syrian citizens.<sup>14</sup> Text embedded in images depicting this theme questioned how the passports could have survived the heat of the blasts and fire. This conspiracy theory was then used by miscreants to label the attacks as a *false flag* operation, influencing citizens to question the policies and motives of their own government. The popularity of such memes on OSN platforms can have undesirable outcomes in the real world, like protests and mass unrest. It is therefore vital to be able to identify such content and counter / control its flow to avoid repercussions in the real world. Interestingly, 8,273 of these 31,869 images

<sup>14</sup><http://www.aljazeera.com/news/2015/11/paris-attacks-give-rise-conspiracy-theories-151118093352559.html>

(25.95%) did not contain any user-generated textual content otherwise, indicating that most prior work on event analysis using text on OSNs would have entirely missed this set of data during their analysis, as discussed previously [64, 79, 133, 150].

Table 6.4: Mutually exclusive set of 20 most frequently occurring relevant keywords in post and image text, with their normalized frequency. We identified some potentially sensitive topics among image text, which were not present in post text. Word frequencies are normalized independently by the total sum of frequencies of the top 500 words in each class.

	Top words in posts		Top words in images	
	Word	Norm. freq.	Word	Norm. freq.
1.	retweeted	0.0055	house	0.0045
2.	time	0.0052	safety	0.0044
3.	prayers	0.0050	washington	0.0042
4.	news	0.0047	sisters	0.0039
5.	prayfortheworld	0.0044	learned	0.0038
6.	life	0.0043	mouth	0.0038
7.	let	0.0042	stacy	0.0037
8.	support	0.0042	passport	0.0037
9.	god	0.0040	americans	0.0036
10.	war	0.0039	refugee	0.0035
11.	thoughts	0.0038	japan	0.0028
12.	need	0.0038	texas	0.0027
13.	last	0.0037	born	0.0026
14.	lives	0.0037	dear	0.0026
15.	said	0.0034	syrians	0.0026
16.	place	0.0034	similar	0.0025
17.	country	0.0033	deadly	0.0025
18.	city	0.0032	services	0.0025
19.	everyone	0.0032	accept	0.0025
20.	live	0.0032	necessary	0.0025

**Text sentiment:** Researchers in the past have looked at text sentiment to draw inferences about the overall sentiment and emotion of users about a topic on OSNs. Since most modern content monitoring techniques also focus on textual content, obfuscating sensitive textual content like hate speech and propaganda by embedding it in images is a lucrative way for malicious entities to avoid detection. Thus, we hypothesize that the sentiment of text embedded in images would be different from the sentiment of textual content posted by users in the conventional form. To confirm our hypothesis, we employed Linguistic Inquiry and Word Count (LIWC) [124] to determine and compare the emotion of the image text and post text in our dataset. We found that text embedded in images was negative on average, and twice in magnitude as post emotion (Mann-Whitney U test:  $p$ -value < 0.01). Figure 6.6 shows the distribution of emotions across the two classes (image text and post text). Although we found more negative emotion in both images and posts as compared

to positive emotion, the magnitude of negative emotion as compared to positive emotion was much higher in images as compared to text. We also noticed that positive emotion in posts was 2.6 times higher in magnitude than positive emotion in images. For negative emotion, this magnitude dropped down to 1.25.

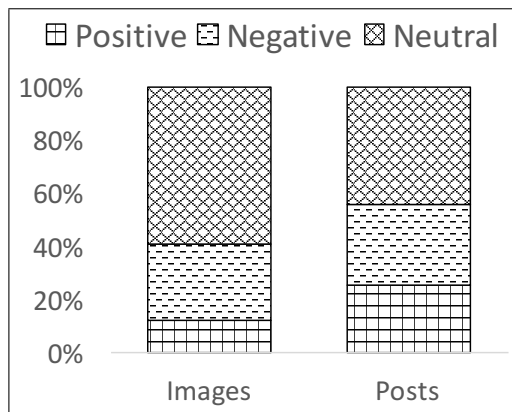


Figure 6.6: Distribution of positive, negative, and neutral text emotion in our dataset of images containing text, and posts. We found more negative emotion in both categories, but the magnitude of negative emotion as compared to positive emotion was higher in images.

These results indicate that textual content flowing on the network in the form of images is a critical source of information that can provide intriguing and detailed insights into crisis event related content posted on social networks. Such insights beg to be taken into consideration and analyzed thoroughly while making a judgment on the pulse and sentiment of the audience about the event.

### 6.4.3 Images inspired positive sentiment

Inferring image sentiment through the sentiment of text embedded in it (as discussed previously), is a small part of understanding the sentiment associated with an image. Text is only a part of the overall sentiment that an image may reflect. Moreover, there may be no text present in an image at all. Researchers have acknowledged the problem of understanding image sentiment, and come up with some solutions recently [168, 170, 171]. Using some of the most advanced techniques in the domain of image sentiment extraction including deep convolutional neural networks and domain transfer learning, we performed sentiment analysis of images in our dataset.

Contrary to text sentiment, we found that images, on average, portrayed a positive sentiment. We observed that close to 60% of the 57 thousand images in our dataset depicted a positive sentiment. However, as already discussed, the accuracy of our image sentiment model was not too high (approximately 70%). Therefore, to verify the validity of our observations, we recruited human annotators to manually mark a small random sample of 2,545 images from our dataset as positive,

negative, or neutral. Participants were also given an option to skip. Each image was annotated by at least 2 (and at most 3) participants. After removing the skipped images and using majority voting, we found that 50.95% images were marked as positive, whereas only 16.21% images were marked as negative. The remaining images were marked as neutral. This exercise confirmed our findings and affirmed the dominance of positive sentiment images in our dataset.

This observation can be attributed to the large number of pictures depicting support and solidarity for the victims of the attacks, which included posters, banners, people holding lit candles and lamps, the famous Peace for Paris symbol, etc. Such images inspire a positive sentiment on the viewer, as confirmed by our human annotators as well as the pre-trained sentiment prediction model.

Interestingly, we came across a substantial number of instances where image sentiment conflicted with the sentiment of the text present in the post. Consider the post shown in Figure 6.7 for example. While the text in the post reads, “*Horrible news.. No words :( :(*” reflecting highly negative sentiment, the image depicts the Eiffel tower lit up in French colors, signifying support for the victims and reflecting a positive sentiment. We observed that, out of the 19,954 posts in our dataset which contained user-generated textual content as well as an image, 25.33% of the posts (5,056 posts) had conflicting image and text sentiments. Out of these, 10.98% of the posts (2,192 posts) contained an image depicting a negative sentiment, whereas the textual content present in the post reflected positive sentiment. Similarly, 14.35% of the posts (2,864 posts) contained an image depicting a positive sentiment, whereas the textual content present in the post reflected negative sentiment.

We studied sentiment across post text, image text, and images temporally in order to understand how public sentiment changed across these three categories over time (see Figure 6.8). Images were found to depict positive sentiment throughout the 12 days of our observation period after the attacks. We observed that post text sentiment was negative during the first few hours, but gradually moved to the positive side over time. On the contrary, text embedded in images reflected positive sentiment initially but moved towards negative after the first few hours. This trend persisted for the rest of our observation period.

Upon manual inspection, we observed that for the first few hours after the attacks, news and updates mentioning words like “*killed, blast, attack, shooting, explosion*”, etc. in post text were prominent in the data, resulting in negative sentiment. This trend changed towards positive sentiment after a few hours, when prayers, support, condolences, solidarity, etc. for the victims started pouring in. On the other hand, image text started off with positive sentiment, dominated by a variety of images with text offering support (like “*Pray for Paris*”) posted initially. However, after a few hours, backlash against a terrorist organization, Syrian refugees, conspiracy theories, etc. embedded as text in images skewed image text sentiment towards the negative side.

Through this analysis, we uncovered a new dimension for mining sentiment from user-generated





Figure 6.7: Example of a post published during the Paris attacks, showing conflicting sentiments across image and text. This post was published by a verified page, garnering over 65 thousand likes and was shared 1,774 times.

content on OSNs, which has been largely unexplored in prior OSN related literature. Our results shed light on the varying sentiment depicted by images and text during the Paris attacks. While textual sentiment analysis revealed negative sentiment, we found that images shared on Facebook during the event depicted positive sentiment. We also found a considerable proportion of posts where textual sentiment and image sentiment depicted opposite polarity (8.75% of all images in our dataset). It is important to note that while text has been widely accepted in literature as a means to infer user sentiment during crisis events (and otherwise) on OSNs, the sentiment perceived by users is not restricted to text only. Instead, given the affinity of the human mind towards visual content, images are likely to contribute much more to the perceived sentiment of users as compared to text.

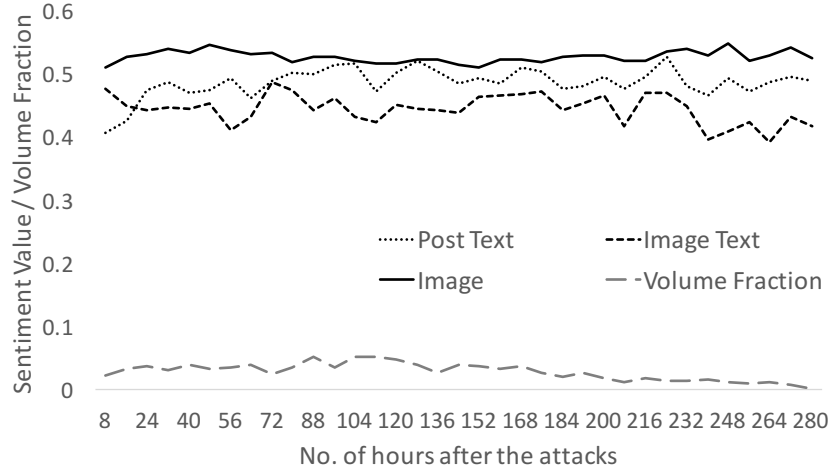


Figure 6.8: Sentiment values across post text, image text, and images over time. We observed images to inspire positive sentiment throughout the 12 days after the event. A value of 0.5 depicts neutral sentiment. Volume fraction represents the fraction of data from our dataset posted over time.

## 6.5 Case Studies: Instances of Misinformation

Detecting rumors, fake content, misinformation, and other similar categories of malicious content on OSNs has been a widely studied problem [67,68]. However, most modern techniques for these problems suffer the same limitations as most other OSN measurement and analysis studies discussed previously. A large section of malicious content detection techniques rely on text [67], while techniques focusing on images are limited by scalability issues due to human involvement for manually identifying malicious images [68]. To this end, we utilized multiple image analysis and processing techniques as discussed in Section 6.3, to identify instances of misinformation spread in the form of images, which are otherwise unidentifiable. We identified at least five independent cases of misinformation and rumors. We noticed that most of these examples had the potential to spark a religiously or politically driven propaganda and create disturbance in the real world, impacting thousands of lives. It is thus important to understand such examples in detail, learn from them, and try to identify and avoid their spread early in the future. Statistics for all the five rumors are summarized in Table 6.5. We now discuss them in detail.

**Diesel the dog killed by terrorists:** One of the most popular rumors spread during the attack was about a police dog named Diesel, who died in the line of duty. The dog was rumored to have died because of an explosion caused by one of the terrorists. However the real cause of death, as later clarified by French police, was multiple gunshot wounds caused by the French police forces’ “Brenneke” bullets.<sup>15</sup> Figure 6.9(a) shows an image of the dog, incorrectly quoting the cause of

<sup>15</sup><http://www.dailymail.co.uk/news/article-3446511/Confirmed-Diesel-hero-police-dog-Paris-attacks-shot-dead-wounded-innocent-neighbours-reckless-shooting.html>

death. This, and many other such pictures of the dog went viral on Facebook, and all of them still exist on the network.

Using OCR, we were able to identify at least 170 images in our dataset with the same rumor, embedded in them as text. These 170 images were a collection of recurring instances of 5 distinct images, which we identified using the dhash image hashing algorithm. We then identified 12 different Facebook pages (including 1 verified page with over 17 million likes) which posted one of these 5 distinct images, and found that the rumor image garnered a total of over 1 million *likes*, 38 thousand *comments*, and 320 thousand *shares*. Diesel the dog belonged to a breed of dogs known as Malinois. Our image characterization module based on the Inception model (Section 6.3.2) was able to identify this breed, and labeled all images of the dog as 'malinois'. This allowed us to further expand our dataset associated with this rumor, and we found an additional set of 825 images of the dog, labeled as 'malinois' in our dataset. Although these images did not contain the same rumor in the form of embedded text, given the immense popularity of the rumor, it is possible that users posting and engaging with these images did so under false impressions of the rumor.

**Muslim guard stopped terrorist from entering stadium:** Another popular rumor during the Paris Attacks was about a muslim 'hero' security guard named Zouheir, who was claimed to stop a stop a suicide bomber from entering the Stade de France stadium, thus saving hundreds of innocent lives of people inside the stadium. It was later confirmed by BBC that it wasn't actually him who turned away the bomber. Instead, Zouheir was stationed elsewhere in the stadium, and related what he heard from colleagues who were closer to the bomb blast.<sup>16</sup> Figure 6.9(b) shows such an image, with the rumor embedded on top as text. This rumor was also identified through embedded text extracted using OCR. Table 6.5 shows the total number of instances, and other statistics associated with this rumor.

**Eiffel Tower turns off lights for the first time since 1952:** Figure 6.9(d) shows another viral rumor image claiming that the Eiffel tower turned off its lights for the first time since 1952, in memory of the people who lost their lives in the November 2015 attacks. However, this claim turned out to be false, since lights at the Eiffel tower are turned off every night at 1 a.m. as standard practice.<sup>17</sup> We also found some other similar images of the Eiffel tower in the dark, claiming that this was the first time since 1889 that lights at the Eiffel tower went out. More images showed monuments from around the world light up in the colors of the French flag, with the dark Eiffel tower image in the middle, and captions like, "*When Paris turned out its lights, the rest of the world turned them on*". Using OCR, we were able to identify all such rumor images, which became extremely popular, accumulating over 700 thousand likes and 500 thousand shares (Table 6.5).

**Donald Trump's insensitive tweet:** An old tweet from January, 2015 posted by the U.S.

<sup>16</sup><http://www.bbc.com/news/blogs-trending-34845882>

<sup>17</sup><http://www.toureiffel.paris/en/everything-about-the-tower/the-illuminations.html>



(a) Diesel the dog who was allegedly killed by terrorists



We've all seen video of the France vs. Germany soccer game when the sound of an explosion can be heard causing the players to look at each other in curiosity and confusion. That explosion took place OUTSIDE the Stade de France. It was supposed to take place INSIDE and potentially kill hundreds of people including French President Francois Hollande. But it didn't.

It didn't because a security guard detected the bomber's vest and confronted him. The bomber then detonated his vest OUTSIDE the stadium. Everyone INSIDE was safe.

That security guard who saved so many hundreds, if not thousands of people, including the President of France himself, has been completely ignored and only his first name has been reported only a handful of times.

His name is Zouheir. He is a Muslim.

No one cares.

(b) Zouheir, the security guard who was claimed to have stopped a terrorist from entering the stadium



(c) Image of Donald Trump's insensitive tweet



(d) Eiffel Tower goes dark for the first time in years



(e) Screenshot of tweet predicting death toll before the attack took place

Figure 6.9: Rumors spread on Facebook in the form of images, during the Paris Attacks in 2015. We used modern image analysis techniques to identify these rumors.

presidential candidate Donald Trump<sup>18</sup> resurfaced in the form of an image just after the Paris Attacks, and received a lot of backlash and criticism from the public (Figure 6.9(c)). Large section of users on Facebook mistook the tweet as Donald Trump's reaction to the attacks that took place in November, more than 10 months after the tweet was posted. Since the tweet was being popularized

<sup>18</sup><https://twitter.com/realDonaldTrump/status/552955167533174785>

Table 6.5: Statistical summary of all rumors we identified. Table captures, for each rumor separately, number of images in our dataset, number of distinct images, count of Facebook users / pages who posted these images, and sum total of all *likes*, *comments*, and *shares* on these images on Facebook.

Rumor	Instances in our dataset	Total distinct images	Posted by	Likes	Comments	Shares
Diesel the dog	170	5	12	1,101,115	38,291	321,706
Zouheir, security guard	251	17	35	21,944	450	11,207
Eiffel Tower lights out	24	6	14	727,977	5,581	566,565
Donald Trump’s tweet	42	2	3	1,311	77	284
@PZBooks tweet predicting death toll	21	8	9	952	136	1,109

in the form of an image, we used OCR to extract the text from the image to identify this rumor. Although the rumor wasn’t as popular as the ones previously discussed, it still managed to get over 1,000 engagements (Table 6.5).

