5-2015

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

Meet Rajdev
Utah State University

FAKE AND SPAM MESSAGES: DETECTING MISINFORMATION DURING NATURAL DISASTERS ON SOCIAL MEDIA

by

Meet Rajdev

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Computer Science

Approved:

Dr. Kyumin Lee Major Professor	Dr. Amanda Lee Hughes Committee Member
Dr. Young-Woo Kwon Committee Member	Dr. Mark R. McLellan Vice President for Research and Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2015

ABSTRACT

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

by

Meet Rajdev, Master of Science

Utah State University, 2015

Major Professor: Dr. Kyumin Lee
Department: Computer Science

During natural disasters or crises, users on social media tend to easily believe contents of postings related to the events, and retweet the postings, hoping that the postings will be reached by 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 the 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.

(38 pages)

PUBLIC ABSTRACT

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

Meet Rajdev

During natural disasters or crises, users on social media tend to easily believe contents of postings related to the events, and retweet the postings, hoping that the postings will be reached by 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 the 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.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Billions of people have used social media sites such as Facebook and Twitter where they post messages related to various topics such as politics, economics, entertainment, sports and personal stories. These postings have been mined by researchers for stock market prediction [1], opinion mining [2], location prediction [3] and so on.

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 [4]. These messages were retweeted by many users who believed retweeting the messages would help the victims affected by the Hurricane Sandy.

Researchers [4, 5] analyzed fake contents or studied a fake image detection problem. Other researchers [6–8] 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. Do fake message posters and spammers have different behaviors from legitimate users? Do fake, spam and legitimate messages have distinguishing patterns? Can we automatically detect fake and spam messages?

To answer these questions, we make the following contributions in this paper:

- First, we collect tweets posted during 2013 Moore Tornado and Hurricane Sandy on Twitter. Then we analyze properties of fake, spam and legitimate messages.
- Second, we propose two classification approaches – (i) flat classification; and (ii) hierarchical classification – to automatically detect fake and spam messages. To our knowledge, this is the first research to detect both fake and spam messages at once.
- Third, we conduct fake and spam message detection experiments with the two classification approaches when Moore Tornado and Hurricane Sandy datasets are given.
- Finally, we measure consistency of the flat and hierarchical approaches when we combine two different natural disaster datasets.

CHAPTER 2

RELATED WORK

In this section, we summarize some of the previous research work related to spam and fake accounts or contents on social media.

Researchers have studied how to identify spammers or spam messages on social media for several years. For example, researchers [7,9–12] focused on analyzing behaviors of social spammers and detecting these spammers. Lee et al. [13] conducted a long-term study of content polluters, analyzed their behaviors and detected them. An online review spam detection problem has been studied [8, 14–16]. Ghosh et al. [6] analyzed how shortened URLs on Twitter have been used to link malware and spam links. McCord and Chuah [17] used machine learning approach to detect social spammers.

While the above researchers focused on individual spammer, other researchers focused on groups of spammers and their tactics. Mukherjee et al. [18] proposed a frequent pattern mining technique based approach to detect group review spammers. The Truthy system [19] detects astroturf political campaigns on Twitter. Gao et al. [20] studied spam campaigns on Facebook. Lee et al. [21] proposed a group spam detection approach based on near-duplicate detection methods and graph mining techniques.

Recently, researchers [22–25] have paid attention on rumor and fake information. Gupta et al. [4] discussed the identification of fake images on twitter which were spread during Hurricane Sandy. A scope of this work was limited to identifying fake images. They [5] also analyzed tweets posted on Twitter during the Boston Bombing, and found that 29% of the most viral content were rumors and fake content. Giatsoglou et al. [26] explored the identification of fraudulent and genuine retweet threads.

Compared with the previous research work, we collect two natural disaster datasets on Twitter, and analyze properties of fake, spam and legitimate messages. Then, we propose

flat and hierarchical classification approaches. To our knowledge, this is the first work to detect both fake and spam tweets at once on social media with even classifying between them. This research will complement the existing research work.

