Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy

Aditi Gupta*, Hemank Lamba**, Ponnurangam Kumaraguru*, Anupam Joshi†
*Indraprastha Institute of Information Technology, Delhi, India
**IBM Research Labs, Delhi, India
†University of Maryland Baltimore County, Maryland, USA
{aditig, pk}@iiitd.ac.in, helamba1@in.ibm.com, joshi@cs.umbc.edu

ABSTRACT

In today's world, online social media plays a vital role during real world events, especially crisis events. There are both positive and negative effects of social media coverage of events, it can be used by authorities for effective disaster management or by malicious entities to spread rumors and fake news. The aim of this paper, is to highlight the role of Twitter, during Hurricane Sandy (2012) to spread fake images about the disaster. We identified 10,350 unique tweets containing fake images that were circulated on Twitter, during Hurricane Sandy. We performed a characterization analysis, to understand the temporal, social reputation and influence patterns for the spread of fake images. Eighty six percent of tweets spreading the fake images were retweets, hence very few were original tweets. Our results showed that top thirty users out of 10,215 users (0.3%) resulted in 90% of the retweets of fake images; also network links such as follower relationships of Twitter, contributed very less (only 11%) to the spread of these fake photos URLs. Next, we used classification models, to distinguish fake images from real images of Hurricane Sandy. Best results were obtained from Decision Tree classifier, we got 97% accuracy in predicting fake images from real. Also, tweet based features were very effective in distinguishing fake images tweets from real, while the performance of user based features was very poor. Our results, showed that, automated techniques can be used in identifying real images from fake images posted on Twitter.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Online social media, Twitter, crisis, fake pictures

1. INTRODUCTION

Over the past few years there has been increase in the usage of Online Social Media (OSM) services as a medium for people to share, coordinate and spread information about events while they are going on. Though a large volume of content is posted on OSM, not all of the information is of good quality with respect to the event, like it may be fake, incorrect or noisy. Extracting good quality information is one of the biggest challenges in utilizing information from OSM. Over last few years, people have highlighted how OSM can be used to help in extracting useful information about real life events. But, on the other hand, there have been many instances which have highlighted the negative effects on content on online social media on real life events. The information shared and accessed on social media such as Twitter, is in real-time, the impact of any malicious intended activity, like spreading fake images and rumors needs to be detected and curbed from spreading immediately. Such false and incorrect information can lead to chaos and panic among people on the ground. Since detecting whether images posted are fake or not, using traditional image analysis methods, can be highly time and resource consuming, we explore the option of using Twitter specific features, like the content of the tweet and the user details, in identifying fake images from real.

Hurricane Sandy: Hurricane Sandy caused mass destruction and turmoil in and around USA from October 22nd to October 31st, 2012. According to NBC News, the death toll in the U.S. was 109, including at least 40 in New York City. NBC also reported that damages from Hurricane Sandy exceeded $50 billion. Online social media such as Twitter and Facebook were widely used by people to keep abreast about latest updates of the storm. [1] Social media was also widely exploited by malicious entities during Sandy, to spread rumors and fake pictures in real-time. [2] [3] Such fake images and news became extremely viral on OSM and caused panic and chaos among the people affected by the hurricane. Hence, it is an ideal event, to analyze the spread and impact of fake and incorrect information on social media. Figure 1 shows some of the fake images that were spread during Hurricane Sandy, which we also found in our dataset.

There is dire need to build automated solutions that can help people judge the quality of information appearing on

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

[1] http://www.guardian.co.uk/world/us-news-blog/2013/feb/20/mta-conedison-hurricane-sandy-social-media-week
[2] http://news.yahoo.com/10-fake-photos-hurricane-sandy-075500934.html
[3] http://www.guardian.co.uk/news/datablog/2012/nov/06/fake-sandy-pictures-social-media