**Tweet predicts death toll from Paris attacks before they occur:** Twitter bot @PZbooks posted a tweet which said, “*BREAKING: Death toll from Paris terror attack rises to at least 120 with 270 others injured*” on November 11, 2015, two days before the Paris attacks took place. As the attacks took place on November 13, this tweet became popular and Internet users around the world were startled by this “prediction”. Multiple users took to OSNs like Facebook and Twitter to voice their concerns, with some even labelling it a ‘false flag’ – a term used by conspiracy theorists for an act they think has been staged by a government or other authority.

It was later clarified that this Twitter account (@PZbooks) was a spambot that regurgitates actual news from @PZF – which is an actual breaking news account. In a bizzare co-incidence, @PZBooks appeared to have taken two official @PZF breaking news tweets – one dating back to January’s Charlie Hebdo attack in Paris: “*BREAKING NEWS: Death toll from Paris terror attack rises to at least 12 - Reuters*”<sup>19</sup> and the other from an attack on a Nigerian mosque in November 2014: “*AFP: Death toll from Nigerian mosque attack rises to at least 120 with 270 others injured*”<sup>20</sup>, and combined them to create what appeared to be a “tweet from the future”.

Images and screenshots of this tweet (Figure 6.9(e)) made rounds on Facebook, luring users into believing that the attacks were staged. Again, we were able to identify such images using OCR.

<sup>19</sup><https://twitter.com/pzf/status/552801974991998976>

<sup>20</sup><https://twitter.com/pzf/status/538422123094888448>

## 6.6 Discussion, Limitations, and Future work

We understand that the accuracy of our image labeling and sentiment detection models is limited. However, these models are trained using CNNs, which form the basis for one of the most advanced state-of-the-art techniques for learning features from visual media in the modern age. These models can be further improved by feeding them true positive datasets of images. Generating a big enough dataset for such models is, however, a challenging task, and out of the scope of this work. Moreover, validation through manual inspection and human annotations revealed that our models sufficed for producing high-level summaries for images, which was the primary objective of this work.

Text extracted using Tesseract is limited by the performance of modern OCR techniques. We came across instances in our dataset where we were manually able to recognize text, but the OCR failed to identify this text. Most such instances involved the presence of calligraphic text, and noisy background. The percentage of such instances was low as compared to the number of images for which we were able to identify and extract text.

Images are an integral part of OSN content and are naturally more appealing to the human mind than text. OSN statistics explicitly support this phenomenon and makes images an interesting avenue for researchers to explore. In this chapter, we discovered a new dimension of content on Facebook through images. This study highlights the vast amount of information that can be mined from images alone, by making use of some modern image analysis techniques. We also highlight how this information may be different from textual content that has been previously used in literature to infer users' pulse and sentiment.

Various researchers have used text to measure the sentiment and mood of users in diverse contexts like natural calamities, politics, sports, etc. With the large volume and popularity of images on OSNs in recent times, it is imminent that results drawn from text alone fail to capture the pulse of the audience accurately. We believe that the results drawn from past studies can be improved by taking visual content into consideration.

Brands and organizations invest heavily in social media marketing and rely on textual responses generated by users to gauge their reactions, and in turn, the performance of their products. Being able to understand the users' pulse through images is likely to help such organizations measure the response to their products much better, and cover a larger section of the audience. Moreover, while analyzing sentiment and emotion through text is largely limited by language, such a barrier does not exist for images.

Pictures posted on OSNs can be a critical source of information for law and order organizations to understand public sentiment, especially during crisis events. With the enormous volume and velocity of data being generated on OSNs, it is extremely tough to monitor visual media at present because of lack of automated methods for measurement and analysis of such content. Our proposed

pipeline can be used in such scenarios for mining knowledge from visual content and identifying popular themes and citizens' pulse during crisis events. Although this methodology has its limitations, it can be very effective for producing high-level summaries and reducing the search space for organizations in terms of content that may need attention. We also described how our methodology can be used for identifying popular instances of misinformation spread through images, which may lead to major implications in the real world.

## Chapter 7

# Analyzing Social and Stylometric Features to Identify Spear phishing Emails

This chapter is joint work with Anand Kashyap (Lead Security Researcher at Symantec between 2009-2015) and is largely a reproduction of a paper published at IEEE APWG eCrime Research Summit (eCRS) 2014 [35].

### 7.1 Introduction

In the previous chapters, we focussed on automatically identifying poor quality context-specific content in scenarios where the attacker and victim are part of the same network ecosystem (Facebook). The attacker generated poor quality content in the form of misinformation, rumors, hoaxes, scams, etc. and exploited the context of a news-making event to lure victims into consuming this attacker-generated content. However, this may not always be the case. In this chapter, we explore a scenario where the attacker uses information from one social networking platform to attack a victim on a different platform, viz. email (Figure 7.1). The poor quality context-specific content in this case are targeted spear-phishing emails which may have been crafted by the attacker by utilizing information available about the victim on a social networking platform, viz. LinkedIn. Through this chapter, we take a step towards analyzing whether social features of the victim can be utilized to better identify targeted spear phishing emails sent to her.

A new race of insidious threats called Advanced Persistent Threats (APTs) have joined the league of eCrime activities on the Internet, and caught a lot of organizations off guard in the fairly recent times. Critical infrastructures and the governments, corporations, and individuals supporting them



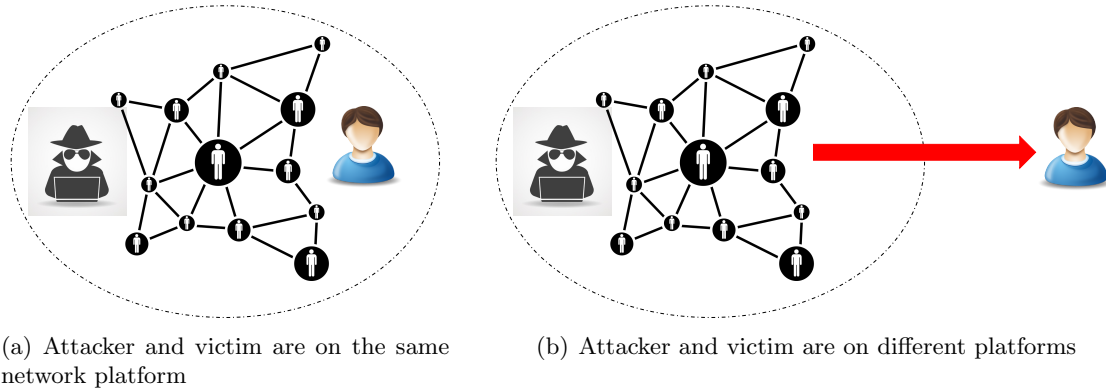


Figure 7.1: Attack scenario where attacker leverages context-specific information from one platform (social network) to attack victim with poor quality content (targeted spear-phishing) on another platform.

are under attack by these increasingly sophisticated cyber threats. The goal of the attackers is to gain access to intellectual property, personally identifiable information, financial data, and targeted strategic information. This is not simple fraud or hacking, but intellectual property theft and infrastructure corruption on a grand scale [30]. APTs use multiple attack techniques and vectors that are conducted by stealth to avoid detection, so that hackers can retain control over target systems unnoticed for long periods of time. Interestingly, no matter how sophisticated these attack vectors may be, the most common ways of getting them inside an organization’s network are social engineering attacks like phishing, and targeted spear phishing emails. There have been numerous reports of spear phishing attacks causing losses of millions of dollars in the recent past.<sup>1 2</sup> Although there exist antivirus, and other similar protection software to mitigate such attacks, it is always better to stop such vectors at the entry level itself [94]. This requires sophisticated techniques to deter spear phishing attacks, and identify malicious emails at a very early stage itself.

In this chapter, we focus on identifying such spear phishing emails, wherein the attacker targets an individual or company, instead of anyone in general. According to Trend Micro, spear phishing is a phishing method that targets specific individuals or groups within an organization.<sup>3</sup> Spear phishing emails usually contain victim-specific context instead of general content. Since it is targeted, a spear phishing attack looks much more realistic, and thus, harder to detect [83]. A typical spear phishing attack can broadly be broken down into two phases. In the first phase, the attacker tries to gather as much information about the victim as possible, in order to craft a scenario which looks realistic, is believable for the victim, and makes it very easy for the attacker to attain the victim’s trust. In the second phase, the attacker makes use of this gained trust, and draws the

<sup>1</sup><http://businesstech.co.za/news/internet/56731/south-africas-3-billion-phishing-bill/>  
<sup>2</sup><http://www.scmagazine.com/stolen-certificates-used-to-deliver-trojans-in-spear-phishing-campaign/article/345626/>  
<sup>3</sup><https://www.trendmicro.com/vinfo/us/security/definition/spear-phishing>

victim into giving out sensitive / confidential information like a user name, password, bank account details, credit card details, etc. The attacker can also exploit the victim's trust by infecting the victim's system, through luring them into downloading and opening malicious attachments [83]. While spear phishing may be a timeworn technique, it continues to be effective even in today's Web 2.0 landscape. A very recent example of such a spear phishing attack was reported by FireEye. Here, attackers exploited the news of the disappearance of Malaysian Airlines Flight MH370, to lure government officials across the world into opening malicious attachments (Figure 7.2) sent to them over email [109]. In 2011, security firm RSA suffered a breach via a targeted attack; analysis revealed that the compromise began with the opening of a spear phishing email. <sup>4</sup> That same year, email service provider Epsilon also fell prey to a spear phishing attack that caused the organization to lose an estimated US\$4 billion. <sup>5</sup> These examples indicate that spear phishing has been, and continues to be one of the biggest forms of eCrime over the past few years, especially in terms of the monetary losses incurred.

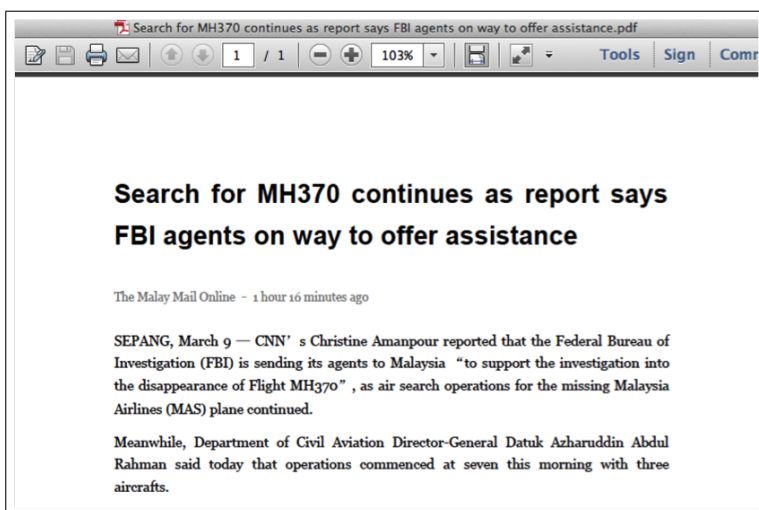


Figure 7.2: Example of a malicious PDF attachment sent via a spear phishing email. The PDF attachment was said to contain information about the missing Malaysian Airlines Flight 370.

Spear phishing was first studied as context aware phishing by Jakobsson et al. in 2005 [82]. A couple of years later, Jagatic et al. performed a controlled experiment and found that the number of victims who fell for context aware phishing / spear phishing is 4.5 times the number of victims who fell for general phishing [81]. This work was preliminary proof that spear phishing attacks have a much higher success rate than normal phishing attacks. It also highlighted that, what separates a regular phishing attack from a spear phishing attack is the additional information / context. Online social media services like LinkedIn, which provide rich professional information about individuals,

<sup>4</sup><http://blogs.rsa.com/rivner/anatomy-of-an-attack/>

<sup>5</sup><http://www.net-security.org/secworld.php?id=10966>

can be one such source for extracting contextual information, which can be used against a victim. In 2013, the Federal Bureau of Investigation (FBI) also warned that spear phishing emails typically contain accurate information about victims often obtained from data posted on social networking sites, blogs or other websites.<sup>6</sup>

In this work, we investigate if publicly available information from LinkedIn can help in differentiating a spear phishing from a regular phishing email received by an individual. The attack model in this work consists of individuals (victims) who received one or more spear phishing emails on their official email address between March 2009 and December 2013, and had a user profile on the LinkedIn social network. Publicly available information from this LinkedIn profile of the victims may or may not have been used to craft the targeted spear phishing email that they received. We attained a dataset of true positive targeted spear phishing emails, and a dataset of a mixture of non targeted, spam and phishing attack emails from the Symantec’s enterprise email scanning service, which is deployed at multiple international organizations around the world. To conduct the analysis at the organizational level, we extracted the most frequently occurring domains from the “to” fields of these emails, and filtered out 14 most targeted organizations, where the first name and last name could be derived from the email address. The availability of the first name and last name from the email address was crucial since this information was necessary to identify and pin-point the victims’ LinkedIn profiles (we discuss this in more detail in Section 7.3.3). Our final dataset consisted of 4,742 spear phishing emails sent to 2,434 unique employees, and 9,353 non targeted spam / phishing emails to 5,914 unique employees. For a more exhaustive analysis, we also used a random sample of 6,601 benign emails from the Enron email dataset [29] sent to 1,240 unique employees with LinkedIn profiles.

We applied 4 classification algorithms, and were able to achieve a maximum accuracy of 97.04% for classifying spear phishing, and non spear phishing emails using a combination of *email* features, and *social* features. However, without the *social* features, we were able to achieve a slightly higher accuracy of 98.28% for classifying these emails. We then looked at the most informative features, and found that *email* features performed better than *social* features at differentiating targeted spear phishing emails from non targeted spam / phishing emails, and benign Enron emails. To the best of our knowledge, this is the first attempt at making use of a social media profile of a user to distinguish targeted spear phishing emails from non targeted attack emails, received by her. Having found that *social* features extracted from LinkedIn profiles do not help in distinguishing spear phishing and non spear phishing emails, our results encourage to explore other social media services like Facebook, and Twitter. Such studies can be particularly helpful in mitigating APTs, and reducing the chances of attacks to an organization at the entry level itself.

The rest of the chapter is arranged as follows. Section 7.2 discuss the related work, Section 7.3

---

<sup>6</sup><http://www.computerweekly.com/news/2240187487/FBI-warns-of-increased-spear-phishing-attacks>

describes our email and LinkedIn profile datasets, and the data collection methodology. The analysis and results are described in Section 7.4. We conclude our findings, and discuss the limitations, contributions, and scope for future work in Section 7.5.

## 7.2 Background and Related work

The concept of targeted phishing was first introduced in 2005 as *social phishing* or *context-aware phishing* [82]. Authors of this work argued that if the attacker can infer or manipulate the context of the victim before the attack, this context can be used to make the victim volunteer the target information. This theory was followed up with an experiment where Jagatic et al. harvested freely available acquaintance data of a group of Indiana University students, by crawling social networking websites like Facebook, LinkedIn, MySpace, Orkut etc. [81]. This contextual information was used to launch an actual (but harmless) phishing attack targeting students between the age group of 18-24 years. Their results indicated about 4.5 times increase in the number of students who fell for the attack which made use of contextual information, than the generic phishing attack. However, authors of this work do not provide details of the kind and amount of information they were able to gather from social media websites about the victims.

**Who falls for phish** Dhamija et al. provided the first empirical evidence about which malicious strategies are successful at deceiving general users [39]. Kumaraguru et al. conducted a series of studies and experiments on creating and evaluating techniques for teaching people not to fall for phish [92–94]. Lee studied data from Symantec’s enterprise email scanning service, and calculated the odds ratio of being attacked for these users, based on their area of work. The results of this work indicated that users with subjects “*Social studies*”, and “*Eastern, Asiatic, African, American and Australasian Languages, Literature and Related Subjects*” were both positively correlated with targeted attacks at more than 95% confidence [96]. Sheng et al. conducted an online survey with 1,001 participants to study who is more susceptible to phishing based on demographics. Their results indicated that women are more susceptible than men to phishing, and participants between the ages 18 and 25 are more susceptible to phishing than other age groups [140]. In similar work, Halevi et al. found a strong correlation between gender and response to a prize phishing email. They also found that neuroticism is the factor most correlated to responding to the email. Interestingly, authors detected no correlation between the participants’ estimate of being vulnerable to phishing attacks and actually being phished. This suggests susceptibility to phishing is not due to lack of awareness of the phishing risks, and that real-time response to phishing is hard to predict in advance by online users [71].