CHAPTER 3

DATASET

To conduct research for detecting misinformation on social media, collecting and labeling datasets are the first step. We now present our data collection and labeling strategy.

3.1 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. Then, we collected 1% sample tweets posted during a period of each event by using Twitter streaming API. In order to extract tweets relevant to an event, we pre-selected a set of keywords for each event. If a tweet contains one or more keywords relevant to an event, we considered it is relevant to the event.

Detailed data collection strategy is described as follows:

2013 Moore Tornado. The Moore Tornado struck Moore, Oklahoma on May 20, 2013. It killed 24 people, leaving behind extensive damage to homes and businesses over 5 days. 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. 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). It was one of the deadliest in the history of United States, and affected 24 states covering east and west sides. There were long power cuts in most of the affected areas. News media reported that 148 direct and 138 indirect fatalities were occurred. The total damaged amount was around 65 billion dollars. 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.

3.2 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.

Definition of the three categories – (i) spam; (ii) fake; and (iii) legitimate – is as follows:

- **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
  - 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. A tweet classified as spam or fake may be intentional or unintentional.

Based on the definition of the three categories, we show three example tweets relevant to "2013 Moore Tornado":

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

Table 3.1: Datasets.

- **Sample 1 Labeled as Legitimate:**

  "RT @5NEWS: An elementary school many have been hit by the #tornado, according to @KFOR `http://t.co/iyVKztNiUX` *"*

  This tweet is labeled as a legitimate message because it contains correct information, and no link to misleading/ spam content.

- **Sample 2 Labeled as Fake:**

  "@garyeOK I AIN'T SCARED OF NO STORM #oklahomaprobs #tornado `http://t.co/ivxz9CWoUd` *"*

  This tweet is labeled as fake because it contains a link "http://t.co/ivxz9CWoUd" which leads to a misleading image.

- **Sample 3 Labeled as Spam:**

  "#RealEstate #Tornado #OklahomaCity 15613 Ivy Hill Dr, Oklahoma City, OK 73170, $107,500 3 beds, 2 baths Find this RE & More: `http://t.co/LfAfoSBf6g` *"*

  This tweet is labeled as spam because the content focuses on personal agenda and monetary gain on the grounds of the event.

Examples of legitimate, fake and spam tweets.

With the above definition of the three categories, we labeled 2013 Moore Tornado and Hurricane Sandy datasets. 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. The ratio of Legitimate : Non-Legitimate (2:1) was maintained in dataset for both the events. 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 3.1 shows our datasets that are used in the rest of the paper.

CHAPTER 4

ANALYSIS

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

First, we analyze how many users, who posted different type of tweets, disclosed their location information in their profiles. Figure 4.1 shows what percent of users in each category disclosed their location information. While 66.6% legitimate users disclosed their location information, 38.1% fake message posters and 61.4% spammers disclosed their location information. Interestingly, fake message posters less likely shared their location information.

Since a tweet can only contain 140 characters, a tweet may not be enough for non-legitimate users (i.e., fake message posters and spammers) to post spam contents or fake information. Do non-legitimate tweets (i.e., fake and spam tweets) most likely contain URLs compared with legitimate tweets? To answer this question, we analyzed how many spam, fake and legitimate tweets in our dataset contain URLs. Figure 4.2 shows the percentage