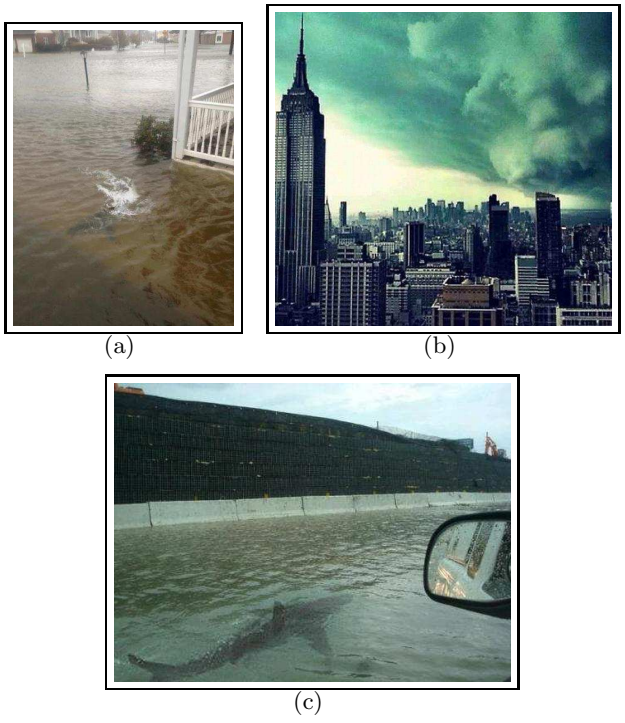


Figure 1: Some of the fake pictures of Hurricane Sandy that were shared on Twitter. (a) Picture of shark in New Jersey (b) Faked image of stormy New York skyline (c) Another picture of shark in the streets.

OSM in real-time. The aim of this work is to characterize and identify the propagation of fake pictures on OSM, Twitter. These fake images, created panic and chaos among the people. The affect of the spreading such false information can be multifold in case of crisis situations. Hence, we analyzed the propagation of fake images URLs during Hurricane Sandy. The power and impact of online social media in shaping real world events has been widely studied by researchers across the globe. To the best of our knowledge this is the first paper to study the diffusion and spread of fake pictures on OSM. The main contributions of this work are:

- We performed in-depth characterization of tweets sharing fake images on Twitter during Hurricane Sandy. We found that the tweets containing the fake images URLs were mostly retweets (86%), hence very few users posted original tweets with fake images. Also, we found that social network of a user on Twitter had little impact on making these fake images viral, there was just 11% overlap between the retweet and follower graphs of tweets containing fake images.
- We used classification algorithms to distinguish between tweets containing fake and real images. We primarily used two kinds of features: user level and tweet level features. Best accuracy of 97% was achieved using decision tree classifier, using tweet based features.

The rest of the paper is organized as follows: Section 2, describes the closely related work to this paper. Section 3 explains methodology that we used in collecting data, analyzing and classifying the tweets. Section 4 describes the analysis performed. Section 5 summarizes the results from our analysis and highlights the implications of our results. The last section presents the limitations, and future work of the paper.

2. RELATED WORK

2.1 Role of OSM during Real World Events

Role of social media has been analyzed by computer scientists, psychologists and sociologists for impact in the real-world. The OSM has progressed from being merely a medium to share users' opinions; to an information sharing and dissemination agent; to propagation and coordination of relief and response efforts. Palen et al. presented a path breaking vision on how Internet resources (technology and crowd based) can be used for support and assistance during mass emergencies and disasters [20]. They viewed people collectively as an important resource that can play a critical role in crisis. In a followup work to the above research proposal, Palen et al. studied two real world events, to understand and characterize the wide scale interaction on social networking websites with respect to the events [21]. The two events considered by them were: Northern Illinois University (NIU) shootings of February 14, 2008 and Virginia Tech (VT) tragedy 10 months earlier. Sakaki et al. used tweets as social sensors to detect earthquake events. They developed a probabilistic spatio-temporal model for predicting the center and trajectory of an event using Kalman and particle filtering techniques. Based upon the above models, they created an earthquake reporting application for Japan, which detected the earthquake occurrences based on tweets and sent users alert emails [23]. Sakaki et al. in a different research work, analyzed tweet trend to extract the events that happen during a crisis from the Twitter log of user activity analyzed Japanese tweets on all earthquakes during 2010 - 2011 [24]. Some of the prominent results obtained by them via statistical analysis, like tweet frequencies of feature phones and smart-phones were dominant just after the earthquake, although those of PCs was dominant in less-damaged areas. Cheong et al. performed social network analysis on Twitter data during Australian floods of 2011 to identify active players and their effectiveness in disseminating critical information [6].