**Phishing email detection techniques** To keep this work focused, we concentrate only on techniques proposed for detecting phishing emails; we do not cover all the techniques used for detecting phishing URLs or phishing websites in general. Abu-Nimeh et al. [1] studied the performance of different classifiers used in text mining such as Logistic regression, classification and regression trees, Bayesian additive regression trees, Support Vector Machines, Random forests, and Neural networks. Their dataset consisted of a public collection of about 1,700 phishing mails, and 1,700 legitimate mails from private mailboxes. They focused on richness of word to classify phishing email based on 43 keywords. The features represent the frequency of “bag-of-words” that appear in phishing and legitimate emails. However, the ever-evolving techniques and language used in phishing emails might make it hard for this approach to be effective over a long period of time.

Various feature selection approaches have also been recently introduced to assist phishing detection. A lot of previous work [1,23,54] has focused on email content in order to classify the emails as either benign or malicious. Chandrasekaran et al. [23] presented an approach based on natural structural characteristics in emails. The features included number of words in the email, the vocabulary, the structure of the subject line, and the presence of 18 keywords. They tested on 400 data points which were divided into five sets with different type of feature selection. Their results were the best when more features were used to classify phishing emails using Support Vector Machine. Authors of this work proposed a rich set of stylometric features, but the dataset they used was very small as compared to a lot of other similar work. Fette et al. [54] on the other hand, considered 10 features which mostly examine URL and presence of JavaScript to flag emails as phishing. Nine features were extracted from the email and the last feature was obtained from WHOIS query. They followed a similar approach as Chandrasekaran et al. but using larger datasets, about 7,000 normal emails and 860 phishing emails. Their filter scored 97.6% F-measure, false positive rate of 0.13% and a false negative rate of 3.6%. The heavy dependence on URL based features, however, makes this approach ineffective for detecting phishing emails which do not contain a URL, or are attachment based attacks, or ask the user to reply to the phishing email with potentially sensitive information. URL based features were also used by Chhabra et al. to detect phishing using short URLs [27]. Their work, however, was limited to only URLs, and did not cover phishing through emails. Islam and Abawajy [80] proposed a multi-tier phishing detection and filtering approach. They also proposed a method for extracting the features of phishing email based on weights of message content and message header. The results of their experiments showed that the proposed algorithm reduces the false positive problems substantially with lower complexity.

Behavior-based approaches have also been proposed by various researchers to determine phishing messages [152,173]. Zhang et al. [173] worked on detecting abnormal mass mailing hosts in network layer by mining the traffic in session layer. Toolan et al. [152] investigated 40 features that have been used in recent literature, and proposed behavioral features such as number of words in *sender* field, total number of characters in *sender* field, difference between sender’s domain and reply-to

domain, and difference between sender’s domains and the email’s modal domain, to classify ham, spam, and phishing emails. Ma et al. [102] attempted to identify phishing email based on hybrid features. They derived 7 features categorized into three classes, i.e. content features, orthographic features, and derived features, and applied 5 machine learning algorithms. Their results stated that Decision Trees worked best in identifying phishing emails. Hamid et al. [73] proposed a hybrid feature selection approach based on combination of content and behaviour. Their approach mined attacker behavior based on email header, and achieved an accuracy of 94% on a publicly available test corpus.

All of the aforementioned work concentrates on distinguishing phishing emails from legitimate ones, using various types of features extracted from email content, URLs, header information etc. To the best of our knowledge, there exists little work which focuses specifically on targeted spear phishing emails. Further, there exists no work which utilizes features from the social media profiles of the victim in order to distinguish an attack email from a legitimate one. In this work, we direct our focus on a very specific problem of distinguishing targeted spear phishing emails from general phishing, spam, and benign emails. Further, we apply *social* features extracted from the LinkedIn profile of recipients of such emails to judge whether an email is a spear phishing email or not; which has never been attempted before, to the best of our knowledge. We performed our entire analysis on a real-world dataset derived from Symantec’s enterprise email scanning service.

## 7.3 Data collection methodology

The dataset we used for the entire analysis, is a combination of two separate datasets, viz. a dataset of emails (combination of targeted attack and non targeted attack emails), and a dataset of LinkedIn profiles. We now explain both these datasets in detail.

### 7.3.1 Email dataset

Our email dataset consisted of a combination of targeted spear phishing emails, non targeted spam and phishing emails, and benign emails. We obtained the targeted spear phishing emails from Symantec’s enterprise email scanning service. *Symantec* collects data regarding targeted attacks that consist of emails with malicious attachments. These emails are identified from the vast majority of non-targeted malware by evidence of there being prior research and selection of the recipient, with the malware being of high sophistication and low copy number. The process by which Symantec’s enterprise mail scanning service collects such malware has already been described elsewhere [97,151]. The corpus almost certainly omits some attacks, and most likely also includes some non-targeted attacks, but nevertheless it represents a large number of sophisticated targeted attacks compiled according to a consistent set of criteria which render it a very useful dataset to

Table 7.1: Top 20 most frequently occurring attachment names, and their corresponding percentage share in our spear phishing and spam / phishing datasets. Attachment names in the spear phishing emails dataset look much more realistic and genuine as compared to attachment names in spam / phishing emails dataset.

Spear phishing Attachment Name	%	Spam / phishing Attachment Name	%
work.doc	3.46	100A_0.txt	20.74
More detail Chen Guangcheng.rar	3.01	100_5X_AB_PA1_MA-OCTET-STREAM_..form.html	9.02
ARMY_600_8_105.zip	2.54	./attach/100_4X_AZ-D_PA2_..FedEx=5FInvoice=5FN56=2D141.exe	4.19
Strategy_Meeting.zip	1.58	100_2X_PM3_EMS_MA-OCTET=2DSTREAM_..apply.html	2.66
20120404 H 24 year annual business plan 1 quarterly.zip	1.33	100_4X_AZ-D_PA2_..My=5Fsummer=5Fphotos=5Fin=5FEgypt=5F2011.exe	1.87
The extension of the measures against North Korea.zip	1.30	./attach/100_2X_PM2_EMS_MA-OCTET=2DSTREAM_..ACC01291731.rtf	1.40
Strategy_Meeting_120628.zip	1.28	100_5X_AB_PA1_MH_..NothernrockUpdate.html	1.28
image.scr	1.24	./attach/100_2X_PM2_EMS_MA-OCTET=2DSTREAM_..invoice.rtf	1.15
Consolidation Schedule.doc	0.98	100_6X_AZ-D_PA4_..US=2DCERT=20Operations=20Center=20Report=2DJan2012.exe	1.12
DARPA-BAA-11-65.zip	0.93	100_4X_AZ-D_PA2_..I=27m=5Fwith=5Fmy=5Ffriends=5Fin=5FEgypt.exe	1.11
Head Office-Y drive.zip	0.93	100_4X_AZ-D_PA2_..I=27m=5Fon=5Fthe=5FTurkish=5Fbeach=5F2012.exe	0.80
page 1-2.doc	0.90	100_5X_AB_PA1_MA-OCTET-STREAM_..Lloyds=R01TSB=R01- =R01Login=R01Form.html	0.69
Aircraft Procurement Plan.zip	0.90	100_6X_AZ-D_PA4_..Fidelity=20Investments=20Review=2Dfrom=2DJan2012.exe	0.68
Overview of Health Reform.doc	0.74	100_4X_AZ-D_PA2_..FedEx=5FInvoice=20=5FCopy=5FIDN12=2D374.exe	0.64
page 1-2.pdf	0.64	100_4X_AZ-D_PA2_..my=5Fphoto=5Fin=5Fthe=5Fdominican=5FRepublic.exe	0.63
fisa.pdf	0.58	100_2X_PM4_EMQ_MH_..message.htm	0.60
urs.doc	0.52	/var/work0/attach/100_4X_AZ-D_PA2_..document.exe	0.58
script.au3	0.50	100_6X_AZ-D_PA4_..Information.exe	0.58
install_reader10_en_air_gtbd_aih.zip	0.48	/var/work0/attach/100_4X_AZ-D_PA2_..Ticket.exe	0.57
dodi-3100-08.pdf	0.43	100_4X_AZ-D_PA2_..Ticket.exe	0.57

study.

The non targeted attack emails were also obtained from Symantec’s email scanning service. These emails were marked as *malicious*, and were a combination of malware, spam, and phishing. Both these datasets contained an enormously large number of emails received at hundreds of organizations around the world, where Symantec’s email scanning services are being used. Before selecting a suitable sample for organization level analysis, we present an overview of this entire data. Table 7.1 shows the top 20 most frequently occurring attachment names in the complete spear phishing and spam / phishing datasets. We found distinct differences in the type of attachment names in these two datasets. While names in spear phishing emails looked fairly genuine and personalized, attachment names in spam / phishing emails were irrelevant, and long. It was also interesting to see that the attachment names associated with spear phishing emails were less repetitive than those associated with spam / phishing emails. As visible in Table 7.1, the most commonly occurring attachment name in spear phishing emails was found in less than 3.5% of all spear phishing emails, while in the case of spam / phishing emails, the most common attachment name was present in over 20% of all spam / phishing emails. This behavior reflects that attachments in spear phishing emails are named more carefully, and with more effort to make them look genuine.

We also looked at the most frequently spread file types in spear phishing, spam, and phishing emails. Table 7.2 shows the top 15 most frequently occurring file types in both the spear phishing and spam / phishing email datasets. Not surprisingly, both these datasets had notable presence of executable (.exe, .bat, .com), and compressed (.rar, .zip, .7z) file types. In fact, most of the file types spread through such emails were among the most frequently used file types in general, too. Microsoft Word, Excel, PowerPoint, and PDF files were also amongst the most frequently spread

files. It was, however, interesting to note that lesser percentage of targeted spear phishing emails contained executables than spam / phishing emails. This reflects that attackers prepare targeted attacks smartly as compared to spammers / phishers, and avoid using executables, which are more prone to suspicion.

Table 7.2: Top 15 most frequently occurring attachment types, and their corresponding percentage share in our spear phishing and spam / phishing datasets. Only 5 file types were common in the top 15 in these datasets.

Spear phishing Attachment Type	%	Spam / phishing Attachment Type	%
Zip archive data (zip)	19.59	Windows Executable (exe)	38.39
PDF document (pdf)	13.73	ASCII text (txt)	21.73
Composite Document File	13.63	Hypertext (html)	18.08
Windows Executable (exe)	11.20	Hypertext (htm)	7.06
Rich Text Format data (rtf)	10.40	Rich Text Format data (rtf)	3.04
RAR archive data (rar)	9.47	PDF document (pdf)	2.04
Screensaver (scr)	5.06	Zip archive data (zip)	1.75
data (dat)	3.00	Microsoft Word	1.27
JPEG image data (jpg)	1.64	Screensaver (scr)	1.14
CLASS	1.56	Microsoft Excel (xls)	0.81
Microsoft Word 2007+	1.15	Program Info. file (pif)	0.80
7-zip archive data (7z)	1.12	Dynamic-link Library (dll)	0.30
Shortcut (lnk)	1.08	Windows Batch file (.bat)	0.24
ASCII text (txt)	0.80	JavaScript (js)	0.17
Dynamic-link Library (dll)	0.54	Microsoft HTML Help (chm)	0.16

All the emails present in our full dataset were collected over a period of slightly under 4 years, from March 2009 to December 2013. Figure 7.3 presents a time line of the “received time” of all these emails. The spam / phishing emails were collected over a period of 3 years, from March 2009 to March 2012. The targeted spear phishing emails were also collected during a period of about 3 years, but from January 2011, to December 2013. The two datasets, thus, had data for a common time period of about 15 months, from January 2011, to March 2012. It was interesting to observe that during this period, while the number of spam and phishing emails saw a tremendous rise, the number of spear phishing emails did not vary too much. This characteristic was observed for the entire 3-year period for spear phishing emails. The number of spear phishing emails received in the beginning and end of our three year observation period saw a 238% rise, as compared to a rise of 35,422% percent in the number of spam / phishing emails.

In addition to the attachment information and time line, we also looked at the “subject” fields of all the emails present in our datasets. Table 7.3 shows the top 20 most frequently occurring “subject lines” in our datasets. Evidently, “subjects” in both these datasets were very different in terms of context. Targeted spear phishing email subjects seemed to be very professional, talking about jobs, meetings, *unclassified* information etc. Spam / phishing email subjects, however, were observed to follow a completely different genre. These emails’ subjects were found to follow varied themes, out



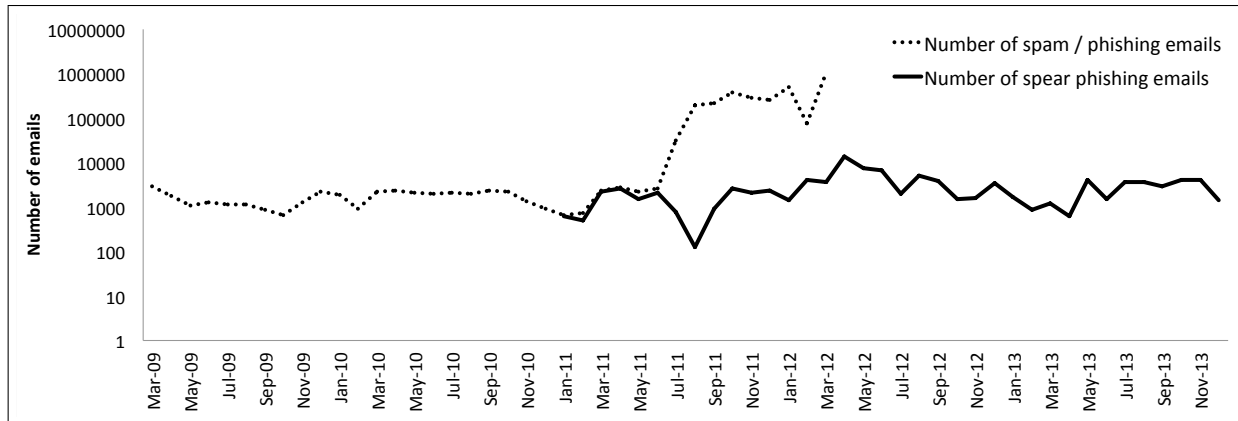


Figure 7.3: Time line of the number of spear phishing and spam / phishing emails in our dataset. The X axis represents time, and the Y axis represents the number of emails on a logarithmic scale. The period of May 2011 - September 2011 saw an exponential increase in the number of spam / phishing emails in our dataset.

of which, three broad themes were fairly prominent: a) fake email delivery failure error messages, which lure victims into opening these emails to see which of their emails failed, and why; b) arrival of packages or couriers by famous courier delivery services – victims tend to open such messages out of curiosity, even if they are not expecting a package; and c) personalized messages via third party websites and social networks (Hallmark E-Card, hi5 friend request, and Facebook message in this case). Most of such spam / phishing emails have generic subjects, to which most victims can relate easily, as compared to spear phishing email subjects, which would seem irrelevant to most common users.

It is important to note that these statistics are for the complete datasets we obtained from Symantec. The total number of emails present in the complete dataset was of the order of hundred thousands. However, we performed our entire analysis on a sample picked from this dataset. The analysis in the rest of the chapter talks about only this sample. To make our analysis more exhaustive, we also used a sample of benign emails from the Enron email dataset for our analysis [29]. All the three email datasets had duplicates, which we identified and removed by using a combination of 5 fields, viz. *from ID*, *to ID*, *subject*, *body*, and *timestamp*. On further investigation, we found that these duplicate email messages were different instances of the same email. This happens when an email is sent to multiple recipients at the same time. A globally unique *message-id* is generated for each recipient, and thus results in duplication of the message. Elimination of duplicates reduced our email sample dataset by about 50%. Our final sample email dataset that we used for all our analysis was, therefore, a mixture of targeted attack emails, non targeted attack emails, and benign emails. We now describe this sample.

Table 7.3: Top 20 most frequently occurring subjects, and their corresponding percentage share in our spear phishing, and spam / phishing email datasets. Spear phishing email subjects appear to depict that these emails contain highly confidential data. Spam / phishing emails, on the other hand, are mainly themed around email delivery error messages, and courier or package receipts.