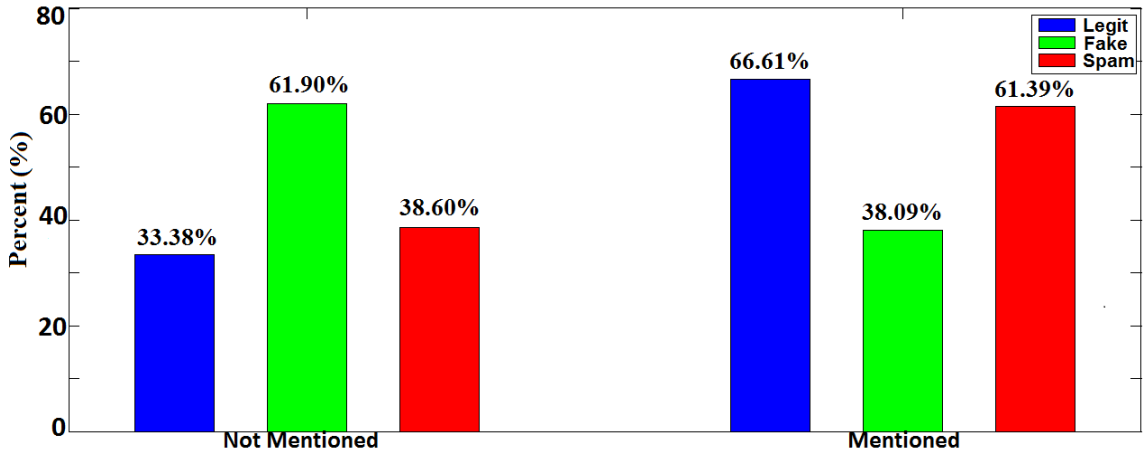

Figure 4.1: How many users have shared their location information?

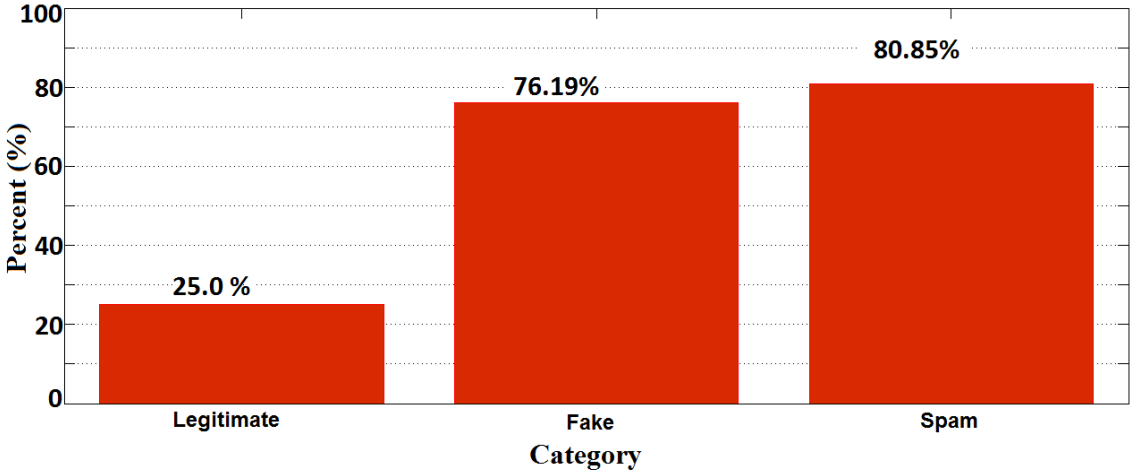

Figure 4.2: Tweets containing URLs.

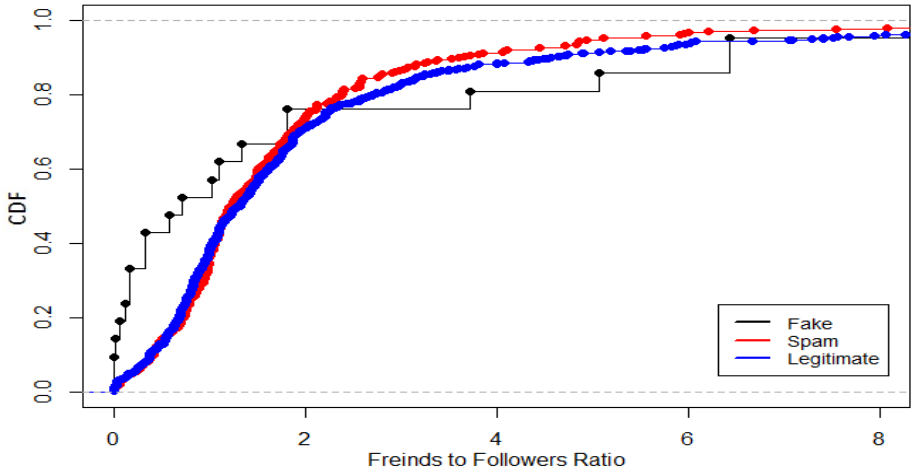

Figure 4.3: User-focused feature: Friends to followers ratio of users.

