Fake and Spam Messages: Detecting Misinformation during Natural Disasters on Social Media

Meet Rajdev and Kyumin Lee
Department of Computer Science
Utah State University
Logan, UT 84322
meetrajdev@aggiemail.usu.edu, kyumin.lee@usu.edu

***Abstract*—During natural disasters or crises, users on social media tend to easily believe contents of postings related to the events, and retweet the postings with hoping them to be reached to many other users. Unfortunately, there are malicious users who understand the tendency and post misinformation such as spam and fake messages with expecting wider propagation. To resolve the problem, in this paper we conduct a case study of 2013 Moore Tornado and Hurricane Sandy. Concretely, we (i) understand behaviors of these malicious users; (ii) analyze properties of spam, fake and legitimate messages; (iii) propose flat and hierarchical classification approaches; and (iv) detect both fake and spam messages with even distinguishing between them. Our experimental results show that our proposed approaches identify spam and fake messages with 96.43% accuracy and 0.961 F-measure.**

***Keywords*-fake; spam; social media; detection**

I. Introduction

A property of social media sites is real-time communication. People post news or their opinions in near-real time. Especially, when natural disasters (e.g., hurricane and tornado) and outbreaks (e.g., Ebola) happened, they post news and information regarding the events, express concerns, and pray for victims. People pay more attention on postings related to these crises and tend to easily believe contents of the postings. Unfortunately, there are malicious users who know the tendency, and post and propagate misinformation such as fake and spam information. For example, when hurricane Sandy happened, malicious users posted relevant messages with fake images [1]. These messages were retweeted by many users who believed retweeting the messages would help the victims affected by the Hurricane Sandy.

Researchers [1], [2] analyzed fake contents or studied a fake image detection problem. Other researchers [3], [4] studied s spam message detection problem. However, they focused on only one event in a narrow scope or only one problem (either fake image or spam message detection). In practice, fake and spam messages should be detected at once, and even distinguishing fake and spam messages is required.

To resolve the problem, in this paper we conduct a case study of 2013 Moore Tornado and Hurricane Sandy by answering following research questions: (i) Do fake message posters and spammers have different behaviors from legitimate users?; (ii) Do fake, spam and legitimate messages have distinguishing patterns?; and (iii) Can we automatically detect fake and spam messages?

II. Dataset

A. Collecting Dataset

Since we are interested in analyzing and detecting misinformation during natural disasters, we selected two well-known natural disasters: (i) Moore Tornado; and (ii) Hurricane Sandy. Detailed data collection strategy is described as follows:

2013 Moore Tornado. The Moore Tornado struck Moore, Oklahoma on May 20, 2013. Twitter users had posted about the Moore Tornado between 19th and 24th May (i.e., right before and after the tornado reaching to Moore, Oklahoma). Initially, we collected 158,000 tweets posted between 19th and 24th May by using Twitter streaming API. After removing irrelevant tweets, 9,284 relevant tweets were left.

Hurricane Sandy. The hurricane Sandy developed on October 22, 2012 and lasted 10 days (i.e., until October 31). Initially, we collected 3,251,083 tweets posted between October 22 and 31. Then we removed irrelevant tweets. Out of 3.2 million tweets, 34,054 tweets were relevant to Hurricane Sandy.

B. Labeling Dataset

Given a set of tweets relevant to each event, in this subsection, we first define three categories of the tweets and then label them based on the categories.

- **Fake tweet:** A tweet is defined as fake if it satisfies at least one of the following conditions:
  - incorrect location related to the event
  - incorrect time/date related to the event
  - some other incorrect information related to the event
  - link to misleading/ fake image
- **Spam tweet:** A tweet is labeled as spam if it satisfies at least one of the following conditions:
  - link to a spam page (pharmacy, loans, etc)
  - link to a pornographic content

Name	|Relevant Tweets|	|Labeled Tweets|
Moore Tornado	9,284	1,050
Hurricane Sandy	34,054	1,051

Table I
DATASETS.

- link to advertisements (personal agendas, etc)

- **Legitimate tweet:** A tweet is neither fake nor spam.

We also define a tweet is non-legitimate if the tweet is either spam or fake. In other words, non-legitimate tweets consist of spam and fake tweets.

Two human labelers independently classified each tweet to either non-legitimate or legitimate. If a tweet was labeled as a non-legitimate tweet, they further labeled it to either spam or fake. Finally, the two human labelers achieved 96% agreement of the labeling.