Spear phishing subjects	%	Spam / phishing subjects	%
Job Opportunity	3.45	Mail delivery failed: returning message to sender	10.95
Strategy Meeting	3.09	Delivery Status Notification (Failure)	6.71
What is Chen Guangcheng fighting for?	3.00	Re:	2.59
FW: [2] for the extension of the measures against North Korea	1.70	Re	2.56
[UNCLASSIFIED] 2012 U.S.Army orders for weapons	1.27	Become A Paid Mystery Shopper Today! Join and Shop For Free!	1.28
FW:[UNCLASSIFIED]2012 U.S.Army orders for weapons	1.27	failure notice	1.09
<blank subject line>	1.17	Delivery Status Notification (Delay)	1.06
FW: results of homemaking 2007 annual business plan (min quarter 1 included)	1.02	Returned mail: see transcript for details	0.95
[UNCLASSIFIED]DSO-DARPA-BAA-11-65	0.93	Get a job as Paid Mystery Shopper! Shop for free and get Paid!	0.85
Wage Data 2012	0.90	Application number: AA700003125331	0.82
U.S.Air Force Procurement Plan 2012	0.90	Your package is available for pickup	0.78
About seconded expatriate management in overseas offices	0.80	Your statement is ready for your review	0.75
FW:[CLASSIFIED] 2012 USA Government of the the Health Reform	0.74	Unpaid invoice 2913.	0.71
T.T COPY	0.62	Track your parcel	0.70
USA to Provide Declassified FISA Documents	0.58	You have received A Hallmark E-Card!	0.59
FY2011-12 Annual Merit Compensation Guidelines for Staff	0.55	Your Account Opening is completed.	0.57
Contact List Update	0.45	Delivery failure	0.57
DOD Technical Cooperation Program	0.43	Undelivered Mail Returned to Sender	0.56
DoD Protection of Whistleblowing Spies	0.43	Laura would like to be your friend on hi5!	0.56
FW:UK Non Paper on arrangements for the Arms Trade Treaty (ATT) Secretariat	0.42	You have got a new message on Facebook!	0.55

### 7.3.2 Email Sample Dataset Description

To focus our analysis at the organization level, we identified and extracted the most attacked organizations (excluding free email providing services like Gmail, Yahoo, Hotmail etc.) from the domain names of the victims' email addresses, and picked 14 most frequently attacked organizations. We were however, restricted to pick only those organizations, where the first names and last names were easily extractable from the email addresses. The first name and last name were required to obtain the corresponding LinkedIn profiles of these victims (this process is discussed in detail in Section 7.3.3). This restriction, in addition to removal of duplicates, left us with a total of 4,742

targeted spear phishing emails sent to 2,434 unique victims (referred to as *SPEAR* in the rest of the chapter); 9,353 non targeted attack emails sent to 5,912 unique non victims (referred to as *SPAM* in the rest of the chapter), and 6,601 benign emails from the Enron dataset, sent to 1,240 unique Enron employees (referred to as *BENIGN* in the rest of the chapter). Further details of this dataset can be found in Table 7.4. Table contains the number of victims, and non victims in each of the 15 companies (including Enron), and the number of emails sent to them. The victim and non victim employee sets are mutually exhaustive, and each employee in these datasets received at least one email, and had at least one LinkedIn profile. To maintain anonymity, we do not include the name of the organizations we picked; we only mention the operation sector of these companies.

Table 7.4: Detailed description of our dataset of LinkedIn profiles and emails across 15 organizations including Enron.

Sector	#Victims	#Emails	#Non Victims	#Emails	No. of Employees
Govt. & Diplomatic	206	511	572	1,103	10,001+
Info. & Broadcasting	150	326	240	418	10,001+
NGO	131	502	218	472	1001-5000
IT/Telecom/Defense	158	406	68	157	1001-5000
Pharmaceuticals	120	216	589	862	10,001+
Engineering	396	553	1000	1,625	10,001+
Automotive	153	601	891	1,204	10,001+
Aviation/Aerospace	281	355	161	187	1001-5000
Agriculture	94	138	173	264	10,001+
IT & Telecom	11	12	543	943	5001-10,000
Defense	388	651	123	147	10,001+
Oil & energy	201	212	680	1,017	10,001+
Finance	89	129	408	608	10,001+
Chemicals	56	130	248	346	10,001+
Enron	NA	NA	1,240	6,601	10,001+
Total	2,434	4,742	7,154	15,954	

Figures 7.4(a), 7.4(c), and 7.4(d) represent the tag clouds of the 100 most frequently occurring words in the “subject” fields of our *SPEAR*, *SPAM*, and *BENIGN* datasets respectively. We noticed considerable differences between subjects from all the three datasets. While all three datasets were observed to have a lot of *forwarded* emails (represented by “fw”, and “fwd” in the tag clouds), *SPAM* and *BENIGN* datasets were found to have much more *reply* emails (signified by “re” in the tag clouds) as compared to *SPEAR* emails. These characteristics of whether an email is forwarded, or a reply, have previously been used as boolean features by researchers to distinguish between phishing and benign emails [152]. The difference in vocabularies used across the three email datasets is also notable. The *SPEAR* dataset (Figure 7.4(a)) was found to be full of attention-grabbing words like *strategy*, *unclassified*, *warning*, *weapons*, *defense*, *US Army* etc. Artifact 7.5 shows an example of

ARTIFACT 7.5: A spear phishing email from our SPEAR dataset. The email shows a seemingly genuine conversation, where the attacker sent a malicious compressed (.rar) attachment to the victim in the middle of the conversation.

<b>Attachment:</b> All information about mobile phone.rar
<b>Subject:</b> RE: Issues with Phone for help
<b>Body:</b> <name>, Thanks for your replying.I contacted my supplier,but he could not resolved it.Now I was worried, so I take the liberty of writing to you.I collect all information including sim card details,number,order record and letters in the txt file.I hope you can deal with the issues as your promised. Best, <name>  -----Original Message----- From: Customer Care [mailto:Customer_Care@<companyDomain>] Sent: 2011-12-8 0:35 To: <name> Cc: Subject: RE: Issues with Phone for help  Dear <name>,  Thank you for your E-mail. I am sorry to hear of your issues. Please can you send your SIM card details or Mobile number so that we can identify your supplier who can assist you further?  Thank you  Kind regards,  <name> Customer Service Executive  <Company Name>, <Company Address> United Kingdom  Tel: <telephone number> Fax : <Fax number> <company website>  -----Original Message----- From: <name> [mailto:<email address>] Sent: 08 December 2011 08:27 To: support@<companyDomain> Subject: Issues with Phone for help  Hello, I purchased order for your IsatPhone Pro six months ago.Now I have trouble that it can't work normally.It often automatic shuts down.Sometimes it tells some information that i can't understand.How to do?Can you help me? Best, <name>  ----- This e-mail has been scanned for viruses by Verizon Business Internet Managed Scanning Services - powered by MessageLabs. For further information visit <a href="http://www.verizonbusiness.com/uk">http://www.verizonbusiness.com/uk</a>

the attachment name, subject and body of such an email. We removed the received address and other details to maintain anonymity.

SPAM emails in our dataset (Figure 7.4(c)) followed a completely different genre, dominated by words like *parcel*, *order*, *delivery*, *tracking*, *notification*, *shipment* etc. We also found mentions of famous courier service brand names like FedEx and DHL which seem to have been used for targeting victims. Such attacks have been widely talked about in the recent past; users have also been warned about scams, and infected payloads (like spyware or malware), that accompany such emails.<sup>7 8</sup> Some examples of attachment names, and subjects of such non targeted SPAM emails are shown in Artifact 7.6. BENIGN subjects comprised of diverse keywords like *report*, *program*,

<sup>7</sup><http://nakedsecurity.sophos.com/2013/03/20/dhl-delivery-malware/>

<sup>8</sup><http://www.spamfighter.com/News-13360-FedEx-and-DHL-Spam-Attack-with-Greater-Ferocity.htm>

*meeting, migration, energy*, which did not seem specific to a particular theme (Figure 7.4(d)). These keywords were fairly representative of the kind of typical internal communication that may have been going on in the company.

ARTIFACT 7.6: Examples of *subject* and *attachment* names of two spam emails from our SPAM dataset. The *body* field of the emails was not available in this dataset.

<b>Attachment:</b> 100A_0.txt
<b>Subject:</b> DHL Express Notification for shipment 15238305825550113
<b>Attachment:</b> ./attach/100_4X_AZ-D_PA2__FedEx=5FInvoice=5FN 56=2D141.exe
<b>Subject:</b> FEDEX Shipment Status NR-6804

We also compared the body content of SPEAR and BENIGN emails. Figures 7.4(b) and 7.4(e) represent the tag clouds of the 100 most frequently occurring words in the body fields of the SPEAR and BENIGN datasets respectively. Contrary to our observations from the subject content in the SPEAR dataset (Figure 7.4(a)), the body content of the SPEAR emails (Figure 7.4(b)) did not look very attention-grabbing or themed. SPEAR bodies contained words like *attached, please, email, dear, materials, phone* etc., which commonly occur in routine email communications too. The BENIGN body content did not contain anything peculiar or alarming either (Figure 7.4(e)). Since Symantec’s email dataset of spear phishing, spam and phishing emails isn’t publicly available, we believe that this characterization of our dataset can be useful for researchers to get a better idea of state-of-the-art, real-world malicious email data which circulates in the corporate environment.

### 7.3.3 LinkedIn profile dataset

Our second dataset consisted of LinkedIn profiles of the receivers of all the emails present in our email dataset. In fact, we restricted our email dataset to only those emails which were sent to employees having at least one LinkedIn profile. This was done to have a complete dataset in terms of the availability of social and stylometric features. There were two major challenges with data collection from LinkedIn; a) Strict input requirements, and b) Rate limited API.

Firstly, to fetch the profiles of LinkedIn users who are outside a user’s network (3<sup>rd</sup> degree connections and beyond), the LinkedIn People Search API requires first name, last name, and company name as a mandatory input.<sup>9</sup> Understandably, none of the users we were looking for, were in our network, and thus, as specified in the previous subsection, we were restricted to pick up emails of only those companies which followed the format *firstName.lastName@companyDomain* or *first-Name.lastName@companyDomain*. Restricting our dataset to such email addresses was the only way we could satisfy the API’s input requirements.

<sup>9</sup>[developer.linkedin.com/documents/people-search-api](https://developer.linkedin.com/documents/people-search-api)



their Vetted API access program.<sup>10</sup> We were able to get access to the Vetted API for two of our applications. Although the new rate limit allowed 100,000 API calls per day, per application, this was still restricted to 100 calls per user, per day, per application. We then created multiple LinkedIn user accounts to make calls to the API. Even with multiple applications, and user accounts, this data collection process took about 4 months. This happened because a lot of our search queries to the API returned no results. On average, we were able to find a LinkedIn profile for only 1 in 10 users in our dataset. This resulted in about 90% of the API calls returning no results, and hence getting wasted. Eventually, we were able to collect a total of 2,434 LinkedIn profiles of victims, 5,914 LinkedIn profiles of non victims, across the 14 organizations; and 1,240 LinkedIn profiles of employees from Enron (Table 7.4). To obtain these profiles for the 9,588 employees (2,434 victims, 5,914 non victims, and 1,240 Enron employees), the number of API calls we had to make was approximately 100,000 (approx. 10 times the number of profiles obtained). Figure 7.5 shows the flow diagram describing our data collection process.

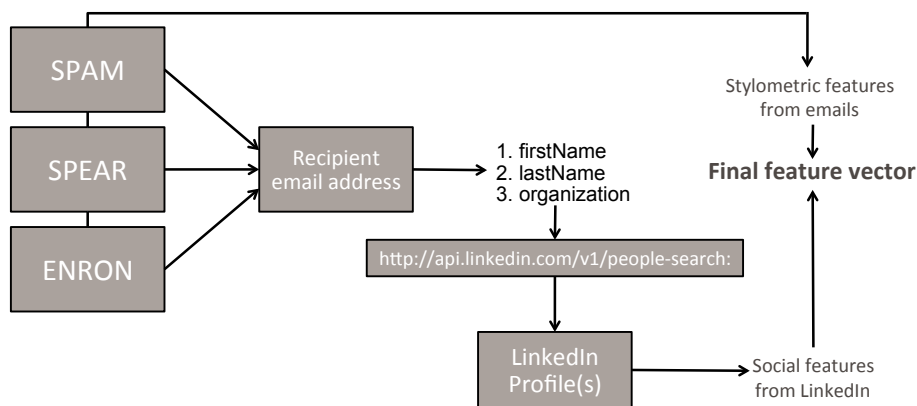


Figure 7.5: Flow diagram describing the data collection process we used to collect LinkedIn data, and create our final feature vector containing stylometric features from emails, and social features from LinkedIn profiles.

Our first choices for extracting *social* features about employees were Facebook, and Twitter. However, we found that identifying an individual on Facebook and Twitter using only the first name, last name, and employer company was a hard task. Unlike LinkedIn, the Facebook and Twitter APIs do not provide endpoints to search people using the work company name. This left us with the option to search for people using first name, and last name only. However, searching for people using only these two fields returned too many results on both Facebook and Twitter, and we had no way to identify the correct user that we were looking for. We then visited the profile pages of some users returned by the API results manually, and discovered that the *work* field for most users on Facebook was private. On Twitter profiles, there did not exist a *work* field at all. It thus became very hard to be able to find Facebook or Twitter profiles using the email addresses in our

<sup>10</sup><https://developer.linkedin.com/blog/vetted-api-access>

dataset.

## 7.4 Analysis and results

To distinguish spear phishing emails from non spear phishing emails using *social* features of the receivers, we used four machine learning algorithms, and a total of 27 features; 18 stylometric, and 9 *social*. The entire analysis and classification tasks were performed using the Weka data mining software [72]. We applied 10-fold cross validation to validate our classification results. We now describe our feature sets, analysis, and results of the classification.

### 7.4.1 Feature set description

We extracted a set of 18 stylometric features from each email in our email dataset, and a set of 9 *social* features from each LinkedIn profile present in our LinkedIn profile dataset, features described in Table 7.7. Features extracted from our email dataset are further categorized into three categories, viz. *subject* features, *attachment* features, and *body* features. It is important to note that we did not have all the three types of these aforementioned features available for all our datasets. While the SPAM dataset did not have *body* features, the BENIGN dataset did not have the *attachment* features. Features marked with \* (in Table 7.7) have been previously used by researchers to classify spam and phishing emails [152]. The *richness* feature is calculated as the ratio of the number of words to the number of characters present in the text content under consideration. We calculate richness value for the email *subject*, email *body*, and LinkedIn profile *summary*. The *Body\_hasAttach* features is a boolean variable which is set to true, if the body of the email contain the word “attached” or “attachment”, indicating that an attachment is enclosed with the email. This feature helped us to capture the presence of attachments for the BENIGN dataset, which did not have attachment information. The *Body\_numFunctionWords* feature captures the total number of occurrences of function words present in the email body, from a list of function words which includes the words: *account, access, bank, credit, click, identity, inconvenience, information, limited, log, minutes, password, recently, risk, social, security, service, and suspended*. These features have been previously used by Chandrasekaran [23].

The *social* features we extracted from the LinkedIn profiles, captured three distinct types of information about an employee, viz. location, connectivity, and profession. The *Location* was a text field containing the state / country level location of an employee, as specified by her on her LinkedIn profile. We extracted and used the country for our analysis. The *numConnections* was a numeric field, and captured the number of connections that a user has on LinkedIn. If the number of connections for a user is more than 500, the value returned is “500+” instead of the actual number of connections. These features captured the location and connectivity respectively. In addition to



TABLE 7.7: List of features used in our analysis. We used a combination of stylometric features in addition to normal features. Features marked with \* have been previously used for detecting spam and phishing emails.

Feature	Data Type	Source
Subject_IsReply*	Boolean	Email
Subject_hasBank*	Boolean	Email
Subject_numWords*	Numeric	Email
Subject_numChars*	Numeric	Email
Subject_richness*	Decimal (0-1)	Email
Subject_isForwarded*	Boolean	Email
Subject_hasVerify*	Boolean	Email
Length of attachment name	Numeric	Email
Attachment size (bytes)	Numeric	Email
Body_numUniqueWords*	Numeric	Email
Body_numNewlines	Numeric	Email
Body_numWords*	Numeric	Email
Body_numChars*	Numeric	Email
Body_richness*	Decimal (0-1)	Email
Body_hasAttach	Boolean	Email
Body_numFunctionWords*	Numeric	Email
Body_verifyYourAccount*	Boolean	Email
Body_hasSuspension*	Boolean	Email
Location	Text (country)	LinkedIn
numConnections	Numeric (0-500)	LinkedIn
SummaryLength	Numeric	LinkedIn
SummaryNumChars	Numeric	LinkedIn
SummaryUniqueWords	Numeric	LinkedIn
SummaryNumWords	Numeric	LinkedIn
SummaryRichness	Decimal (0-1)	LinkedIn
jobLevel	Numeric (0-7)	LinkedIn
jobType	Numeric (0-9)	LinkedIn

these two, we extracted 5 features from the *Summary* field, and 2 features from the *headline* field returned by LinkedIn’s People Search API. The *Summary* field is a long, free-text field comprising of a summary about a user, as specified by her, and is optional. The features we extracted from this field were similar to the ones we extracted from the subject and body fields in our email dataset. These features were, the *summary length*, *number of characters*, *number of unique words*, *total number of words*, and *richness*. We introduced two new features, *job\_level* and *job\_type*, which are numeric values ranging from 0 to 7, and 0 to 9 respectively, describing the position and area of work of an individual. We looked for presence of certain level and designation specific keywords in the “headline” field of a user, as returned by the LinkedIn API. The job levels and job types, and

their numeric equivalents are as follows:

- Job\_level; maximum of the following:
  - 1 - Support
  - 2 - Intern
  - 3 - Temporary
  - 4 - IC
  - 5 - Manager
  - 6 - Director
  - 7 - Executive
  - 0 - Other; if none of the above are found.
  