of tweets in each category which contain at least one URL. We can see that 80.85% spam and 76.19% fake tweets contained URLs while 25% legitimate tweets contained URLs. The result makes sense because non-legitimate users (e.g., spammers and malicious users) want to link malware pages or external pages containing misinformation toward tempting users to access the pages. We also noticed that many legitimate users posted short messages with hashtags to spread news, opinions and relevant information to friends and fellows.

Next, we analyze two user properties (i.e., a ratio of the number of friends and followers, and the number of favorited tweets) and one tweet property.

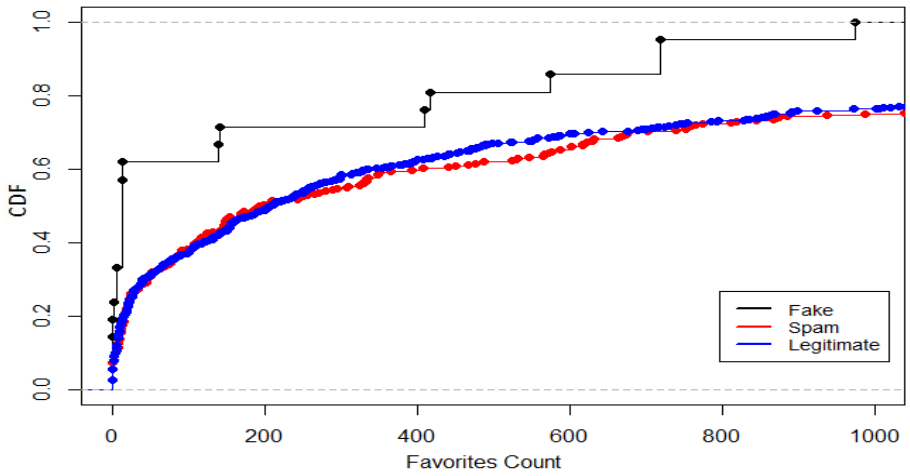

Figure 4.4: User-focused feature: Favorites count of users.

Figure 4.3 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 4.4 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 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:

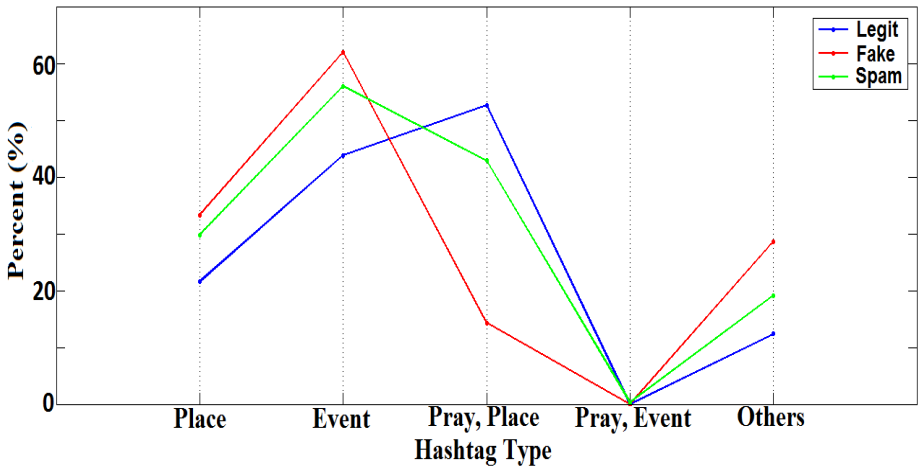

Figure 4.5: Tweet-focused feature: a hashtag type.

- **Place:**#Oklahoma, #Moore...
- **Event:**#Hurricane, #Tornado...
- **Pray, place:**#prayforoklahoma...
- **Pray, event:**#praytornadookc...
- **Others:**#help, #save...