Work has ben done to extract situational awareness information from the vast amount of data posted on OSM during real-world events. Vieweg et al. analyzed the Twitter logs for the Oklahoma Grassfires (April 2009) and the Red River Floods (March and April 2009) for presence of situational awareness content. An automated framework to enhance situational awareness during emergency situations was developed by Vieweg et al. They extracted geo-location and location-referencing information from users' tweets; which helped in increasing situation awareness during emergency events [26]. Verma et al. used natural language techniques to build an automated classifier to detect messages on Twitter that may contribute to situational awareness [25]. Another closely related work was done by Oh et al., where they analyzed Twitter stream during the 2008 Mumbai terrorist attacks [19]. Their analysis showed how information available on online social media during the attacks aided the terrorists in their decision making by increasing their *social awareness*. Corvey et al. analyzed one of the important aspects of applying computational techniques and algorithms

to social media data to obtain useful information for social media content, i.e. linguistic and behavioral annotations [8]. One important conclusion obtained by them was that during emergency situations, users use a specific vocabulary to convey tactical information on Twitter, as indicated by the accuracy achieved using bag-of-words model for situational awareness tweets classification. Mendoza et al. used the data from 2010 earthquake in Chile to explore the behavior of Twitter users for emergency response activity [15]. Their results showed that propagation of tweets related to rumors versus true news differed and could be used to develop automated classification solutions to identify correct information. Longueville et al. analyzed Twitter feeds during forest Marseille fire event in France. They showed information from location based social networks can be used to acquire spatial temporal data that can be analyzed to provide useful localized information about the event [9]. A team at National ICT Australia Ltd. (NICTA) has been working on developing a focused search engine for Twitter and Facebook that can be used in humanitarian crisis situation. [4] Hughes et al. in their work compared the properties of tweets and users during an emergency to normal situations [1]. They performed empirical and statistical analysis on their data collected during disaster events and showed an increase in the use of URLs in tweets and a decrease in @-mentions during emergency situations.

2.2 Assessing Quality of Information on OSM

Presence of spam, compromised accounts, malware, and phishing attacks are major concerns with respect to the quality of information on Twitter. Techniques to filter out spam / phishing on Twitter have been studied and various effective solutions have been proposed. Chhabra et al. highlighted the role of URL shortener services like *bit.ly* [5] in spreading phishing; their results showed that URL shorteners are used for not only saving space but also hiding the identity of the phishing links [7]. In a followup study Aggarwal et al. further analyzed and identified features that indicate to phishing tweets [2]. Using them, they detected phishing tweets with an accuracy of 92.52%. One of the major contributions of their work, was the Chrome Extension they developed and deployed for real-time phishing detection on Twitter. Grier et al. characterized spam spread on Twitter via URLs. They found that 8% of 25 million URLs posted on Twitter point to phishing, malware, and scams listed on popular blacklists [12]. Ghosh et al. characterized social farming on Twitter, and also proposed a methodology to combat link farming [11]. Yang et al. analyzed community or ecosystem of cyber criminals and their supporters on Twitter [28]. Yardi et al. applied machine learning techniques to identify spammers [29]. They used features (1) searches for URLs; (2) username pattern matches; and, (3) keyword detection; and obtained 91% accuracy. Benevenuto et al. classified real YouTube users, as spammers, promoters, and legitimates [3]. They used techniques such as supervised machine learning algorithms to detect promoters and spammers; they achieved higher accuracy for detecting promotors; the algorithms were less effective for detecting spammers. Nazir et al. provided insightful characterization of phantom profiles for gaming applications on Facebook [17]. They proposed a classification framework using SVM classifier for detecting phantom profiles of users from real profiles based on certain social network related features.

Now, we discuss some of the research work done to assess, characterize, analyze and compute trust and credibility of content on online social media. Truthy [6], was developed by Ratkiewicz et al. to study information diffusion on Twitter and compute a trustworthiness score for a public stream of micro-blogging updates related to an event to detect political smears, astroturfing, misinformation, and other forms of social pollution [22]. It works on real-time Twitter data with three months of data history. Castillo et al. showed that automated classification techniques can be used to detect news topics from conversational topics and assessed their credibility based on various Twitter features [5]. They achieved a precision and recall of 70-80% using J48 decision tree classification algorithms. They evaluated their results with respect to data annotated by humans as ground truth. Canini et al. analyzed usage of automated ranking strategies to measure credibility of sources of information on Twitter for any given topic [4]. The authors define a credible information source as one which has trust and domain expertise associated with it. Gupta et al. in their work on analyzing tweets posted during the terrorist bomb blasts in Mumbai (India, 2011), showed that majority of sources of information are unknown and with low Twitter reputation (less number of followers) [14]. This highlights the difficulty in measuring credibility of information and the need to develop automated mechanisms to assess credibility of information on Twitter. The authors in a follow up study applied machine learning algorithms (SVM Rank) and information retrieval techniques (relevance feedback) to assess credibility of content on Twitter [13]. They analyzed fourteen high impact events of 2011; their results showed that on average 30% of total tweets posted about an event contained situational information about the event while 14% was spam. Only 17% of the total tweets posted about the event contained situational awareness information that was credible. Another, very similar work to the above was done by Xia et al. on tweets generated during the England riots of 2011 [27]. They used a supervised method of Bayesian Network is used to predict the credibility of tweets in emergency situations. Donovan et al focussed their work on finding indicators of credibility during different situations (8 separate event tweets) were considered. Their results showed that the best indicators of credibility were URLs, mentions, retweets and tweet length [18]. A different methodology, than the above papers was followed by Morris et al., who conducted a survey to understand users perceptions regarding credibility of content on Twitter [16]. They asked about 200 participants to mark what they consider are indicators of credibility of content and users on Twitter. They found that the prominent features based on which users judge credibility are features visible at a glance, for example, username and picture of a user. Another approach to detect users with high value users of credibility and trustworthiness was taken by Ghosh et al., they identified the topic based experts on Twitter [10]. Their techniques rely on the wisdom of the Twitter crowds - i.e. they used the Twitter Lists feature to identify experts in various topics.