- Job\_type; minimum of the following:
  - 1 - Engineering
  - 2 - Research
  - 3 - QA
  - 4 - Information Technology
  - 5 - Operations
  - 6 - Human Resources
  - 7 - Legal
  - 8 - Finance
  - 9 - Sales / Marketing
  - 0 - Other; if none of the above are found.

To see if information extracted about a victim from online social media helps in identifying a spear phishing email sent to her, we performed classification using a) *email* features <sup>11</sup>; b) *social* features, and c) using a combination of these features. We compared these three accuracy scores across a combination of datasets viz. SPEAR versus SPAM emails from Symantec’s email scanning service, SPEAR versus benign emails from BENIGN dataset, and SPEAR versus a mixture of emails from BENIGN, and SPAM from the Symantec dataset. As mentioned earlier, not all *email* features mentioned in Table 7.7 were available for all the three email datasets. The BENIGN dataset did not have attachment related features, and the *body* field was missing in the SPAM email dataset. We thus used only those features for classification, which were available in both the targeted, and non targeted emails.

---

<sup>11</sup>We further split email features into *subject*, *body*, and *attachment* features for analysis, wherever available.

### 7.4.2 SPEAR versus SPAM emails from Symantec

Table 7.8 presents the results of our first analysis where we subjected SPEAR and SPAM emails from Symantec, to four machine learning classification algorithms, viz. Random Forest [15], J48 Decision Tree [127], Naive Bayesian [85], and Decision Table [91]. Feature vectors for this analysis were prepared from 4,742 SPEAR emails, and 9,353 SPAM emails, combined with *social* features extracted from the LinkedIn profiles of receivers of these emails. Using a combination of all *email* and *social* features, we were able to achieve a maximum accuracy of 96.47% using the Random Forest classifier for classifying SPEAR and SPAM emails. However, it was interesting to notice that two out of the four classifiers performed better *without* the social features. Although the Decision Table classifier seemed to perform equally well with, and without the social features, it performed much better using only *email* features, as compared to only *social* features.<sup>12</sup> In fact, the Decision Table classifier achieved the maximum accuracy using *attachment* features, which highlights that the attachments associated with SPEAR and SPAM emails were also substantially different in terms of name and size. We achieved an overall maximum accuracy of 98.28% using the Random Forest classifier trained on only email features. This behavior revealed that the public information available on the LinkedIn profile of an employee in our dataset, does not help in determining whether she will be targeted for a spear phishing attack or not.

TABLE 7.8: Accuracy and weighed false positive rates for SPEAR versus SPAM emails. Social features reduce the accuracy when combined with email features.

Feature set	Classifier	Random Forest	J48 Decision Tree	Naive Bayesian	Decision Table
Subject (7)	Accuracy (%)	83.91	83.10	58.87	82.04
	FP rate	0.208	0.227	0.371	0.227
Attachment (2)	Accuracy (%)	97.86	96.69	69.15	<b>95.05</b>
	FP rate	0.035	0.046	0.218	0.056
All email (9)	Accuracy (%)	<b>98.28</b>	<b>97.32</b>	68.69	<b>95.05</b>
	FP rate	0.024	0.035	0.221	0.056
Social (9)	Accuracy (%)	81.73	76.63	65.85	70.90
	FP rate	0.229	0.356	0.445	0.41
Email + Social (18)	Accuracy (%)	96.47	95.90	<b>69.35</b>	<b>95.05</b>
	FP rate	0.052	0.054	0.232	0.056

To get a better understanding of the results, we looked at the information gain associated with each feature using the InfoGainAttributeEval Attribute Evaluator package.<sup>13</sup> This package calculates the *information gain*<sup>14</sup> associated with each feature, and ranks the features in descending order of

<sup>12</sup>This happened because the Decision Table classifier terminates search after scanning for a certain (fixed) number of non-improving nodes / features.

<sup>13</sup><http://weka.sourceforge.net/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html>

<sup>14</sup>This value ranges between 0 and 1, where a higher value represents a more discriminating feature.

the information gain value. The ranking revealed that the attachment related features were the most distinguishing features between SPEAR and SPAM emails. This phenomenon was also highlighted by the Decision Table classifier (Table 7.8). The attachment size was the most distinguishing feature with an information gain score of 0.631, followed by length of attachment name, with an information gain score of 0.485. As evident from Table 7.9, attachment sizes associated with SPAM emails have very high standard deviation values, even though the average attachment sizes of SPAM and SPEAR emails are fairly similar. It is also evident that attachments associated with SPAM emails tend to have longer names; on average, twice in size as compared to attachments associated with SPEAR emails. Among subject features, we found no major difference in the length (number of characters, and number of words) of the subject fields across the two email datasets.

TABLE 7.9: Information gain, mean and standard deviation of the 10 most informative features from SPEAR and SPAM emails.

Feature	Info. Gain	SPEAR		SPAM	
		Mean	Std Dev.	Mean	Std. Dev.
Attachment size (Kb)	0.6312	285	531	262	1,419
Len. attachment name	0.4859	25.48	16.03	51.08	23.29
Subject_richness	0.2787	0.159	0.05	0.177	0.099
Subject_numChars	0.1650	29.61	17.77	31.82	23.85
Location	0.0728	-	-	-	-
Subject_numWords	0.0645	4.74	3.28	4.59	3.97
numConnections	0.0219	158.68	164.31	183.82	171.45
Subject_isForwarded	0.0219	-	-	-	-
Subject_isReply	0.0154	-	-	-	-
SummaryRichness	0.0060	0.045	0.074	0.053	0.078

It was interesting to see that apart from the Location, number of LinkedIn connections, and SummaryRichness, none of the other social features were ranked amongst the top 10 informative features. Figure 7.6 shows the top 25 *Locations* extracted from the LinkedIn profiles of employees of the 14 companies who received SPAM and SPEAR emails. We found a fairly high correlation of 0.88 between the number of SPAM and SPEAR emails received at these locations, depicting that there is not much difference between the number of SPAM and SPEAR emails received by most locations. This falls in line with the low information gain associated with this feature. Among the top 25, only 3 locations viz. France, Australia, and Afghanistan received more SPEAR emails than SPAM emails.

The number of LinkedIn connections of the recipients of SPEAR and SPAM emails in our dataset are presented in Figure 7.7. There wasn't much difference between the number of LinkedIn connections of recipients of SPEAR emails, and the number of LinkedIn connections of recipients of SPAM emails. We grouped the number of LinkedIn connections into 11 buckets as represented by the

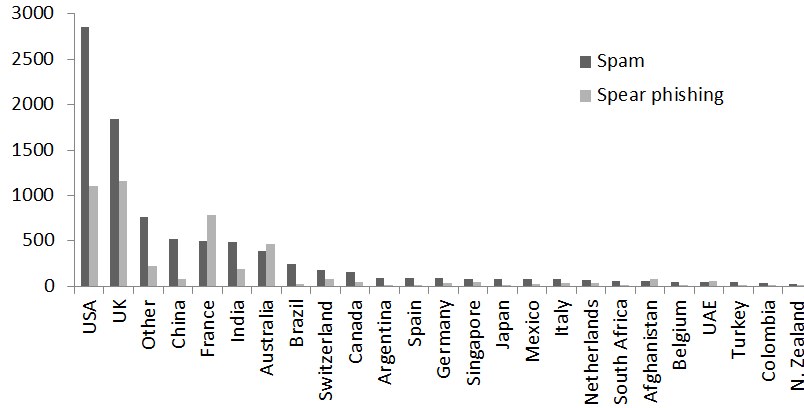


Figure 7.6: Number of SPEAR and SPAM emails received by employees in the top 25 locations extracted from their LinkedIn profiles. Employees working in France, Australia, and Afghanistan received more SPEAR emails than SPAM emails.

X axis in Figure 7.7, and found a strong correlation value of 0.97 across the two classes (SPEAR and SPAM). This confirmed that the number of LinkedIn connections did not vary much between recipients of SPEAR and SPAM emails, and thus, is not an informative feature for distinguishing between SPEAR and SPAM emails.

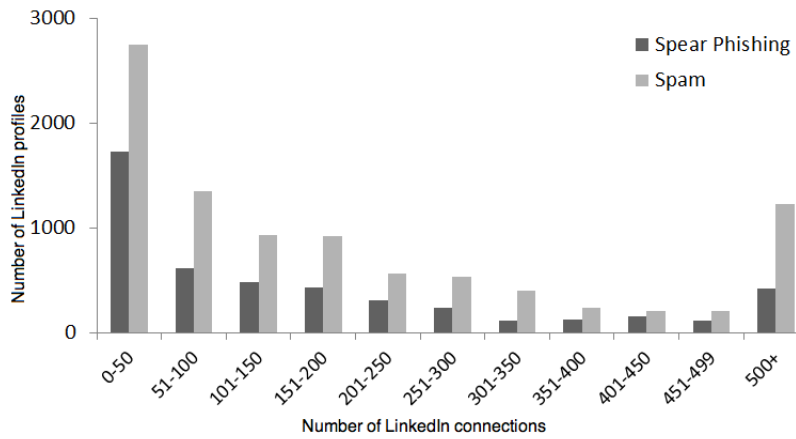


Figure 7.7: Number of LinkedIn connections of the recipients of SPEAR and SPAM emails. The number of connections are plotted on the X axis, and the number of employee profiles are plotted on the Y axis. Maximum number of employee profiles had less than 50 LinkedIn connections.

### 7.4.3 SPEAR emails versus BENIGN emails

Similar to the analysis performed in Section 7.4.2, we applied machine learning algorithms on a different dataset containing SPEAR emails, and BENIGN emails. This dataset contained 4,742

SPEAR emails, and 6,601 benign emails from BENIGN. Since BENIGN mostly contains internal email communication between Enron’s employees, we believe that it is safe to assume that none of these emails would be targeted spear phishing emails, and can be marked as benign. Similar to our observations in Section 7.4.2, we found that, in this case too, only *email* features performed slightly better than a combination of *email* and *social* features, at distinguishing spear phishing emails from non spear phishing emails. We were able to achieve a maximum accuracy of 97.04% using the Random Forest classifier trained on a set of 25 features; 16 *email*, and 9 *social* features. However, the overall maximum accuracy that we were able to achieve for this dataset was 97.39%, using only *email* features. Table 7.10 shows the results of our analysis in detail. Three out of the four classifiers performed best with *email* features; two classifiers performed best using a combination of *subject* and *body* features, while one classifier performed best using only *body* features. The Naive Bayes classifier worked best using *social* features.

TABLE 7.10: Accuracy and weighed false positive rates for SPEAR emails versus BENIGN emails. Similar to SPEAR versus SPAM, social features decrease the accuracy when combined with email features in this case too.

Feature set	Classifier	Random Forest	J48 Decision Tree	Naive Bayesian	Decision Table
Subject (7)	Accuracy (%)	81.19	81.11	61.75	79.55
	FP rate	0.210	0.217	0.489	0.228
Body (9)	Accuracy (%)	97.17	95.62	53.81	<b>90.85</b>
	FP rate	0.031	0.048	0.338	0.082
All email (16)	Accuracy (%)	<b>97.39</b>	<b>95.84</b>	54.14	89.80
	FP rate	0.029	0.044	0.334	0.090
Social (9)	Accuracy (%)	94.48	91.79	<b>69.76</b>	83.80
	FP rate	0.067	0.103	0.278	0.198
Email + Social (25)	Accuracy (%)	97.04	95.28	57.27	89.80
	FP rate	0.032	0.052	0.316	0.090

Table 7.11 presents the 10 most informative features, along with their information gain, mean and standard deviation values from the SPEAR and BENIGN datasets. The *body* features were found to be the most informative in this analysis, with only 2 *social* features among the top 10. Emails in the BENIGN dataset were found to be much longer than SPEAR emails in our Symantec dataset in terms of number of words, and number of characters in their “body”. The “subject” lengths, however, were found to be very similar across SPEAR and BENIGN.

The Random Forest classifier was also able to achieve an accuracy rate of 94.48% using only *social* features; signifying that there exist distinct differences between the LinkedIn profiles of Enron employees, and the LinkedIn profiles of the employees of the 14 companies in our dataset. The *location* attribute was found to be the most distinguishing feature among the *social* features. This

TABLE 7.11: Information gain, mean and standard deviation of the 10 most informative features from SPEAR and BENIGN emails. The *body* features performed best at distinguishing SPEAR emails from BENIGN emails.

Feature	Info. Gain	SPEAR		BENIGN	
		Mean	Std Dev.	Mean	Std. Dev.
Body_richness	0.6506	0.134	0.085	0.185	0.027
Body_numChars	0.5816	313.60	650.48	1735.5	8692.6
Body_numWords	0.4954	53.12	107.53	312.81	1572.1
Body_numUniqueWords	0.4766	38.08	49.70	149.93	416.40
Location	0.3013	-	-	-	-
Body_numNewlines	0.2660	11.29	32.70	43.58	215.77
Subject_richness	0.2230	0.159	0.051	0.174	0.056
numConnections	0.1537	158.68	164.31	259.89	167.14
Subj_numChars	0.1286	29.61	17.77	28.54	15.23
Body_numFunctionWords	0.0673	0.375	1.034	1.536	5.773

was understandable since most of the Enron employees were found to be based in the US (as Enron was an American services company). However, we also found a considerable difference in the average number of LinkedIn connections of Enron employees, and employees of the 14 organizations from our dataset (mean values for *numConnections* feature in Table 7.11).

#### 7.4.4 SPEAR versus a mixture of BENIGN and SPAM

While analyzing SPEAR with SPAM, and BENIGN emails separately, we found similar results where *social* features were not found to be very useful in both the cases. So we decided to use a mixture of SPAM and BENIGN emails against SPEAR emails, and perform the classification tasks again. We found that in this case, two out of the four classifiers performed better with a combination of email and social features, while two classifiers performed better with only *email* features. However, the overall maximum accuracy was achieved using a combination of *email* and *social* features (89.86% using Random Forest classifier). This result is in contradiction with our analysis of SPEAR versus SPAM, and SPEAR versus BENIGN separately, where *email* features always performed better independently, than a combination of *email* and *social* features. Our overall maximum accuracy, however, dropped to 89.86% (from 98.28% in SPEAR versus SPAM email classification) because of the absence of *attachment* features in this dataset. Although the *attachment* features were available in the SPAM dataset, their unavailability in BENIGN forced us to remove this feature for the current classification task. Eventually, merging the SPAM email dataset with BENIGN reduced our email dataset to only 7 features, all based on the email “subject”. Table 7.12 presents the detailed results from this analysis.

As mentioned earlier, combining the SPAM email dataset with BENIGN largely reduced our *email*

TABLE 7.12: Accuracy and weighed false positive rates for SPEAR emails versus mix of SPAM emails and BENIGN emails. Unlike SPEAR versus SPAM, or SPEAR versus BENIGN, *social* features increased the accuracy when combined with email features in this case.

Feature set	Classifier	Random Forest	J48 Decision Tree	Naive Bayesian	Decision Table
Subject (7)	Accuracy (%)	86.48	86.35	<b>77.99</b>	<b>85.46</b>
	FP rate	0.333	0.352	0.681	0.341
Social (9)	Accuracy (%)	88.04	84.69	74.46	80.61
	FP rate	0.241	0.371	0.454	0.432
Email + Social (16)	Accuracy (%)	<b>89.86</b>	<b>88.38</b>	73.97	84.14
	FP rate	0.202	0.248	0.381	0.250

feature set. We were left with 7 out of a total of 18 email features described in Table 7.7. Understandably, due to this depleted *email* feature set, we found that the email features did not perform as good as *social* features in this classification task. Despite being fewer in number, the *subject* features, viz. *Subject\_richness* and *Subject\_numChars* were found to be two of the most informative features (Table 7.13). However, the information gain value associated with both these features was fairly low. This shows that even being the best features, the *Subject\_richness* and *Subject\_numChars* were not highly distinctive features amongst spear phishing, and non spear phishing emails. Similar mean and standard deviation values for both these features in Table 7.13 confirm these outcomes.