Figure 4.5 shows a percentage of tweets containing a type of hashtags in each category. 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.

CHAPTER 5

PROPOSED APPROACH AND FEATURES

So far we have analyzed distinguishing patterns among legitimate, fake and spam tweets. We now turn to propose flat and hierarchical classification approaches toward detecting spam and fake tweets. Then, we present our features which are used to develop classifiers.

5.1 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 as shown in Figure 5.1. Unlike the flat classification approach, hierarchical classification approach consists of two steps as shown in Figure 5.2. 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.

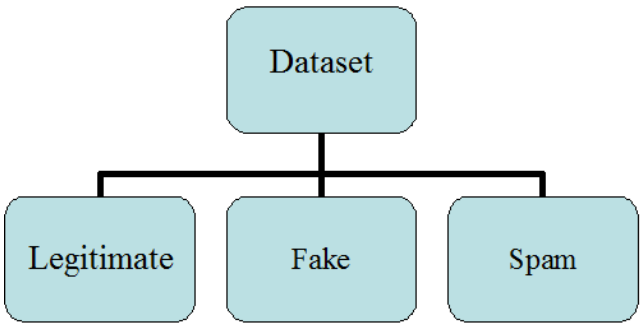

Figure 5.1: Flat classification approach.

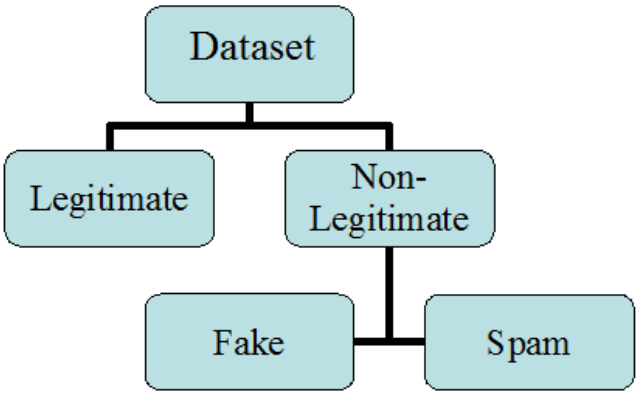

Figure 5.2: Hierarchical classification approach.

5.2 Features

To build classifiers, we now present our proposed features. Since our aim is to determine whether a tweet is spam, fake or legitimate, we extract features from a tweet and a user profile.

User features are as follows:

- The number of tweets the user has favorited in the user account's lifetime
- Ratio of the number of friends and followers
- The number of followers
- The number of friends
- Did the user enable the possibility of geotagging his Tweets?
- Length of the screen name of the user
- The number of public lists that the user is member of
- Did the user define his location in his profile?
- The number of tweets that the user has posted
- Time zone that the user declares himself within
- Does the user profile include a URL?
- Longevity of the user account (i.e., when was it created?)
- Does the user profile contain a user name?

- Is the user account verified by Twitter?

Tweet features are as follows:

- n-gram features
- A category of a hashtag (e.g., place, event, pray with a place, pray with an event name, and others) as we mentioned in the previous section
- Tweet creation time (UTC time)
- 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 2013 Moore Tornado and hurricane Sandy datasets, respectively. After feature selection, we finally used 675 and 343 features for 2013 Moore Tornado and hurricane Sandy, respectively.

CHAPTER 6

DETECTING SPAM AND FAKE TWEETS

Based on the proposed features, we now build flat and hierarchical classifiers to detect spam and fake tweets. We develop and test these classifiers in 2013 Moore Tornado and Hurricane Sandy datasets.

6.1 Experimental Settings

As we mentioned in dataset section, 2013 Moore Tornado dataset consists of 1,050 tweets and Hurricane Sandy dataset consists of 1,051 tweets. As evaluation metrics, we used accuracy, F-measure and confusion matrix. Table 6.1 shows a confusion matrix in which we can measure a recall value of a particular category. For example, a recall value of spam is z / (x + y + z).

