Analyzing and Measuring the Spread of Fake Content on Twitter during High Impact Events

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. Malicious content is posted online during these events, which can result in damage, chaos and monetary loss in the offline world. In our paper, we highlight the role of Twitter in *two* major crisis events: *Hurricane Sandy* and *Boston Marathon Bombings* in spreading fake content about the events. We performed a characterization analysis, to understand the temporal, social reputation and influence patterns for the spread of such fake information. Our results indicate that automated techniques can be used to identify characteristics of fake information on Twitter.

1. INTRODUCTION

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 of content posted on online social media during real life events. The information shared and accessed on social media such as Twitter, is in real-time, and the impact of any malicious intended activity, like spreading fake images and rumors needs to be detected and curbed from spreading immediately. We analyze and characterize the various type of fake content that emerged during these two events - images, text, suspended profiles.

2. DATA

We collected data from Twitter using the Streaming API. We queried the Twitter Trends API after every hour for the current trending topics, and collect tweets corresponding to these topics as query search words for the Streaming API.

Table 1: Descriptive statistics of datasets.

Event	Boston Blasts	Hurricane Sandy
Total tweets	7,888,374	1,782,526
Total users	3,677,531	1,174,266
Tweets with URLs	3,420,228	622,860

Boston Marathon Blasts: Twin blasts occurred during the Boston Marathon on April 15th, 2013 at 18:50 GMT. Three people were killed and 264 were injured in the incident. We annotated the top 20 most popular tweets to the following three categories: Fake / Rumor , True and Not Applicable (NA).

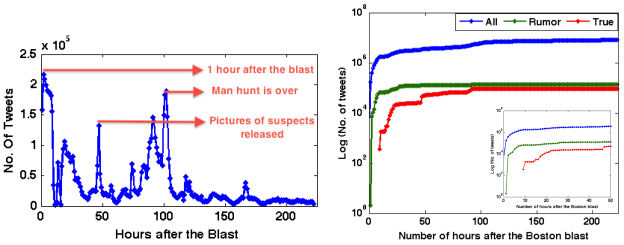
Hurricane Sandy: Hurricane Sandy caused mass destruction and turmoil in and around USA from October 22nd to October 31st, 2012. Social media was widely exploited by malicious entities during Sandy, to spread rumors and fake pictures in real-time. We identified eight unique fake images of Sandy that were spread on Twitter in our dataset, we collected about 10,350 tweets for these URLs.

(a) (b)

Figure 1: Some of the fake tweets and pictures that were shared on Twitter during the two events.

3. CHARACTERIZATION ANALYSIS

To analyze the temporal distribution of the tweets posted during the events, we calculate the number of tweets posted in each hour after the event occurred. In case of Boston blasts, we observed that for the first 7-8 hours, NA and Fake tweets were observed. The spread of true tweets started only after eight hours from the time of the blasts. For Sandy event, we saw that fake tweets pick up propagation 10 hours after the event.



(a) (b)

Figure 2: (a) Timeline for tweets collected for the Boston marathon blasts. (b) The log distribution for the number of the total, fake, true information and NA category tweets.

3.1 Suspended Account Analysis

We aim to identify the characteristics and activities of malicious new accounts created during the Boston marathon blasts. We identified 31,919 new Twitter accounts that were created during the Boston blasts tragedy [Apr. 15th - Apr. 20th], that also tweeted atleast one tweet about the event. Out of these 19% [6,073 accounts] were deleted or suspended by Twitter, when we checked two months after the blasts. Some of these accounts were quite influential during the Boston tragedy too. Next, we tried to find out how affective were these accounts during the Boston marathon events. We constructed a network graph $G = (V, E)$ for the interaction between these newly created malicious profiles. Where each node in V represents a suspended user profile, and an edge between two suspended nodes represents a retweet, reply or mention action by them.

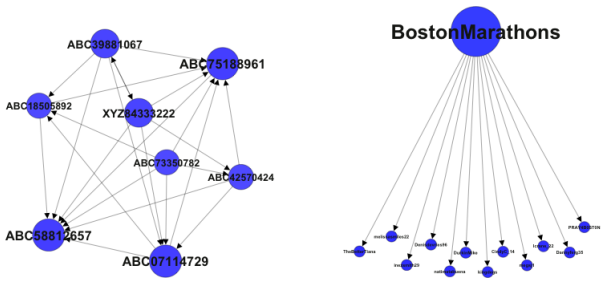


Figure 3: Network of suspended accounts created during Boston blasts. We show two different forms of interactions amongst suspended profiles (left to right): Closed Community and Star Topology.

- *Closed Community:* We observed a community of users who retweet and mention each other, and form a closed community. All these nodes had similar usernames, all usernames have the same prefix and only numbers in the suffixes are different. This indicates that either these profiles were created by same or similar minded people for posting common propaganda posts. These twelve accounts were all tweeting the same propaganda and hate filled tweet.

- *Star Topology:* A fake account *BostonMarathons* was created similar to the original Twitter account *bostonmarathon*, resulting in users getting confused between the two, leading to a lot of propagation of content by the fake BostonMarathons profile. Impersonation or creating fake profiles is a crime that results in identity theft and is punishable by law in many countries.

3.2 Twitter Follower Network Analysis

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 2 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, irrespective of whether they follow the user or not.

Algorithm 1 Compute_Overlap

```
Create_Graph_Retweets()
Create_Graph_Followers()
for each edge in the retweet network do
    num_retweet_edges++
    Insert edge into hashmap, H[1..n]
end for
for each edge in the follower network do
    Insert each edge in hashmap, H[1..n]
    if collision then
        intersections++
    end if
end for
%overlap = (intersections/num_retweet_edges) * 100
```

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

3.3 Classification Results

The next important step is to 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). 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 3 summarizes the results from the classification experiment.

	User Features	**Tweet Features**	**Total**
Naive Bayes	56.32%	91.97%	91.52%
Decision Tree	53.24%	**97.65%**	96.65 %

Table 3: Classification results for tweets containing fake image and real images during Hurricane Sandy.

4. CONCLUSIONS

Our research work provided insights into the behavioral pattern of spread of fake images and rumors on Twitter. Our results provided a proof of concept that, automated techniques can be used in identifying fake content posted on Twitter.

5. ACKNOWLEDGMENTS

The work presented in this paper has been published as:
Gupta, A., Lamba, H., and Kumaraguru, P. $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing Fake Content on Twitter. Accepted at IEEE APWG eCrime Research Summit (eCRS), 2013.
Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy. Accepted at Workshop on Privacy and Security in Online Social Media (PSOSM), in conjunction with World Wide Web Conference (WWW) (2013). Best Paper Award