[4]http://leifhanlen.wordpress.com/2011/07/22/crisis-management-using-twitter-and-facebook-for-the-greater-good/

[5]https://bitly.com/

[6]http://truthy.indiana.edu/

3. METHODOLOGY

In this section, we discuss our research methodology in detail. First we describe the methodology of collecting data from Twitter, followed by the various analytical techniques applied in this paper.

3.1 Data

For data collection from Twitter we have a $24 * 7$ setup, which has been functional for about last 20 months. We collected data from Twitter using the *Streaming API*. [7] This API enables researchers to extract tweets in real-time, based on certain query parameters like words in the tweet, time of posting of tweet, etc. We queried the Twitter *Trends API* after every hour for the current trending topics, [8] and collect tweets corresponding to these topics as query search words for the *Streaming API*.

Hurricane Sandy's impact lasted from Oct. 20th to Nov. 1st, 2012, hence from all the tweets collected during this period, we filtered out tweets containing the words '*sandy*' and '*hurricane*'. We filtered out about 1.8 million tweets by 1.2 million unique users on Hurricane Sandy from Oct. 20th to Nov. 1st, 2012. Table 1 gives the descriptive statistics of the tweets and users data collected to the event, and Figure 2 shows the spatial distribution of these tweets (about 19K tweets had geo-location embedded in them).

Table 1: Descriptive statistics of the Twitter dataset for Hurricane Sandy.

Total tweets	1,782,526
Total unique users	1,174,266
Tweets with URLs	622,860

Using certain online resources (articles, tweets and blogs) we were able to identify certain URLs that belonged to fake pictures of Hurricane Sandy. One of the prominent data sources used by us was the list of fake and real images made public by the Guardian news media company. [9] The list provided by Gaurdian, classified the top image URLs shared during the hurricane as fake or real image URLs, which we used to form our dataset. There were many other articles and blogs that covered the real and fake images that were spread on Twitter. [10] [11] [12] Table 2 describes the statistics for data related to tweets containing fake and real image URLs. We identified eight unique fake images of Sandy that were spread on Twitter in our dataset, we collected about 10K tweets for these URLs.

3.2 Characterization Analysis

We performed characterization of the tweets containing fake images URLs and their propagation, to understand how they became viral. First we performed temporal analysis on the fake images tweets. We analyzed how many such tweets were shared per hour on Twitter. Also, we analyzed the sudden peaks (from $x1$ hour to $x1+1$) in the graph more closely. We constructed the retweet graph for the sudden peak in the temporal analysis, to find out what changes in the network topology lead to the viral spread of these images. We obtained certain useful insights, about the nature and spread of fake image URLs on Twitter, which are summarized in the next section

Next, we analyzed what role the social network graph of a user on Twitter plays in propagation of fake URLs. The explicit social network of a user on Twitter, is that of his follower graph. We wanted to analyze what percentage of information diffusion takes place via this follower network graph of a user. The details of the algorithm used to compute are summarized in Algorithm 1.

Figure 2: Spatial distribution of total tweets on Hurricane Sandy. Here we have plotted about 19K tweets, which had embedded geo-location data in them.

Table 2: Descriptive statistics of the tweets with fake and real images URLs.