To measure which classification algorithm would work well, we chose three classification algorithms such as Functional Trees (FT), NBTree (a decision tree with Naive Bayes classifiers at the leaves) and Random Forest.

6.2 2013 Moore Tornado

We first develop and evaluate classifiers to detect spam and fake tweets in 2013 Moore Tornado dataset. As we mentioned in the previous section, we apply flat and hierarchical classification approaches, and compare their classification performance.

Table 6.1: Confusion Matrix.

		Predicted		
		Legit	Fake	Spam
True	Legit	**r**	s	t
	Fake	u	**v**	w
	Spam	x	y	**z**

Table 6.2: 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

Table 6.3: Confusion matrix of flat classification based on Functional Trees in 2013 Moore Tornado dataset.

		Predicted		
		Legit	Fake	Spam
True	Legit	91.14%	0.57%	8.29%
	Fake	19.05%	57.14%	23.80%
	Spam	13.98%	0%	86.02%

6.2.1 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. In 10-fold cross-validation, the dataset was divided into 10 subsets. Then we used 9 subsets as training set and the remaining subset as testing set. We repeated 10 times by choosing each subset as testing set and remaining 9 subsets as training set.

We developed FT classifier, NBTree classifier and Random Forest classifier based on flat classification approach. Table 6.2 shows classification results. FT classifier outperformed NBTree classifier and Random Forest classifier by achieving 88.67% accuracy and 0.887 F-measure.

Next, we further analyze confusion matrix of FT classifier as shown in Table 6.3. We can observe 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.

The benefit of flat approach is that we can label the tweets one time, run the classifier and get the predictions with a good accuracy. We can detect spam and legitimate tweets with a high recall (90%), but we still want to achieve a better accuracy for detecting fake

Table 6.4: 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 6.5: Step 1: Confusion Matrix of hierarchical classification based on Functional Trees in 2013 Moore Tornado dataset.

		Predicted	
		Legitimate	Non-Legitimate
True	Legitimate	95.30%	4.70%
	Non-Legitimate	15.52%	84.48%

tweets. For this, we now turn to hierarchical classification approach.

6.2.2 Hierarchical Classification

To conduct hierarchical classification, we randomly split each dataset into training (containing 2/3 data) and testing sets (containing 1/3 data). 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 6.4 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.

Table 6.5 shows confusion matrix of the first step. 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.

Table 6.6: 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

Table 6.7: Step 2: Confusion Matrix of hierarchical classification based on Functional Trees in 2013 Moore Tornado dataset.

		Predicted	
		Fake	Spam
True	Fake	83.33%	16.67%
	Spam	0.00%	100.00%

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 6.6.

Table 6.7 shows confusion matrix of the classification results. 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.

6.3 Hurricane Sandy

In the previous subsection, we performed flat and hierarchical classification. The experimental results showed promising results. Now we apply the same approaches to Hurricane

Table 6.8: 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 6.9: Confusion matrix of flat classification based on NBTree in Hurricane Sandy dataset

		Predicted		
		Legit	Fake	Spam
True	Legit	97.00%	1.71%	1.28%
	Fake	68.11%	30.43%	1.45%
	Spam	23.49%	1.78%	74.73%

Sandy dataset to verify whether our proposed approaches consistently work in another dataset. We follow the same experimental scenarios of 2013 Moore Tornado by conducting flat and hierarchical classifications.

6.3.1 Flat Classification

As mentioned previously, flat classification approach classifies tweets into legitimate, fake and spam categories. Experimental results with 10-fold cross validation are shown in Table 6.8. NBTree achieved 86.68% accuracy and 0.857 F-measure, outperforming FT and Random Forest classifiers. Table 6.9 shows confusion matrix of NBTree classifier. 97% legitimate tweets, 74.3% spam tweets were correctly classified. But, only 30.43% fake tweets were correctly classified.

6.3.2 Hierarchical Classification

To improve fake tweet recall, we run hierarchical classification based on NBTree algorithm which performed the best in the previous experiment.