TABLE 7.13: Information gain, mean and standard deviation of the 10 most informative features from SPEAR and a combination of BENIGN and SPAM emails. The *subject* features performed best at distinguishing SPEAR emails from non SPEAR emails.

Feature	Info. Gain	SPEAR		SPAM + BENIGN	
		Mean	Std Dev.	Mean	Std. Dev.
Subject_richness	0.1829	0.159	0.051	0.176	0.084
Subject_numChars	0.1050	29.61	17.77	30.46	20.79
Location	0.0933	-	-	-	-
numConnections	0.0388	158.68	164.31	215.30	173.76
Subject_numWords	0.0311	4.74	3.28	4.75	3.57
Subject_isForwarded	0.0188	-	-	-	-
Subject_isReply	0.0116	-	-	-	-
SummaryNumChars	0.0108	140.98	308.17	198.62	367.81
SummaryRichness	0.0090	0.045	0.074	0.057	0.080
jobLevel	0.0088	3.41	2.40	3.71	2.49

Contrary to our observations in SPEAR versus SPAM, and SPEAR versus BENIGN emails, we found five *social* features among the top 10 features in this analysis. These were the *Location*, *numConnections*, *SummaryNumChars*, *Richness*, and *jobLevel* features. Although there was a



significant difference between the average number of LinkedIn connections in the two datasets, this feature did not have much information gain associated with it due to the very large standard deviation.

## 7.5 Discussion

In this chapter, we attempted to utilize *social* features from LinkedIn profiles of employees from 14 organizations, to distinguish between spear phishing and non spear phishing emails. We extracted LinkedIn profiles of 2,434 employees who received a 4,742 targeted spear phishing emails; 5,914 employees who received 9,353 spam or phishing emails; and 1,240 Enron employees who received 6,601 benign emails. We performed our analysis on a real-world dataset from Symantec’s enterprise email scanning service, which is one of the biggest email scanning services used in the corporate organizational level. Furthermore, we targeted our analysis completely on corporate employees from 14 multinational organizations instead of random real-world users. The importance of studying spear phishing emails in particular, instead of general phishing emails, has been clearly highlighted by Jagatic et al. [81]. We performed three classification tasks viz. spear phishing emails versus spam / phishing emails, spear phishing emails versus benign emails from Enron, and spear phishing emails versus a mixture of spam / phishing emails and benign Enron emails. We found that in two out of the three cases, social features extracted from LinkedIn profiles of employees did not help in determining whether an email received by them was a spear phishing email or not. Classification results from a combination of spam / phishing, and benign emails showed some promise, where *social* features were found to be slightly helpful. The depleted *email* feature sets in this case, however, aided the enhancement in classifier performance. We believe that it is safe to conclude that publicly available content on an employee’s LinkedIn profile was not used to send her targeted spear phishing emails in our dataset. However, we cannot rule out the possibility of such an attack outside our dataset, or in future. These attacks may be better detected with access to richer *social* features. This methodology of detecting spear phishing can be helpful for safeguarding soft targets for phishers, i.e. those who have strong social media footprint. Existing phishing email filters and products can also exploit this technique to improve their performance, and provide personalized phishing filters to individuals.

There can be multiple reasons for our results being non-intuitive. Firstly, the amount of social information we were able to gather from LinkedIn, was very limited. These limitations have been discussed in Section 7.3.3. It is likely that in a real-world scenario, an attacker may be able to gain much more information about a victim prior to the attack. This could include looking for the victim’s profile on other social networks like Facebook, Twitter etc., looking for the victim’s presence on the Internet in general, using search engines (Google, Bing etc.), and profiling websites

like Pipl <sup>15</sup>, Yasni <sup>16</sup> etc. The process of data collection by automating this behavior was a time consuming process, and we were not able to take this approach due to time constraints. Secondly, it was not clear that which all aspect(s) of a user’s social profiles were most likely to be used by attackers against them. We tried to use all the features viz. textual information (summary and headline), connectivity (number of connections), work information (job level, and job type) and location information, which were made available by LinkedIn API, to perform our classification tasks. However, it is possible that none of these features were used by attackers to target their victims. In fact, we have no way to verify that the spear phishing emails in our dataset were even crafted using features from social profiles of the victims. These reasons, however, only help us in better understanding the concept of using social features in spear phishing emails.

In terms of research contributions, this work is based on a rich, true positive, real-world dataset of spear phishing, spam, and phishing emails, which is not publicly available. We believe that characterization of this data can be very useful for the entire research community to better understand the state-of-the-art spear phishing emails that have been circulated on the Internet over the past two years. To maintain anonymity and confidentiality, we could not characterize this data further, and had to anonymize the names of the 14 organizations we studied. Also, after multiple reports highlighting and warning about social media features being used in spear phishing, there does not exist much work in the research community which studies this phenomenon.

We would like to emphasize that the aim of this work is not to try and improve the existing state-of-the-art phishing email detection techniques based on their header, and content features, but to see if the introduction of social media profile features can help existing techniques to better detect spear phishing emails. We believe that this work can be a first step towards exploring threats posed by the enormous amount of contextual information about individuals, that is present on online social media. In future, we would like to carry out a similar analysis using the same email dataset, with more social features, which we were not able to collect in this attempt due to time constraints. We would also like to apply more machine learning and classification techniques like Support Vector Machines, Stochastic Gradient boosting techniques etc. on this dataset to get more insights into why social features did not perform well.

## 7.6 Conclusion

Prior research has established that introducing social context into attack vectors makes users more vulnerable to online cyber attacks in case of phishing over emails [81]. Through the analysis and experiments conducted in this chapter, we attempted to devise supervised learning based techniques to better identify such context-specific threats automatically. We attempted to utilize social

---

<sup>15</sup><https://pipl.com/>

<sup>16</sup><http://www.yasni.com/>

context in the form of LinkedIn profiles of victims to see if this social context can augment email level features to identify spear phishing emails better. Our experimental results showed that the social context did not improve detection accuracy. However, we observed that the difference in spear phishing detection accuracy with and without the social context (LinkedIn features) was not drastically different. We believe that exploring more avenues for extracting social context for a victim (like information from other social networks, profile attributes which are not publicly available, etc.) might lead to better results, especially given that past literature has established the increased vulnerability of victims in presence of social context.

What differentiates our approach from past attempts to detect spear phishing is the additional information we bring in from the social network domain. The intersection of the domains of email and social networks has been highly underexplored, especially given how closely the two domains are connected; most social networks require an email address to join the networks.

In the context of this thesis, we enhanced the scope of our work by exploring the use of context-specific social media content as an attack vector outside the social media ecosystem, on the email platform. Similar to our analysis of Facebook content in the previous chapters, we characterized and modeled poor quality context-specific content in the form of spear phishing emails in this chapter. Results from the experiments and analysis conducted in this chapter augment our work and help us better understand the landscape of poor quality context-specific content on social media services.

## Chapter 8

# Conclusion

In this chapter, we summarize the various aspects of our work towards automating quality assessment of context-specific content on online social networks. We selected Facebook as the primary platform because of its popularity and immense user base, as discussed in Chapter 1. The goal of our research was to study and analyze the domain of context-specific poor quality content on Online Social Networks, and propose and evaluate techniques for automatic real-time quality assessment of context-specific content. Section 8.1 presents the summary of our research contributions. The limitation and future directions are discussed in Section 8.2.

### 8.1 Summary

We studied poor quality content on Facebook from three separate aspects, a) individual posts, b) Facebook pages, and c) images. We collected a large dataset of news event related context-specific content from Facebook, and performed characterization studies at the level of posts, pages, and images. Using our findings from these studies, we developed and deployed Facebook Inspector, a tool to assess the quality of context-specific content on Facebook in real-time. Facebook Inspector is novel in the sense that it does not rely on metrics like *likes*, *comments*, and *shares*, which can be vital indicators of content quality, but are populated over time. Instead, Facebook Inspector designed to perform effectively on content that has just appeared on the network, and accurately assess content quality before it goes viral. Facebook Inspector’s evaluation affirms that it is fast, accurate and usable. It may be important to note that we do not focus on distinguishing between human generated versus machine generated poor quality content in this thesis. Although we found some repeating patterns, there is not enough evidence to verify whether the content we studied was generated by a human or a machine. To augment our research, we explored an alternate scenario where the attacker (entity generating poor quality content) and victim (entity consuming

poor quality content) exist on different platforms. Figure 8.1 captures a high-level schema of the problems addressed in this thesis and the techniques used to address each of these problems.

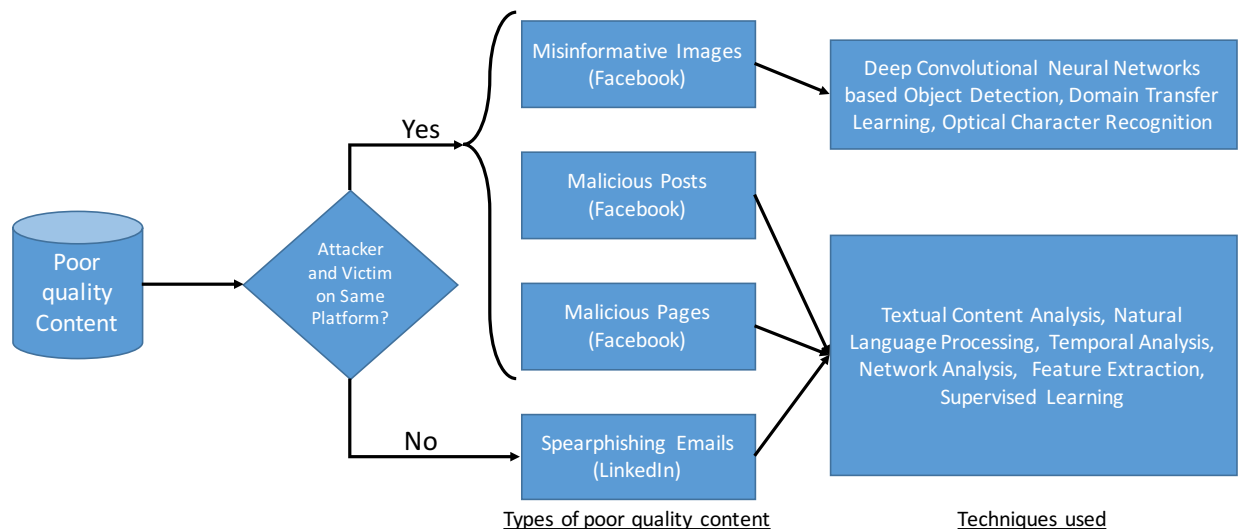


Figure 8.1: High level schema covering the attack scenarios, nature of poor quality content, and techniques used to study this content.

The main contributions of this thesis are summarized as follows:

### 8.1.1 Characterizing poor quality context-specific content

We analyzed over 4.4 million posts generated during 17 news-making events across a period of 16 months. We identified malicious posts and pages from this dataset using URL blacklists and human annotation. We then performed in-depth characterization to mine and understand any statistically significant differences between malicious and benign content. These differences were exploited to model the content quality assessment problem as a binary classification task, and we used supervised learning techniques for effective automation of the quality assessment task. The proposed solution was publicly deployed and evaluated in terms of performance, efficiency, and usability with real users.

We also explored Convolutional Neural Network based image analysis techniques to decipher crisis event using images. Using our generalizable, automated 3-tier pipeline architecture for extracting high-level image descriptors for images, we collected and characterized misinformative images posted on Facebook during a crisis event. This work provides basis for future research towards automatic identification of misinformative images on Facebook during crisis events. We extended our work to explore the possibility of more efficient identification of poor quality context-specific content in the form of spear phishing emails in an alternative scenario where the attacker and

victim exist on different platforms.

### **8.1.2 Effectiveness of automated techniques for identification of poor quality content**

Using our observations from characterization studies, we devised extensive feature sets for Facebook posts and pages separately. We picked features which we found to be differentiating poor quality content and benign content. For example, we observed lesser poor quality content being generated from mobile devices as compared to benign content. Other features that we found discriminating, included the presence of text and URLs together (Facebook posts), entities generating content (user versus page - Facebook posts), temporal activity (daily activity ratio - Facebook pages), textual content and entity names (ngram frequency - Facebook pages), no. of URL domains shared (Facebook pages), etc. In addition, we tried to capture as much publicly available information corresponding to posts and pages as possible while constructing the feature set, in order to make our model robust. Using these feature sets, we experimented with multiple supervised learning techniques ranging from probability based models to artificial neural networks, and obtained highly accurate and robust models for differentiating poor quality content from benign content. We compared our models with previously proposed models and observed that our models achieved better accuracy and recall as compared to previous clustering based models. We also performed experiments to show that models trained for identifying poor quality content in general, fail to perform well while identifying context-specific poor quality content. This result affirmed the need for exclusive models for context-specific poor quality content. Our 3-tier pipeline for associating high-level image descriptors with images proved to be helpful in considerably reducing the search space for spotting misinformative images by grouping similar images together.

### **8.1.3 Deployment and evaluation of a real-world solution for automated real-time assessment of content on Facebook**

We utilized our supervised learning models to develop and deploy Facebook Inspector, an automated real-time framework to assess the quality of context-specific content on Facebook. This framework was made available publicly in the form of a RESTful API and a browser plug-in. Facebook Inspector works on a two-fold filtering approach and combines two supervised learning models to identify poor quality content. The Facebook Inspector browser plug-in has been active for over 18 months, and has processed over 6 million requests. It has had a consistent user base of 200+ daily active users for over 6 months. Facebook Inspector was evaluated in terms of performance, response time, and usability. Users of Facebook Inspector found it to be usable and helpful in various aspects.

## 8.2 Limitations and Future work

The dataset we used for our analysis comprises of only public posts. We were not able to find a way to validate if our dataset is a representative sample of the entire Facebook stream. However, to the best of our knowledge, our dataset of 4.4 million posts generated by 3.3 million users is one of the biggest datasets of Facebook posts ever analyzed as part of academic research.

Facebook Inspector is trained to work on context-specific content. However, in the current setting, Facebook Inspector processes all public posts it comes across, and can produce erratic results for non context-specific content. We would like to develop and integrate a mechanism to filter only context-specific content to process with the framework. We would also like to add support for accessing and processing private Facebook posts by adding a user authentication and authorization module with Facebook Inspector.

Results drawn from our research on studying crisis events on social networks through pictures were based on data accumulated with respect to only one crisis event. We do not claim that similar results can be expected from all similar crisis events. However, the methodology we devised for studying an event on social media (crisis or non-crisis) is generalizable for any dataset of images belonging to a specific domain.

Facebook pages have a lot in common with Facebook groups and events. Groups and events can also be used to target large audiences at once. Our analysis and results can thus be easily extended to study malicious groups and events on Facebook as well.

An intuitive way to attack the problem of automatically identifying poor quality content is to focus on the intent behind the creation of such content, and target the intent itself. However, the exact intent and purposes behind creation of poor quality content are hard to identify without access to an entity that engages in such actions. Some other researchers have cited reasons like monetary gains and degrading user experience to drive away users from the network, behind generation of various types of poor quality content. Given the nature of some of our findings (including presence of politically polarized content), we believe that propaganda might also be one of the driving forces behind creation of poor quality content.

It would have been interesting to study the generalizability of our proposed features for differentiating poor quality content from benign content on other platforms such as Twitter. However, it may be important to note that a lot of features we used in our work are either specific to the Facebook platform, or not publicly available on some other platforms. These factors along with the deliberate choice of network for this thesis push the generalizability of features out of scope of this work.

We believe that insights obtained from this thesis can be utilized by researchers and stakeholders to make social media environment safer and more informative. Based on our experience so far, we

suggest the following directions:

**Address different types of poor quality content individually.** Our analysis revealed the presence of multiple independent categories of content which were deemed as poor quality on the Internet. These categories included child unsafe content, politically polarized content, inaccurate information (rumors, hoaxes, misinformation), etc. In this thesis, we attempted to collectively model content from all these categories as poor quality at once, and devised machine learning techniques to automatically identify them. Although we achieved a certain level of success, it is possible that modeling content in each of these categories separately might produce more accurate models. This avenue can be explored further on Facebook or on any other social media platform.

**Utilize crowdsourcing to personalize and improve the performance of automated techniques for poor quality content identification.** Our evaluation of Facebook Inspector revealed that one of its major limitations was the inability to adapt to the preferences of individual users. Given a Facebook post marked as poor quality by Facebook Inspector, while most users agreed to the post being poor quality, other users felt differently. We realized that given the complex definition and scope of poor quality content we were dealing with, users' definition of "poor quality" differs. For example, a Facebook post appreciating a political agenda might be informative to some, while propaganda to others. For such cases, there needs to be a way to accommodate users' feedback and preferences to enable customization of content filtering techniques based on these preferences. This is likely to make automated tools more widely accepted and thus safeguard a wider audience.