Since labeling all the tweets take a long time, we randomly selected 1,050 out of 9,284 tweets relevant to 2013 Moore Tornado and labeled them. Specifically, the Moore Tornado dataset consisted of 350 non-legitimate (i.e., 21 fake tweets and 329 spam tweets) and 700 legitimate tweets. Likewise, out of 34,054 tweets relevant to Hurricane Sandy, we randomly selected 1,051 tweets. 701 tweets were labeled as legitimate tweets and 350 tweets were labeled as non-legitimate tweets. Out of 350 non-legitimate tweets, 69 tweets were fake and 281 were spam. Table I shows our datasets that are used in the rest of the paper.

III. Analysis

In this section, we analyze the labeled datasets to see whether we can find distinguishing patterns among legitimate, spam and fake messages.

Figure 1 shows cumulative distribution functions (CDFs) of a ratio of number of friends and number of followers in each user category. As we can see in the CDF, the friends to followers ratio of users, who posted spam tweets, is lower than users who posted legitimate tweets. Some of users who posted fake tweets had also low friends to followers ratio. The most number of peaks for this ratio is seen for legitimate tweets. This is because a legitimate user uses the social media platform for getting useful information, and he follows relevant people for this purpose (i.e., friends). But in case of spammers and some fake tweet posters, they uses the social media platform to spread spam and fake contents, i.e., the focus is on getting more and more followers, not friends. As a result these users have a low friends to followers ratio.

Figure 2 shows CDFs of the number of favorited tweets by users in each category. Users who posted fake tweets tend to favorite less number of tweets than legitimate users and spammers. Interestingly, spammers favorited slightly more number of tweets than legitimate users.

Sometimes users post tweets with a hashtag for various reasons (e.g., summarize the tweet, a topic of the tweet

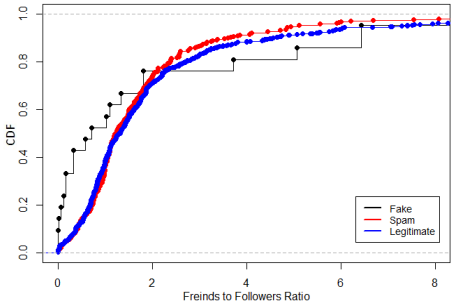

Figure 1. User-focused feature: Friends to followers ratio of users.

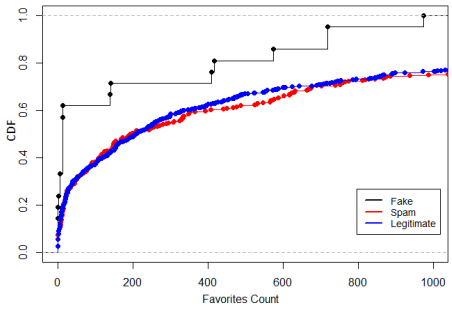

Figure 2. User-focused feature: Favorites count of users.

and a name of the place). We are interested to analyze what kind of hashtags people added to their tweets during the natural disasters. We groups hashtags in tweets to five categories as following: (i) *Place* (e.g., #Oklahoma, #Moore...); (ii) *Event* (e.g., #Hurricane, #Tornado...); (iii) *Pray, place* (e.g., #prayforoklahoma...); (iv) *Pray, event* (e.g., #praytornadookc...); and (v) *Others* (e.g., #save..., #help). Legitimate tweets contained hashtags about praying for the targeted (damaged) place (e.g., #prayforokc). But hashtags in fake and spam tweets were related to a place, an event and other keywords like #Moore, #oklahoma, #Tornado, etc.

IV. Proposed Approach and Features

A. Classification Approach

We propose two classification approaches – (i) flat classification; and (ii) hierarchical classification. Flat classification approach classifies a tweet to a spam, fake or legitimate tweet. Unlike the flat classification approach, hierarchical classification approach consists of two steps. The first step is to classifies a tweet to a legitimate or non-legitimate (again, including spam and fake) tweet. Then, the second step is to classify a predicted non-legitimate tweet to a spam or fake tweet. We develop spam and fake tweet classifiers based on each approach, and test which approach gives us better prediction results.

B. Features

To build classifiers, we extracted user features and tweet features. User features consist of (i) the number of tweets the user has favorited in the user account's lifetime; (ii) ratio

Table II
FLAT CLASSIFICATION RESULTS IN 2013 MOORE TORNADO DATASET.

Classifier	Accuracy	F-measure	Precision	Recall
FT	**88.67%**	**0.887**	0.887	0.887
NBTree	86.38%	0.862	0.874	0.864
Random Forest	88.48%	0.884	0.885	0.885

of the number of friends and followers; (iii) the number of followers; (iv) the number of friends; (v) did the user enable the possibility of geotagging his Tweets?; (vi) length of the screen name of the user; (vii) the number of public lists that the user is member of; (viii) did the user define his location in his profile?; (ix) the number of tweets that the user has posted; (x) time zone that the user declares himself within; (xi) does the user profile include a URL?; (xii) longevity of the user account (i.e., when was it created?); (xiii) does the user profile contain a user name?; and (xiv) is the user account verified by Twitter?.