Step 1. In step 1, we classify tweets to legitimate and non-legitimate categories. Table 6.10 shows the classification results. NBTree classifier achieved 86.04% accuracy and

Table 6.10: 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

Table 6.11: Step 1: Confusion Matrix of hierarchical classification based on NBTree in Hurricane Sandy dataset.

		Predicted	
		Legitimate	Non-Legitimate
True	Legitimate	91.88%	8.12%
	Non-Legitimate	28.21%	71.79%

86.04 F-measure. Table 6.11 shows confusion matrix of the classifier which correctly classified 91.88% legitimate tweets and 71.79% non-legitimate tweets.

Step 2. Now, we further classify predicted non-legitimate tweets to fake and spam categories. The process of generating training and testing sets was the same with experimental settings of 2013 Moore Tornado. Tables 6.12 and 6.13 show classification results. NBTree classifier achieved 96.43% accuracy by correctly identifying 62.50% fake tweets and 100% spam tweets. Compared with the flat classification results, the hierarchical classification approach improved recall of fake tweets.

So far, we applied flat and hierarchical classification approaches to 2013 Moore Tornado and Hurricane Sandy datasets. Both flat and hierarchical classification approaches achieved up to 91.71% accuracy and 0.916% F-measure. Especially, hierarchical classification approach identified fake tweets with higher accuracy than flat classification approach.

6.4 Combining Two Datasets

What if we combine Moore Tornado and Hurricane Sandy datasets, and develop and evaluate classifiers? Do they still achieve a high accuracy even though two different events

Table 6.12: 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

Table 6.13: Step 2: Confusion Matrix of hierarchical classification based on NBTree in Hurricane Sandy dataset.

		Predicted	
		Fake	Spam
True	Fake	62.50%	37.50%
	Spam	0.00%	100.00%

Table 6.14: Flat classification results in the combined dataset.

Classifier	Accuracy	F-measure	Precision	Recall
FT	**88.30%**	**0.878**	0.850	0.883
NBTree	85.02%	0.841	0.847	0.850
Random Forest	87.02%	0.863	0.877	0.870

happened in different timing? To answer these research questions, first we combined Moore Tornado and Hurricane Sandy datasets. Then, we randomly split the combined dataset to training (containing 2/3 dataset) and testing (containing 1/3 dataset) sets. The two sets were stratified.

Flat Classification. We developed flat classifiers and then classified tweets to spam, fake and legitimate categories. Table 6.14 shows experimental results. FT classifier outperformed NBTree and Random Forest classifiers, achieving 88.3% accuracy and 0.878 F-measure.

Hierarchical Classification. Next, we conducted hierarchical classification. In step 1 (again, classifying tweets to legitimate or non-legitimate category) as shown in Table 6.15, FT classifier achieved 86.59% accuracy and 0.864 F-measure. In step 2 (again, classifying predicted non-legitimate tweets to spam or fake category), FT classifier achieved 94.92% accuracy and 0.939 F-measure.

Overall, both flat classification and hierarchical classification approaches performed well in the combined dataset even though Moore Tornado and Hurricane Sandy happened

Table 6.15: Hierarchical classification results (legitimate tweets vs. non-legitimate tweets) in the combined dataset.

Classifier	Accuracy	F-measure	Precision	Recall
FT in Step 1	**86.59%**	**0.864**	0.864	0.866
FT in Step 2	**94.92%**	**0.939**	0.952	0.949

in different time and are different events.

CHAPTER 7

DETECTING FAKE AND SPAM TWEETS BY USING STREAMING DATA

So far, we have developed and evaluated flat and hierarchical classifiers in Moore Tornado and Hurricane Sandy datasets. One missing study is when would be a right time to develop fake and spam tweet predictor while streaming data 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? In other words, train a classifier based on tweets posted within 1 hour after the disaster just happened and test up-coming tweets. 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). In this study, we only show experimental results of flat classification approach and step 1 in hierarchical approach.