Tweets with fake images	10,350
Users with fake images	10,215
Tweets with real images	5,767
Users with real images	5,678

Algorithm 1 Compute_Overlap

```
1: Create_Graph_Retweets()
2: Create_Graph_Followers()
3: for each edge in the retweet network do
4:     num_retweet_edges++
5:     Insert edge into hashmap, H[1..n]
6: end for
7: for each edge in the follower network do
8:     Insert each edge in hashmap, H[1..n]
9:     if collision then
10:        intersections++
11:    end if
12: end for
13: %overlap = (intersections/num_retweet_edges) * 100
```

In the function, *Create_Graph_Followers*, we crawled the follower network of all the unique users that had tweeted the fake images, using the REST API of Twitter. The network created had 10,779,122 edges and 10,215 nodes. In *Create_Graph_Retweets*, we created a retweet network, where an

[7] https://dev.twitter.com/docs/streaming-api.
[8] https://dev.twitter.com/docs/api/1/get/trends
[9] http://www.guardian.co.uk/news/datablog/2012/nov/06/fake-sandy-pictures-social-media
[10] http://now.msn.com/hurricane-sandy-fake-photos
[11] http://mashable.com/2012/10/29/fake-hurricane-sandy-photos/
[12] http://theweek.com/article/index/235578/10-fake-photos-of-hurricane-sandy

edge between two nodes exists if one user had retweeted the other's tweet. A hashmap, $H[1..n]$, is created to compute the overlap between the follower and retweets graphs.

3.3 Classif cation Analysis

We analyzed the effectiveness of machine learning algorithms in detecting tweets containing fake image URLs versus tweets containing real images of Sandy. We performed two-class classification using Naive Bayes and J48 Decision Tree classifiers. We had a dataset of 10,350 tweets containing fake image URLs and 5,767 tweets containing real images URLs. To avoid any bias, due to unequal size of any of the classes, we randomly selected 5,767 tweets from the fake images tweets, and then applied classification.

We used two kinds of features, for the classification algorithm. Table 3 summarizes the features computed by us for each tweet and the user of the tweet.

- **Source or user level features** [F1]: The attributes of the user who posted the tweet. We consider properties such as number of friends, followers and status messages of the user as part of this set.
- **Content or tweet level features** [F2]: The 140 characters posted by users contain data (e.g. words, URLs, hashtags) and meta-data (e.g. is tweet a reply or a retweet) related to it.

User Features [F1]
Number of Friends
Number of Followers
Follower-Friend Ratio
Number of times listed
User has a URL
User is a verified user
Age of user account
Tweet Features [F2]
Length of Tweet
Number of Words
Contains Question Mark?
Contains Exclamation Mark?
Number of Question Marks
Number of Exclamation Marks
Contains Happy Emoticon
Contains Sad Emoticon
Contains First Order Pronoun
Contains Second Order Pronoun
Contains Third Order Pronoun
Number of uppercase characters
Number of negative sentiment words
Number of positive sentiment words
Number of mentions
Number of hashtags
Number of URLs
Retweet count

Table 3: User and tweet based features used for classification of fake and real images of Sandy.

4. RESULTS

In this section, we summarize the results obtained for the characterization and classification analysis performed.

4.1 Characterization Results

We found that out of the 10,350 tweets identified by us, containing fake images URLs, about 86% were retweets. That is, only about 14% people posted original tweeted that contained such URLs. From the temporal analysis, we plotted the per hour tweeting activity of the fake images URLs. From Figure 3 we see that the fake URLs spread spikes at, 12 hours after the introduction of the URLs in the Twitter network. We now analyze the spread of these picture URLs one hour before and after the spike. We construct the reply and retweet graph for the tweets sharing these fake picture URLs on October 29th, at 21 hours and 22 hours, as shown in Figure 5. We see that there are only a few users with very high degree, that is, only a few users results in majority of the retweets. We confirmed this statistically, Figure 4 (CDF) shows that top 30 users (0.3% of the users) resulted in 90% of retweets of the fake images. Combining results from both the graphs, we conclude that though the fake URLs were present in the Twitter network for almost 12 hours before they became viral, also the sudden spike in their propagation via retweets happened only because of a few users.

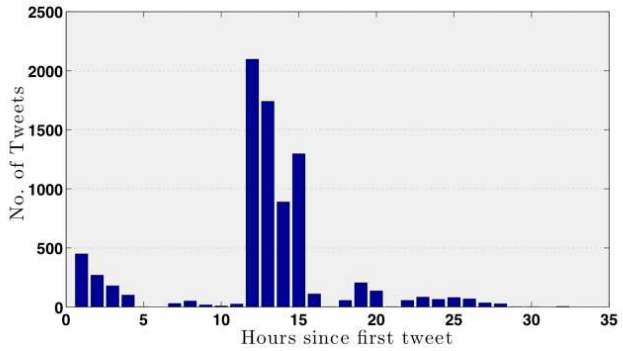
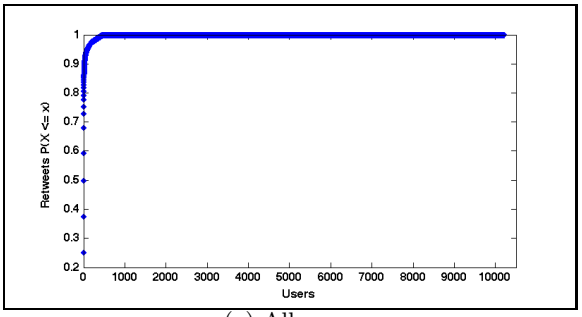


Figure 3: Details of data collected for the fake images URL sharing. Temporal distribution of tweets, hour wise, starting from the first hour that a fake image tweet was posted.