**Explore the impact of images posted during news events.** Results from our analysis of images posted on Facebook during the Paris attacks highlighted multiple interesting visual themes which were popular among image content. However, the impact of such images on end users' perception of the event is largely underexplored. It would be interesting to study whether the end users' perception of an event or topic on social media is driven by text or images. We believe that images would have a significant role to play in the users' understanding of the topic or event. We also believe that the results of such a study would be helpful for brands, government, and all other relevant stakeholders to identify and decide whether existing text based topic and sentiment identification techniques suffice to accurately gauge public perception and sentiment.



# Bibliography

- [1] ABU-NIMEH, S., NAPPA, D., WANG, X., AND NAIR, S. A comparison of machine learning techniques for phishing detection. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit* (2007), ACM, pp. 60–69.
- [2] ACAR, A., AND MURAKI, Y. Twitter for crisis communication: lessons learned from japan’s tsunami disaster. *International Journal of Web Based Communities* 7, 3 (2011), 392–402.
- [3] ACQUISTI, A., AND GROSS, R. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Privacy enhancing technologies* (2006), Springer, pp. 36–58.
- [4] AGGARWAL, A., RAJADESINGAN, A., AND KUMARAGURU, P. Phishari: Automatic realtime phishing detection on twitter. In *eCrime Researchers Summit (eCrime), 2012* (2012), IEEE, pp. 1–12.
- [5] AHMED, F., AND ABULAISH, M. An mcl-based approach for spam profile detection in online social networks. In *IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)* (2012), IEEE, pp. 602–608.
- [6] ANDREASSEN, C. S., TORSHEIM, T., BRUNBORG, G. S., AND PALLESEN, S. Development of a facebook addiction scale 1, 2. *Psychological reports* 110, 2 (2012), 501–517.
- [7] ANTONIADES, D., POLAKIS, I., KONTAXIS, G., ATHANASOPOULOS, E., IOANNIDIS, S., MARKATOS, E. P., AND KARAGIANNIS, T. we. b: The web of short urls. In *Proceedings of the 20th international conference on World Wide Web* (2011), ACM, pp. 715–724.
- [8] BACKSTROM, L., BOLDI, P., ROSA, M., UGANDER, J., AND VIGNA, S. Four degrees of separation. In *Proceedings of the 3rd Annual ACM Web Science Conference* (2012), ACM, pp. 33–42.
- [9] BAY, H., TUYTELAARS, T., AND VAN GOOL, L. Surf: Speeded up robust features. In *European conference on computer vision* (2006), Springer, pp. 404–417.

- [10] BECKER, H., NAAMAN, M., AND GRAVANO, L. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the fifth International AAAI Conference on Weblogs and Social Media* (2011).
- [11] BENEVENUTO, F., MAGNO, G., RODRIGUES, T., AND ALMEIDA, V. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)* (2010), vol. 6, p. 12.
- [12] BOLLEN, J., MAO, H., AND ZENG, X. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8.
- [13] BORTH, D., CHEN, T., JI, R., AND CHANG, S.-F. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia* (2013), ACM, pp. 459–460.
- [14] BOYD, D. Facebook’s privacy trainwreck. *Convergence: The International Journal of Research into New Media Technologies* 14, 1 (2008), 13–20.
- [15] BREIMAN, L. Random forests. *Machine Learning* 45, 1 (2001), 5–32.
- [16] BREIMAN, L., FRIEDMAN, J., STONE, C. J., AND OLSHEN, R. A. *Classification and regression trees*. CRC press, 1984.
- [17] BROOKE, J. SUS-a quick and dirty usability scale. *Usability evaluation in industry* 189 (1996), 194.
- [18] CARLISLE, J. E., AND PATTON, R. C. Is social media changing how we understand political engagement? an analysis of facebook and the 2008 presidential election. *Political Research Quarterly* 66, 4 (2013), 883–895.
- [19] CASTILLO, C., MENDOZA, M., AND POBLETE, B. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (2011), ACM, pp. 675–684.
- [20] CATANESE, S., DE MEO, P., FERRARA, E., FIUMARA, G., AND PROVETTI, A. Extraction and analysis of facebook friendship relations. In *Computational Social Networks*. Springer, 2012, pp. 291–324.
- [21] CATANESE, S. A., DE MEO, P., FERRARA, E., FIUMARA, G., AND PROVETTI, A. Crawling facebook for social network analysis purposes. In *Proceedings of the international conference on web intelligence, mining and semantics* (2011), ACM, p. 52.

- [22] CHA, M., HADDADI, H., BENEVENUTO, F., AND GUMMADI, P. K. Measuring user influence in twitter: The million follower fallacy. *Proceedings of the fourth International AAAI Conference on Weblogs and Social Media 10* (2010), 10–17.
- [23] CHANDRASEKARAN, M., NARAYANAN, K., AND UPADHYAYA, S. Phishing email detection based on structural properties. In *NYS Cyber Security Conference* (2006), pp. 1–7.
- [24] CHEN, L., AND ROY, A. Event detection from flickr data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), ACM, pp. 523–532.
- [25] CHEONG, F., AND CHEONG, C. Social media data mining: A social network analysis of tweets during the 2010–2011 australian floods. *15th Pacific Asia Conference on Information Systems (PACIS)* (2011).
- [26] CHEONG, M., AND LEE, V. C. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via twitter. *Information Systems Frontiers 13*, 1 (2011), 45–59.
- [27] CHHABRA, S., AGGARWAL, A., BENEVENUTO, F., AND KUMARAGURU, P. Phi.sh/\$ocial: the phishing landscape through short urls. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference* (2011), ACM, pp. 92–101.
- [28] CHU, Z., WIDJAJA, I., AND WANG, H. Detecting social spam campaigns on twitter. In *Applied Cryptography and Network Security* (2012), Springer, pp. 455–472.
- [29] COHEN, W. W. Enron email dataset. *Internet: www.cs.cmu.edu/enron/ (Last Accessed: May 25, 2008)* (2009).
- [30] DALY, M. K. Advanced persistent threat. In *23rd Large Installation System Administration Conference. USENIX, Baltimore* (2009).
- [31] DE CHOUDHURY, M., MONROY-HERNANDEZ, A., AND MARK, G. Narco emotions: affect and desensitization in social media during the mexican drug war. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (2014), ACM, pp. 3563–3572.
- [32] DE LONGUEVILLE, B., SMITH, R. S., AND LURASCHI, G. Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *LBSN* (2009), ACM, pp. 73–80.
- [33] DEWAN, P., BAGROY, S., AND KUMARAGURU, P. Hiding in plain sight: Characterizing and detecting malicious facebook pages. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2016), IEEE, pp. 193–196.

- [34] DEWAN, P., BAGROY, S., AND KUMARAGURU, P. *Hiding in Plain Sight: The Anatomy of Malicious Pages on Facebook*. Springer, 2017.
- [35] DEWAN, P., KASHYAP, A., AND KUMARAGURU, P. Analyzing social and stylometric features to identify spear phishing emails. In *APWG Symposium on Electronic Crime Research (eCrime)* (2014), IEEE, pp. 1–13.
- [36] DEWAN, P., AND KUMARAGURU, P. Towards automatic real time identification of malicious posts on facebook. In *13th Annual Conference on Privacy, Security and Trust (PST)* (2015), IEEE, pp. 85–92.
- [37] DEWAN, P., AND KUMARAGURU, P. Facebook inspector (fbi): Towards automatic real-time detection of malicious content on facebook. *Social Network Analysis and Mining* 7, 1 (Apr 2017), 15.
- [38] DEWAN, P., SURI, A., BHARADHWAJ, V., MITHAL, A., AND KUMARAGURU, P. Towards understanding crisis events on online social networks through pictures. *IEEE/ACM Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2017).
- [39] DHAMIJA, R., TYGAR, J. D., AND HEARST, M. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (2006), ACM, pp. 581–590.
- [40] DONAHUE, J., JIA, Y., VINYALS, O., HOFFMAN, J., ZHANG, N., TZENG, E., AND DARRELL, T. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning* (2014), pp. 647–655.
- [41] DOUCEUR, J. R. The sybil attack. In *Peer-to-peer Systems*. Springer, 2002, pp. 251–260.
- [42] DOUGLAS, S., MARUYAMA, M., SEMAAN, B., AND ROBERTSON, S. P. Politics and young adults: The effects of facebook on candidate evaluation. In *Proceedings of the 15th Annual International Conference on Digital Government Research* (New York, NY, USA, 2014), dg.o ’14, ACM, pp. 196–204.
- [43] EARLE, P., GUY, M., BUCKMASTER, R., OSTRUM, C., HORVATH, S., AND VAUGHAN, A. Omg earthquake! can twitter improve earthquake response? *Seismological Research Letters* 81, 2 (2010), 246–251.
- [44] EL-ARINI, K., AND TANG, J. News feed fyi: Click-baiting. *Facebook Newsroom* (2014).
- [45] ELLISON, N. B., STEINFELD, C., AND LAMPE, C. The benefits of facebook “friends:” social capital and college students’ use of online social network sites. *Journal of Computer-Mediated Communication* 12, 4 (2007), 1143–1168.

- [46] FACEBOOK. Facebook company info. <http://newsroom.fb.com/company-info/> (Last Accessed: July 12, 2017) (2014).
- [47] FACEBOOK. What is page spam? <https://www.facebook.com/help/116053525145846> (Last Accessed: September 18, 2015) (2015).
- [48] FACEBOOK, ERICSSON, AND QUALCOMM. A focus on efficiency. *Whitepaper, Internet.org* (2013).
- [49] FACEBOOK DEVELOPERS. Keeping you safe from scams and spam. <https://www.facebook.com/notes/facebook-security/keeping-you-safe-from-scams-and-spam/10150174826745766> (Last Accessed: July 12, 2017) (2011).
- [50] FACEBOOK DEVELOPERS. Facebook graph api search. <https://developers.facebook.com/docs/graph-api/using-graph-api/v1.0#search> (Last Accessed: July 12, 2017) (2013).
- [51] FACEBOOK.COM. Facebook community standards. <https://www.facebook.com/communitystandards> (Last Accessed: July 12, 2017) (2015).
- [52] FALOUTSOS, M. Detecting malware with graph-based methods: traffic classification, botnets, and facebook scams. In *Proceedings of the 22nd international conference on World Wide Web companion* (2013), International World Wide Web Conferences Steering Committee, pp. 495–496.
- [53] FAN, W., AND YEUNG, K.-H. Virus propagation modeling in facebook. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2010), IEEE, pp. 331–335.
- [54] FETTE, I., SADEH, N., AND TOMASIC, A. Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web* (2007), ACM, pp. 649–656.
- [55] FRAUSTINO, J. D., LIU, B., AND JIN, Y. Social media use during disasters: a review of the knowledge base and gaps. *National Consortium for the Study of Terrorism and Responses to Terrorism [START]*. (2012).
- [56] FRIGGERI, A., ADAMIC, L. A., ECKLES, D., AND CHENG, J. Rumor cascades. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* (2014).
- [57] GAO, H., CHEN, Y., LEE, K., PALSETIA, D., AND CHOUDHARY, A. N. Towards online spam filtering in social networks. In *NDSS* (2012).

- [58] GAO, H., HU, J., WILSON, C., LI, Z., CHEN, Y., AND ZHAO, B. Y. Detecting and characterizing social spam campaigns. In *Internet Measurement Conference (2010)*, ACM, pp. 35–47.
- [59] GJOKA, M., KURANT, M., BUTTS, C. T., AND MARKOPOULOU, A. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proceedings IEEE (2010)*, IEEE, pp. 1–9.
- [60] GONZALES, A. L., AND HANCOCK, J. T. Mirror, mirror on my facebook wall: Effects of exposure to facebook on self-esteem. *Cyberpsychology, Behavior, and Social Networking* 14, 1-2 (2011), 79–83.
- [61] GOOGLE. Safe browsing api. <https://developers.google.com/safe-browsing/> (Last Accessed: July 12, 2017) (2014).
- [62] GRIER, C., THOMAS, K., PAXSON, V., AND ZHANG, M. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security (2010)*, ACM, pp. 27–37.
- [63] GUARDIAN, T. Facebook’s failure: Did fake news and polarized politics get trump elected? <https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories> (Last Accessed: July 12, 2017) (2016).
- [64] GUPTA, A., JOSHI, A., AND KUMARAGURU, P. Identifying and characterizing user communities on twitter during crisis events. In *Proceedings of the 2012 Workshop on Data-driven User Behavioral Modelling and Mining from Social Media (New York, NY, USA, 2012)*, DUB-MMSM ’12, ACM, pp. 23–26.
- [65] GUPTA, A., AND KUMARAGURU, P. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st workshop on privacy and security in online social media (2012)*, ACM, p. 2.
- [66] GUPTA, A., KUMARAGURU, P., CASTILLO, C., AND MEIER, P. Tweetcred: Real-time credibility assessment of content on twitter. In *Social Informatics*. Springer, 2014, pp. 228–243.
- [67] GUPTA, A., LAMBA, H., AND KUMARAGURU, P. \$1.00 per rt# bostonmarathon# pray-forboston: Analyzing fake content on twitter. In *eCrime Researchers Summit (eCRS), 2013 (2013)*, IEEE, pp. 1–12.
- [68] GUPTA, A., LAMBA, H., KUMARAGURU, P., AND JOSHI, A. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web (2013)*, ACM, pp. 729–736.

- [69] GUPTA, M., ZHAO, P., AND HAN, J. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining* (2012), SIAM, pp. 153–164.
- [70] GUPTA, S., GUPTA, P., AHAMAD, M., AND KUMARAGURU, P. Exploiting phone numbers and cross-application features in targeted mobile attacks. In *Proceedings of the 6th Workshop on Security and Privacy in Smartphones and Mobile Devices* (2016), ACM, pp. 73–82.
- [71] HALEVI, T., LEWIS, J., AND MEMON, N. Phishing, personality traits and facebook. *arXiv preprint arXiv:1301.7643* (2013).
- [72] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.
- [73] HAMID, I. R. A., ABAWAJY, J., AND KIM, T.-H. Using feature selection and classification scheme for automating phishing email detection. *Studies In Informatics and Control, ISSN* (2013), 1220–1766.
- [74] HARGITTAI, E., ET AL. Facebook privacy settings: Who cares? *First Monday* 15, 8 (2010).
- [75] HILLE, S., AND BAKKER, P. I like news: Searching for the holy grail of social media: The use of facebook by dutch news media and their audiences. *European Journal of Communication* (2013), 0267323113497435.
- [76] HISPASEC SISTEMAS S.L. VirusTotal Public API. <https://www.virustotal.com/en/documentation/public-api/> (Last Accessed: July 12, 2017) (2013).
- [77] HOLCOMB, J., GOTTFRIED, J., AND MITCHELL, A. News use across social media platforms. *Technical report, Pew Research Center*. (2013).
- [78] HU, M., LIU, S., WEI, F., WU, Y., STASKO, J., AND MA, K.-L. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 2751–2754.
- [79] HUGHES, A. L., AND PALEN, L. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management* 6, 3 (2009), 248–260.
- [80] ISLAM, R., AND ABAWAJY, J. A multi-tier phishing detection and filtering approach. *Journal of Network and Computer Applications* 36, 1 (2013), 324–335.
- [81] JAGATIC, T. N., JOHNSON, N. A., JAKOBSSON, M., AND MENCZER, F. Social phishing. *Communications of the ACM* 50, 10 (2007), 94–100.