Tweet features consist of (i) n-gram features; (ii) a hashtag type (e.g., place, event, pray with a place, pray with an event name, and others) as we mentioned in the previous section; (iii) tweet creation time (UTC time); and (iv) the number of URLs in a tweet. As the n-gram features, we extracted unigram, bigram and trigram features from our datasets, and applied feature selection to keep only significant features. For example, we initially extracted 1,334 and 700 features from Moore Tornado and hurricane Sandy datasets, respectively. After feature selection, we finally used 675 and 343 features for Moore Tornado and hurricane Sandy, respectively.

V. DETECTING SPAM AND FAKE TWEETS

Based on the proposed features, we now build flat and hierarchical classifiers to detect spam and fake tweets.

A. 2013 Moore Tornado

1) Flat Classification: In flat classification, given a tweet, our classifier classifies it to a spam, fake or legitimate tweet. We applied 10-fold cross validation on each dataset for flat classification.

We developed FT classifier, NBTree classifier and Random Forest classifier based on flat classification approach. Table II shows classification results. FT classifier outperformed NBTree classifier and Random Forest classifier by achieving 88.67% accuracy and 0.887 F-measure. From confusion matrix of FT classifier, we noticed that 91% legitimate tweets and 86% spam tweets were correctly classified. But, the identification of fake tweets was not good with 57% accuracy. Almost 24% fake tweets were classified to spam tweets. It indicates that users posting fake tweets behaved similar to spammers.

2) Hierarchical Classification: To conduct hierarchical classification, we randomly split each dataset into training (containing 2/3 data) and testing sets (containing 1/3 data).

Table III
STEP 1: HIERARCHICAL CLASSIFICATION RESULTS (LEGITIMATE TWEETS VS. NON-LEGITIMATE TWEETS) IN 2013 MOORE TORNADO DATASET.

Classifier	Accuracy	F-measure	Precision	Recall
FT	**91.71%**	**0.916**	0.917	0.917
NBTree	90.00%	0.900	0.899	0.900
Random Forest	90.00%	0.899	0.899	0.900

Table IV
STEP 2: HIERARCHICAL CLASSIFICATION RESULTS (SPAM TWEETS VS. FAKE TWEETS) IN 2013 MOORE TORNADO DATASET.

Classifier	Accuracy	F-measure	Precision	Recall
FT	94.90	0.931	0.952	0.949

The two sets were stratified, and contained the same ratio of legitimate and non-legitimate tweets. Then we labeled spam and fake tweets in both training and testing sets to non-legitimate tweets. Hierarchical classification approach consists of 2 steps.

Step 1. First, we classify a tweet to either a legitimate or a non-legitimate tweet. By using this approach, we may achieve a higher accuracy in detecting non-legitimate tweets. In addition, sometimes we may want to filter non-legitimate tweets in practice. Table III shows experimental results of identifying legitimate and non-legitimate tweets. Again, FT classifier outperformed NBTree classifier and Random Forest classifier, achieving 91.71% accuracy and 0.916 F-measure. The hierarchical classification approach correctly identified 95.30% legitimate tweets and 84.48% non-legitimate tweets. It increased 4.16% recall of legitimate tweets compared with the flat classification approach.

We also performed 10-fold cross validation for the step 1. FT classifier consistently outperformed the other classifiers.

Step 2. In the second step, predicted non-legitimate tweets are further classified to spam and fake tweets.

To conduct this experiment, we first removed legitimate tweets from the training set and relabeled non-legitimate tweets to spam and fake tweets. The predicted non-legitimate tweets in the testing set was also relabeled to spam and fake tweets so that we can evaluate outcome of step 2. By using the training set only consisting of spam and fake tweets, we developed FT classifier and classified the predicted non-legitimate tweets to spam and fake tweets. The FT classifier achieved 94.9% accuracy and 0.931 F-measure as shown in Table IV.

The hierarchical classification approach was able to detect 83.33% fake tweets and 100% spam tweets correctly. In other words, the hierarchical classification approach significantly improved fake tweet detection rate over the flat classification approach.

B. Hurricane Sandy

The experimental results showed promising results. Now we apply the same approaches to Hurricane Sandy dataset to

Table V
FLAT CLASSIFICATION RESULTS IN HURRICANE SANDY DATASET.

Classifier	Accuracy	F-measure	Precision	Recall
FT	86.20%	0.855	0.861	0.862
NBTree	**86.68%**	**0.857**	0.863	0.867
Random Forest	85.63%	0.845	0.861	0.856

Table VI
STEP 1: HIERARCHICAL CLASSIFICATION RESULTS (LEGITIMATE TWEETS VS. NON-LEGITIMATE TWEETS) IN HURRICANE SANDY DATASET.