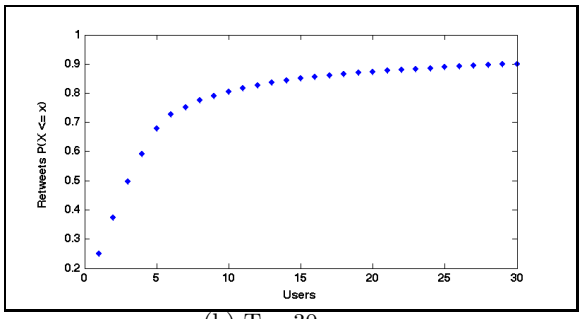
Next, we determine the role of Twitter network graph on the retweets propagation of the fake image tweets. We ran the *Compute_overlap* algorithm discussed above. We found the number of overlapping edges as 1,215, which leads to a percentage overlap of 11% between the retweet and follower graphs. Table 4 summarizes the results of the *Compute_overlap* algorithm. This indicates that there was a very limited retweet activity which originated because of the people in a user's follower graph. Hence, in cases of crisis, people often retweet and propagate tweets that they find in Twitter search or trending topics, irrespective of whether they follow the user or not.

4.2 Classif cation Results

In the above section, we characterized the properties and behavior associated with spread of false information, in form of fake images, on Twitter. The next important step is to



(a) All users



(b) Top 30 users

Figure 4: CDF of retweets of the fake image tweets by the users. It shows that top 30 users (0.3% of the users) resulted in 90% of retweets of the fake images

Total edges in the retweet network	10,508
Total edges in the follower-followee network	10,799,122
Total edges that exist in both retweet network and the follower-followee network	1,215
%age overlap	**11%**

Table 4: Results of the Algorithm *Compute_overlap*. We found only 11% overlap between the follower and retweet graphs for the tweets containing fake images.

explore features and algorithms that can effectively help us is identifying the fake content in real-time. We performed 10-fold cross validation while applying classification models. We applied two standard algorithms used for classification: Naive Bayes and Decision Tree (J48). As described before, we took 5,767 tweets for both fake and real image containing tweets. For each data point, we created user and tweet level feature vectors. Table 5 summarizes the results from the classification experiment. We achieve a good accuracy of above 90% for both classifiers, though Decision Tree out performs the Naives Bayes classifier. We can also see that, user based features, provide very poor accuracy in distinguishing fake image URLs, while tweet based features perfumed very well. We would also like to mention that high accuracy results obtained by us, may be attributed to the similar nature of many tweets (since a lot of tweets are retweets of other tweets in our dataset). We can conclude that, content and property analysis of tweets can help us in identifying real image URLs being shared on Twitter with a high accuracy.

	F1	F2	F1+F2
Naive Bayes	56.32%	91.97%	91.52%
Decision Tree	53.24%	**97.65%**	96.65 %