- [82] JAKOBSSON, M. Modeling and preventing phishing attacks. In *Financial Cryptography* (2005), vol. 5, Citeseer.
- [83] JAKOBSSON, M., AND MYERS, S. *Phishing and countermeasures: understanding the increasing problem of electronic identity theft*. John Wiley & Sons, 2006.
- [84] JIN, X., LIN, C., LUO, J., AND HAN, J. A data mining-based spam detection system for social media networks. *Proceedings of the VLDB Endowment* 4, 12 (2011).
- [85] JOHN, G. H., AND LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence* (San Mateo, 1995), Morgan Kaufmann, pp. 338–345.
- [86] JOINSON, A. N. Looking at, looking up or keeping up with people?: motives and use of facebook. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (2008), ACM, pp. 1027–1036.
- [87] KARMA, F. Study: Average growth of facebook fan pages. <http://blog.fanpagekarma.com/2013/03/20/infographic-average-growths-facebook-fan-pages/> (Last Accessed: July 12, 2017) (2013).
- [88] KHARROUB, T., AND BAS, O. Social media and protests: An examination of twitter images of the 2011 egyptian revolution. *New Media & Society* (2015).
- [89] KIM, J., AND LEE, J.-E. R. The facebook paths to happiness: Effects of the number of facebook friends and self-presentation on subjective well-being. *Cyberpsychology, Behavior, and Social Networking* 14, 6 (2011), 359–364.
- [90] KIREYEV, K., PALEN, L., AND ANDERSON, K. Applications of topics models to analysis of disaster-related twitter data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond* (2009), vol. 1.
- [91] KOHAVI, R. The power of decision tables. In *8th European Conference on Machine Learning* (1995), Springer, pp. 174–189.
- [92] KUMARAGURU, P., CRANSHAW, J., ACQUISTI, A., CRANOR, L., HONG, J., BLAIR, M. A., AND PHAM, T. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (2009), ACM, p. 3.
- [93] KUMARAGURU, P., RHEE, Y., ACQUISTI, A., CRANOR, L. F., HONG, J., AND NUNGE, E. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2007), ACM, pp. 905–914.



- [94] KUMARAGURU, P., SHENG, S., ACQUISTI, A., CRANOR, L. F., AND HONG, J. Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)* 10, 2 (2010), 7.
- [95] KWAK, H., LEE, C., PARK, H., AND MOON, S. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web* (2010), ACM, pp. 591–600.
- [96] LEE, M. Who’s next? identifying risks factors for subjects of targeted attacks. In *Proc. Virus Bull. Conf* (2012), pp. 301–306.
- [97] LEE, M., AND LEWIS, D. Clustering disparate attacks: mapping the activities of the advanced persistent threat. *Last accessed June 26* (2013).
- [98] LEWIS, K., KAUFMAN, J., GONZALEZ, M., WIMMER, A., AND CHRISTAKIS, N. Tastes, ties, and time: A new social network dataset using facebook. com. *Social networks* 30, 4 (2008), 330–342.
- [99] LINDLEY, G. Public conversations on facebook. <http://newsroom.fb.com/news/2013/06/public-conversations-on-facebook/> (Last Accessed: July 12, 2017) (2013).
- [100] LIU, Y., GUMMADI, K. P., KRISHNAMURTHY, B., AND MISLOVE, A. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference* (2011), ACM, pp. 61–70.
- [101] LOWE, D. G. Object recognition from local scale-invariant features. In *The proceedings of the seventh IEEE international conference on Computer Vision* (1999), vol. 2, Ieee, pp. 1150–1157.
- [102] MA, L., OFOGHI, B., WATTERS, P., AND BROWN, S. Detecting phishing emails using hybrid features. In *Ubiquitous, Autonomic and Trusted Computing, 2009. UIC-ATC’09. Symposia and Workshops on* (2009), IEEE, pp. 493–497.
- [103] MAGAZINE, C. R. Facebook & your privacy. who sees the data you share on the biggest social network? <http://www.consumerreports.org/cro/magazine/2012/06/facebook-your-privacy/index.htm> (Last Accessed: July 12, 2017) (2012).
- [104] MANN, H. B., AND WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.
- [105] MARCA.COM. Luis suarez used as bait for facebook scam. <http://www.marca.com/2014/07/18/en/football/barcelona/1405709402.html> (Last Accessed: July 12, 2017) (2014).

- [106] MCAULEY, J. J., AND LESKOVEC, J. Learning to discover social circles in ego networks. In *Advances in neural information processing systems* (2012), vol. 2012, pp. 548–56.
- [107] MENDOZA, M., POBLETE, B., AND CASTILLO, C. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics* (2010), ACM, pp. 71–79.
- [108] MICHAEL BARTHEL, ELISA SHEARER, J. G., AND MITCHELL, A. News use on facebook and twitter is on the rise. *Pew Research Centre* (2015).
- [109] MORAN, N., AND LANSTEIN, A. Spear phishing the news cycle: Apt actors leverage interest in the disappearance of malaysian flight mh 370. <https://www.fireeye.com/blog/threat-research/2014/03/spear-phishing-the-news-cycle-apt-actors-leverage-interest-in-the-disappearance-of-malaysian-flight-mh-370.html> (Last Accessed: July 12, 2017) (2014).
- [110] NEWSROOM, F. News feed fyi: Reducing overly promotional page posts in news feed. <https://newsroom.fb.com/news/2014/11/news-feed-fyi-reducing-overly-promotional-page-posts-in-news-feed/> (Last Accessed: July 12, 2017) (2014).
- [111] OPENDNS. Phishtank api. [http://www.phishtank.com/api\\_info.php](http://www.phishtank.com/api_info.php) (Last Accessed: July 12, 2017) (2014).
- [112] OPSAHL, T. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks* 35, 2 (2013), 159–167.
- [113] OSBORNE, M., AND DREDZE, M. Facebook, twitter and google plus for breaking news: Is there a winner? *Proceedings of the eighth International AAAI Conference on Weblogs and Social Media* (2014).
- [114] OSBORNE, M., PETROVIC, S., MCCREADIE, R., MACDONALD, C., AND OUNIS, I. Bieber no more: First story detection using twitter and wikipedia. In *Proceedings of the Workshop on Time-aware Information Access. TAIA* (2012), vol. 12.
- [115] OSOFSKY, J. Information about trending topics. *Facebook Newsroom* (2016).
- [116] OWENS, E., AND TURITZIN, C. News feed fyi: Cleaning up news feed spam. <http://newsroom.fb.com/news/2014/04/news-feed-fyi-cleaning-up-news-feed-spam/> (Last Accessed: July 12, 2017) (2014).
- [117] OWENS, E., AND WEINSBERG, U. News feed fyi: Showing fewer hoaxes. <https://newsroom.fb.com/news/2015/01/news-feed-fyi-showing-fewer-hoaxes/> (Last Accessed: July 12, 2017) (2015).

- [118] PALEN, L. Online social media in crisis events. *Educause Quarterly* 31, 3 (2008), 76–78.
- [119] PALEN, L., AND VIEWEG, S. The emergence of online widescale interaction in unexpected events: assistance, alliance & retreat. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (2008), ACM, pp. 117–126.
- [120] PAN, S. J., AND YANG, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2010), 1345–1359.
- [121] PARKINSON, M. The power of visual communication. *Billion Dollar Graphics* (2012).
- [122] PATSAKIS, C., ASTHENIDIS, A., AND CHATZIDIMITRIOU, A. Social networks as an attack platform: Facebook case study. In *Eighth International Conference on Networks* (2009), IEEE, pp. 245–247.
- [123] PEMPEK, T. A., YERMOLAYEVA, Y. A., AND CALVERT, S. L. College students’ social networking experiences on facebook. *Journal of Applied Developmental Psychology* 30, 3 (2009), 227–238.
- [124] PENNEBAKER, J. W., CHUNG, C. K., IRELAND, M., GONZALES, A., AND BOOTH, R. J. The development and psychometric properties of liwc2007, 2007.
- [125] PEYSAKHOVICH, A., AND HENDRIX, K. News feed fyi: Further reducing clickbait in feed. *Facebook Newsroom* (2016).
- [126] POULSEN, K. Firsthand reports from california wildfires pour through twitter. *Retrieved February 15* (2007), 2009.
- [127] QUINLAN, R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.
- [128] RAHMAN, M. S., HUANG, T.-K., MADHYASTHA, H. V., AND FALOUTSOS, M. Efficient and scalable socware detection in online social networks. In *USENIX Security Symposium* (2012), pp. 663–678.
- [129] RAHMAN, M. S., HUANG, T.-K., MADHYASTHA, H. V., AND FALOUTSOS, M. Frappe: detecting malicious facebook applications. In *Proceedings of the 8th international conference on Emerging networking experiments and technologies* (2012), ACM, pp. 313–324.
- [130] REDONDO, R. Q. The image use on twitter during the 2015 municipal election campaign in spain. *Revista Latina de Comunicación Social* 71 (2016), 85–107.
- [131] ROOM, F. N. Graph search now includes posts and status updates. <http://newsroom.fb.com/News/728/Graph-Search-Now-Includes-Posts-and-Status-Updates> (Last Accessed: July 12, 2017) (2013).

- [132] ROSS, C., ORR, E. S., SISIC, M., ARSENEAULT, J. M., SIMMERING, M. G., AND ORR, R. R. Personality and motivations associated with facebook use. *Computers in Human Behavior* 25, 2 (2009), 578–586.
- [133] RUDRA, K., BANERJEE, S., GANGULY, N., GOYAL, P., IMRAN, M., AND MITRA, P. Summarizing situational tweets in crisis scenario. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media* (2016), ACM, pp. 137–147.
- [134] RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATY, A., KHOSLA, A., BERNSTEIN, M., BERG, A. C., AND FEI-FEI, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252.
- [135] SAKAKI, T., OKAZAKI, M., AND MATSUO, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web* (2010), ACM, pp. 851–860.
- [136] SAKAKI, T., TORIUMI, F., AND MATSUO, Y. Tweet trend analysis in an emergency situation. In *Proceedings of the Special Workshop on Internet and Disasters* (2011), ACM, p. 3.
- [137] SECURITY, R. B. S. Facebook names dataset. <https://blog.skullsecurity.org/2010/return-of-the-facebook-snatchers> (Last Accessed: July 12, 2017) (2010).
- [138] SEMAAN, B., AND MARK, G. 'facebooking'towards crisis recovery and beyond: disruption as an opportunity. In *Proceedings of the ACM 2012 conference on computer supported cooperative work* (2012), ACM, pp. 27–36.
- [139] SEO, H. Visual propaganda in the age of social media: An empirical analysis of twitter images during the 2012 israeli–hamas conflict. *Visual Communication Quarterly* 21, 3 (2014), 150–161.
- [140] SHENG, S., HOLBROOK, M., KUMARAGURU, P., CRANOR, L. F., AND DOWNS, J. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), ACM, pp. 373–382.
- [141] SHENG, S., WARDMAN, B., WARNER, G., CRANOR, L., HONG, J., AND ZHANG, C. An empirical analysis of phishing blacklists. In *Sixth Conference on Email and Anti-Spam (CEAS)* (2009).
- [142] SPAMHAUS. Domain block list. <http://www.spamhaus.org/dbl/> (Last Accessed: July 12, 2017) (2014).

- [143] STEIN, T., CHEN, E., AND MANGLA, K. Facebook immune system. In *Workshop on Social Network Systems* (2011), ACM, p. 8.
- [144] STIEGLITZ, S., AND DANG-XUAN, L. Political communication and influence through microblogging—an empirical analysis of sentiment in twitter messages and retweet behavior. In *System Science (HICSS), 2012 45th Hawaii International Conference on* (2012), IEEE, pp. 3500–3509.
- [145] STRINGHINI, G., KRUEGEL, C., AND VIGNA, G. Detecting spammers on social networks. In *Proceedings of the 26th annual computer security applications conference* (2010), ACM, pp. 1–9.
- [146] SURBL, URI. Reputation data. <http://www.surbl.org/surbl-analysis> (Last Accessed: July 12, 2017) (2011).
- [147] SZEGEDY, C., VANHOUCKE, V., IOFFE, S., SHLENS, J., AND WOJNA, Z. Rethinking the inception architecture for computer vision. *CoRR abs/1512.00567* (2015).
- [148] SZELL, M., GRAUWIN, S., AND RATTI, C. Contraction of online response to major events. *PLoS ONE 9(2): e89052, MIT* (2014).
- [149] THEGUARDIAN. Facebook spammers make \$200m just posting links, researchers say. <http://www.theguardian.com/technology/2013/aug/28/facebook-spam-202-million-italian-research> (Last Accessed: July 12, 2017) (2013).
- [150] THELWALL, M., BUCKLEY, K., AND PALTOGLOU, G. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology 62, 2* (2011), 406–418.
- [151] THONNARD, O., BILGE, L., O’GORMAN, G., KIERNAN, S., AND LEE, M. Industrial espionage and targeted attacks: Understanding the characteristics of an escalating threat. In *Research in Attacks, Intrusions, and Defenses*. Springer, 2012, pp. 64–85.
- [152] TOOLAN, F., AND CARTHY, J. Feature selection for spam and phishing detection. In *eCrime Researchers Summit (eCrime), 2010* (2010), IEEE, pp. 1–12.
- [153] TRAUD, A. L., KELSIC, E. D., MUCHA, P. J., AND PORTER, M. A. Comparing community structure to characteristics in online collegiate social networks. *SIAM review 53, 3* (2011), 526–543.
- [154] TRAUD, A. L., MUCHA, P. J., AND PORTER, M. A. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications 391, 16* (2012), 4165–4180.

- [155] TUMASJAN, A., SPRENGER, T. O., SANDNER, P. G., AND WELPE, I. M. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review* (2010), 0894439310386557.
- [156] TUMASJAN, A., SPRENGER, T. O., SANDNER, P. G., AND WELPE, I. M. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the fourth International AAAI Conference on Weblogs and Social Media* (2010), pp. 178–185.
- [157] UGANDER, J., KARRER, B., BACKSTROM, L., AND MARLOW, C. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503* (2011).
- [158] VAROL, O., FERRARA, E., OGAN, C. L., MENCZER, F., AND FLAMMINI, A. Evolution of online user behavior during a social upheaval. In *Proceedings of the 2014 ACM conference on Web science* (2014), ACM, pp. 81–90.
- [159] VIEWEG, S., HUGHES, A. L., STARBIRD, K., AND PALEN, L. 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* (2010), ACM, pp. 1079–1088.
- [160] VIS, F., FAULKNER, S., PARRY, K., MANYUKHINA, Y., AND EVANS, L. Twitpic-ing the riots: analysing images shared on twitter during the 2011 uk riots. *Twitter and Society* (2013), 385–398.
- [161] VISWANATH, B., MISLOVE, A., CHA, M., AND GUMMADI, K. P. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks* (2009), ACM, pp. 37–42.
- [162] VITAK, J., ZUBE, P., SMOCK, A., CARR, C. T., ELLISON, N., AND LAMPE, C. It’s complicated: Facebook users’ political participation in the 2008 election. *CyberPsychology, behavior, and social networking* 14, 3 (2011), 107–114.
- [163] WANG, A. H. Don’t follow me: Spam detection in twitter. In *Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT)* (2010), IEEE, pp. 1–10.
- [164] WANG, Y., LEON, P. G., SCOTT, K., CHEN, X., ACQUISTI, A., AND CRANOR, L. F. Privacy nudges for social media: an exploratory facebook study. In *Proceedings of the 22nd international conference on World Wide Web companion* (2013), International World Wide Web Conferences Steering Committee, pp. 763–770.

- [165] WENG, J., AND LEE, B.-S. Event detection in twitter. In *Proceedings of the fifth International AAAI Conference on Weblogs and Social Media* (2011).
- [166] WILSON, C., BOE, B., SALA, A., PUTTASWAMY, K. P., AND ZHAO, B. Y. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems* (2009), ACM, pp. 205–218.
- [167] WOT. Web of trust api. <https://www.mywot.com/en/api> (Last Accessed: July 12, 2017) (2014).
- [168] XU, C., CETINTAS, S., LEE, K.-C., AND LI, L.-J. Visual sentiment prediction with deep convolutional neural networks. *arXiv preprint arXiv:1411.5731* (2014).
- [169] YANG, Z., WILSON, C., WANG, X., GAO, T., ZHAO, B. Y., AND DAI, Y. Uncovering social network sybils in the wild. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 8, 1 (2014), 2.
- [170] YOU, Q., LUO, J., JIN, H., AND YANG, J. Robust image sentiment analysis using progressively trained and domain transferred deep networks. *arXiv preprint arXiv:1509.06041* (2015).
- [171] YUAN, J., MCDONOUGH, S., YOU, Q., AND LUO, J. Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining* (2013), ACM, p. 10.
- [172] ZECH, M. Flight 17 spam scams on facebook, twitter. <http://www.nltimes.nl/2014/07/22/flight-17-spam-scams-facebook-twitter/> (Last Accessed: July 12, 2017) (2014).
- [173] ZHANG, J., DU, Z.-H., AND LIU, W. A behavior-based detection approach to mass-mailing host. In *International Conference on Machine Learning and Cybernetics* (2007), vol. 4, IEEE, pp. 2140–2144.
- [174] ZHANG, X., ZHU, S., AND LIANG, W. Detecting spam and promoting campaigns in the twitter social network. In *IEEE 12th International Conference on Data Mining (ICDM)* (2012), IEEE, pp. 1194–1199.