Classifier	Accuracy	F-measure	Precision	Recall
NBTree	86.04%	86.04	0.873	0.860

verify whether our proposed approaches consistently work in another dataset.

1) Flat Classification: Experimental results with 10-fold cross validation are shown in Table V. NBTree achieved 86.68% accuracy and 0.857 F-measure, outperforming FT and Random Forest classifiers. NBTree classifier correctly classified 97% legitimate tweets and 74.3% spam tweets, but only 30.43% fake tweets.

2) Hierarchical Classification: To improve fake tweet recall, we run hierarchical classification based on NBTree algorithm.

Step 1. In step 1, we classify tweets to legitimate and non-legitimate categories. Table VI shows the classification results. NBTree classifier achieved 86.04% accuracy and 86.04 F-measure. The classifier correctly classified 91.88% legitimate tweets and 71.79% non-legitimate tweets.

Step 2. We further classify predicted non-legitimate tweets to fake and spam categories. Table VII shows classification results. NBTree classifier achieved 96.43% accuracy by correctly identifying 62.50% fake and 100% spam tweets. Compared with the flat classification results, the hierarchical classification approach improved recall of fake tweets.

In summary, both flat and hierarchical classification approaches achieved up to 91.71% accuracy and 0.916% F-measure in 2013 Moore Tornado and Hurricane Sandy datasets. Especially, hierarchical classification approach identified fake tweets with higher accuracy than flat classification approach.

VI. DETECTING FAKE AND SPAM TWEETS BY USING STREAMING DATA

One missing study is when would be a right time to develop fake and spam tweet predictor while streaming data

Table VII
STEP 2: HIERARCHICAL CLASSIFICATION RESULTS (SPAM TWEETS VS. FAKE TWEETS) IN HURRICANE SANDY DATASET.

Classifier	Accuracy	F-measure	Precision	Recall
NBTree	96.43%	0.961	0.966	0.964

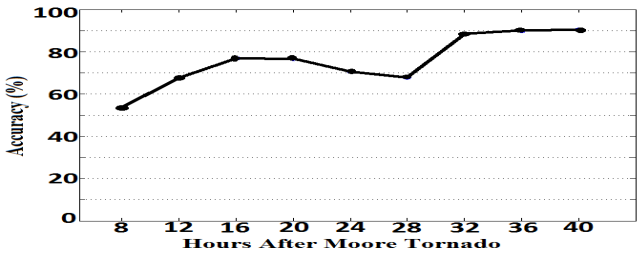

Figure 3. Hierarchical classification results of FT classifier with 4 hour interval in Moore Tornado dataset.

is coming? For example, when a natural disaster (e.g., Hurricane Sandy) happened, 1 hour after the disaster would be the best time to develop a predictor? To answer this research question, we conducted experiments with changing training time (e.g., 1 hour later or 4 hour later after the disaster happened). We conducted flat classification approach and step 1 in hierarchical approach for both 2013 Moore Tornado and Hurricane Sandy. Figure 3 shows the hierarchical approach with 4 hour interval in Moore Tornado dataset. It achieved the first peak in the first 16 hours, went down and then went up since the first 28 hours. Finally it reached to over 90% accuracy. Flat approach in Moore Tornado dataset, and flat and hierarchical approaches in Hurricane Sandy dataset achieved similar results with Figure 3. In summary, when we tested our classifiers in the first 20 hours in both datasets, the classifiers reached the first peak (achieving a reasonable accuracy). When we tested our classifiers in the first 36 hours in both datasets, the classifiers achieved higher accuracy.

VII. CONCLUSION

In this paper, we have conducted a case study of two natural disasters. First, we have collected tweets posted during the natural disasters on Twitter, and then analyzed distinguishing patterns among legitimate, spam and fake tweets. Then, we have proposed flat and hierarchical classification approaches with our proposed features. Finally, we have developed both flat and hierarchical classifiers for each dataset (disaster). Our experimental results show that detecting fake and spam tweets during natural disasters is possible with a high accuracy.

REFERENCES

[1] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in *WWW companion*, 2013.

[2] A. Gupta, H. Lamba, and P. Kumaraguru, "$1.00 per rt #bostonmarathon #prayforboston: Analyzing fake content on twitter," in *Eigth IEEE APWG eCrime Research Summit (eCRS)*, 2013.

[3] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: Social honeypots + machine learning," in *SIGIR*, 2010.

[4] M. Ott, C. Cardie, and J. Hancock, "Estimating the prevalence of deception in online review communities," in *WWW*, 2012.