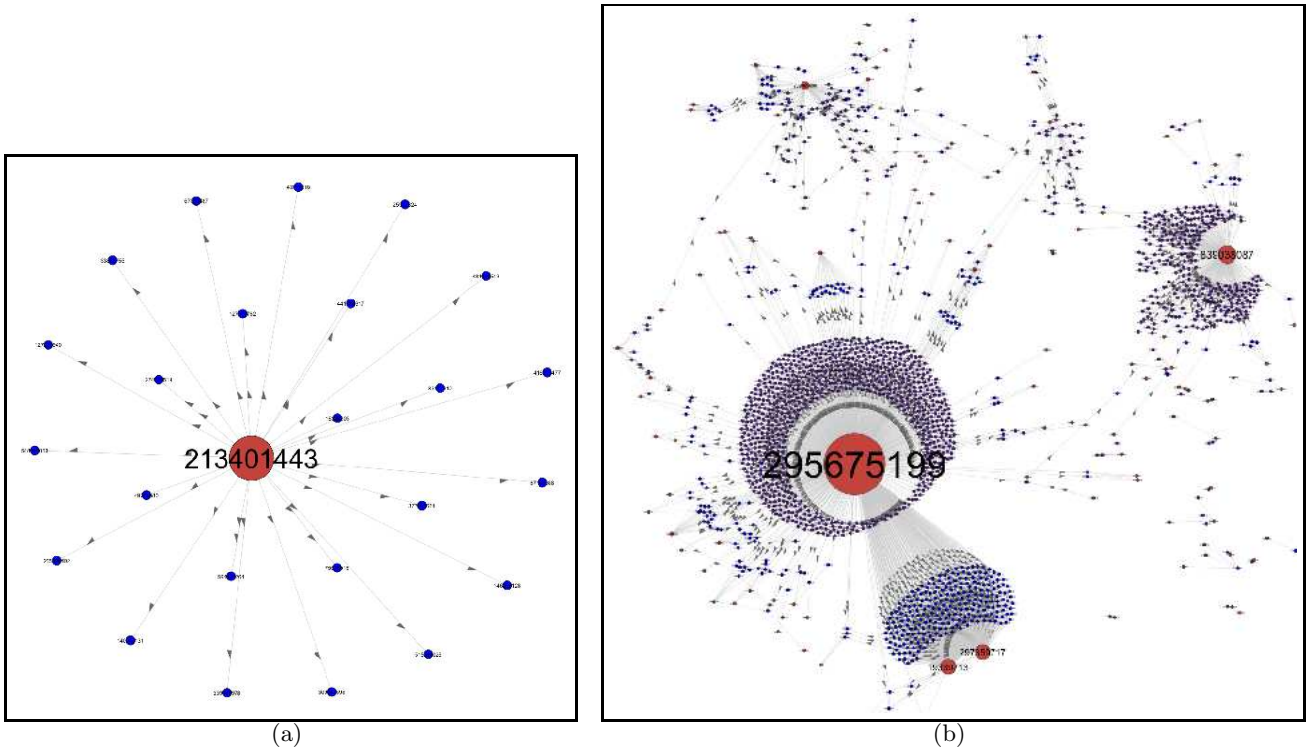
Table 5: Classification results for tweets containing fake image and real images. Our results showed that, tweet based features are more effective in distinguishing the two classes.

5. DISCUSSION

Online social media has the capability of playing the role of, either a life saver or that of a daemon during the times of crisis. In this research work, we highlighted one of the malicious intended usage of Twitter during a real-world event. We analyzed the activity on the online social networking website Twitter, during Hurricane Sandy (2012) that spread fake images. We identified 10,350 unique tweets containing fake images that were circulated on Twitter, during Hurricane Sandy. We performed a characterization analysis, to understand the temporal, social reputation and influence patterns of the spread of these fake images. We found that 86% tweets spreading the fake images were retweets, hence very few were original tweets by users. Also, our results showed that top 30 users (0.3% of the users) resulted in 90% of retweets of the fake image. Hence, we can concluded that only a handful of users contributed to majority of the damage, via the retweeting activity on the Twitter. We analyzed the role of Twitter social graph in propagating the fake images. We crawled the network links, that is, the follower relationships of the users and applied our algorithm to compute the overlap. We found only a 11% overlap between the retweet and follower graphs for the users who tweeted fake images of Sandy. This result highlights the fact that, at the time of crisis, users retweet information from other users irrespective of the fact whether they follow them or not. Next, we used classification models, to identify fake images from real images of Hurricane Sandy. Best results were obtained from Decision Tree classifier, we got 97% accuracy in predicting fake images from real. Tweet based features are very effective in distinguishing fake images tweets from real, while the performance of user based features was very poor. Our research work provided insights into the behavioral pattern of the spread of fake image tweets. Also our results provided a proof of concept that, automated techniques can be used in identifying real images from fake images posted on Twitter.

6. FUTURE WORK

The work done by us, provides a proof of concept that automated techniques can be used to identify malicious or fake content spread on Twitter during real world events. We would like to conduct a larger study with more events for identification of fake images and news propogation. Also, we would like to expand our study, to detecting rumors and other malicious content spread during real world events apart from images. As a next step, we would like to develop a browser plug-in that can detect fake images being shared on Twitter in real-time.



(a) (b)

Figure 5: Spread of fake pictures URLs (retweet and reply graph), the number on the node is user profile ID on Twitter. The figure shows that the fake images became viral very fast, within an hour there was a tremendous growth in the number of people tweeting them. (a) Oct. 29, 2100 hours (b) Oct. 29, 2200 hours.

7. ACKNOWLEDGMENTS

We would like to thank Government of India for funding this project. We would like to express our sincerest thanks to all members of PreCog research group at IIIT, [13] Delhi, for their continued support and feedback on the project.

8. REFERENCES

[1] Leysia Palen Amanda L. Hughes. Twitter Adoption and Use in Mass Convergence and Emergency Events. *ISCRAM Conference*, 2009.

[2] Ponnurangam Kumaraguru Anupama Aggarwal, Ashwin Rajadesingan. Phishari: Automatic realtime phishing detection on twitter. *7th IEEE APWG eCrime Researchers Summit (eCRS)*, 2012.

[3] Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 620–627, New York, NY, USA, 2009. ACM.

[4] Kevin R. Canini, Bongwon Suh, and Peter L. Pirolli. Finding credible information sources in social networks based on content and social structure. In *SocialCom*, 2011.

[5] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 675–684, New York, NY, USA, 2011. ACM.

[6] France Cheong and Christopher Cheong. Social media data mining: A social network analysis of tweets during the 2010-2011 australian floods. In *PACIS*, 2011.

[7] Sidharth Chhabra, Anupama Aggarwal, Fabricio Benevenuto, and Ponnurangam Kumaraguru. Phi.sh/$ocial: the phishing landscape through short urls. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, CEAS '11, pages 92–101, New York, NY, USA, 2011. ACM.

[8] William J. Corvey, Sudha Verma, Sarah Vieweg, Martha Palmer, and James H. Martin. Foundations of a multilayer annotation framework for twitter communications during crisis events. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[9] Bertrand De Longueville, Robin S. Smith, and Gianluca Luraschi. "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 *Proceedings of the 2009 International Workshop on Location Based Social Networks*, LBSN '09, pages 73–80, New York, NY, USA, 2009. ACM.

[10] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi.

[13] precog.iiitd.edu.in

Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, 2012.

[11] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna PhaniGummadi. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, 2012.

[12] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, CCS '10, pages 27–37, New York, NY, USA, 2010. ACM.

[13] Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12, pages 2:2–2:8, New York, NY, USA, 2012. ACM.

[14] Aditi Gupta and Ponnurangam Kumaraguru. Twitter explodes with activity in mumbai blasts! a lifeline or an unmonitored daemon in the lurking? IIIT, Delhi, Technical report, IIITD-TR-2011-005, 2011.

[15] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 71–79, New York, NY, USA, 2010. ACM.

[16] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, New York, NY, USA, 2012. ACM.

[17] Atif Nazir, Saqib Raza, Chen-Nee Chuah, and Burkhard Schipper. Ghostbusting facebook: detecting and characterizing phantom profiles in online social gaming applications. In *Proceedings of the 3rd conference on Online social networks*, WOSN'10, 2010.

[18] J. O'Donovan, B. Kang, G. Meyer, T. HŽllerer, and S. Adali. Credibility in context: An analysis of feature distributions in twitter. *ASE/IEEE International Conference on Social Computing, SocialCom*, 2012.

[19] Onook Oh, Manish Agrawal, and H. Raghav Rao. Information control and terrorism: Tracking the mumbai terrorist attack through twitter. *Information Systems Frontiers*, 13(1):33–43, March 2011.

[20] Leysia Palen, Kenneth M. Anderson, Gloria Mark, James Martin, Douglas Sicker, Martha Palmer, and Dirk Grunwald. A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In *Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference*, ACM-BCS '10, 2010.

[21] Leysia Palen and Sarah Vieweg. The emergence of online widescale interaction in unexpected events: assistance, alliance & retreat. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, CSCW '08, pages 117–126, New York, NY, USA, 2008. ACM.

[22] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. WWW '11, 2011.

[23] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[24] Takeshi Sakaki, Fujio Toriumi, and Yutaka Matsuo. Tweet trend analysis in an emergency situation. In *Proceedings of the Special Workshop on Internet and Disasters*, SWID '11, pages 3:1–3:8, New York, NY, USA, 2011. ACM.

[25] Sudha Verma, Sarah Vieweg, William Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.

[26] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1079–1088, New York, NY, USA, 2010. ACM.

[27] Xin Xia, Xiaohu Yang, Chao Wu, Shanping Li, and Linfeng Bao. Information credibility on twitter in emergency situation. In *Proceedings of the 2012 Pacific Asia conference on Intelligence and Security Informatics*, PAISI'12, 2012.

[28] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, 2012.

[29] Sarita Yardi, Daniel Romero, Grant Schoenebeck, and Danah Boyd. Detecting spam in a Twitter network. *First Monday*, 15(1), January 2010.