Moore Tornado. We chose 4 hours as an experimental interval. In other words, we built a classifier and tested once every 4 hours. Figures 7.1 and 7.2 show experimental results of flat classification approach and hierarchical approach, respectively. In the first 4 hours, there was no fake and spam tweets, so we developed a classifier since 8 hours after Moore Tornado happened. In the first 8 hours, FT classifier based on the flat approach achieved 54% accuracy. As we observed longer (i.e., increase the size of a training set), accuracy was increased until 12 hours, then went down and went up, and then we reached to stable accuracy since 32 hours. Hierarchical approach 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.

Hurricane Sandy. For Hurricane Sandy dataset, we used the same 4 hour experimental interval. There was no spam and fake tweets until the first 8 hours, so we only report classification accuracy since the first 12 hours. Figures 7.3 and 7.4 show experimental

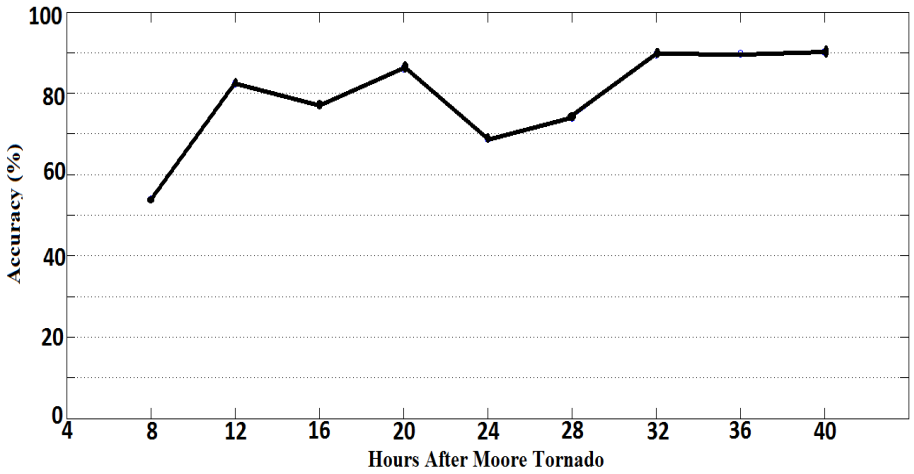

Figure 7.1: Flat classification results of FT classifier with 4 hour interval in Moore Tornado dataset.

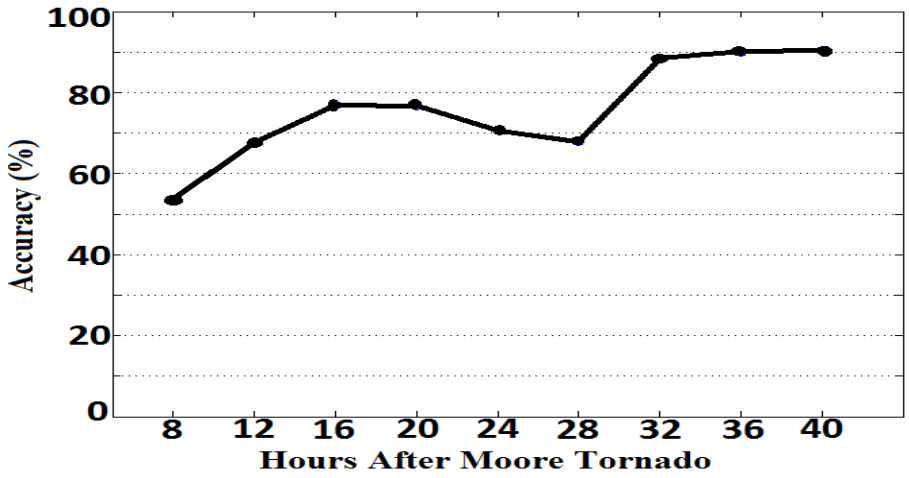

Figure 7.2: Hierarchical classification results of FT classifier with 4 hour interval in Moore Tornado dataset.

results in Hurricane Sandy Dataset. Both flat and hierarchical classification approaches show a similar pattern in which accuracy went up and down and then went up since the first 28 hours.

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.

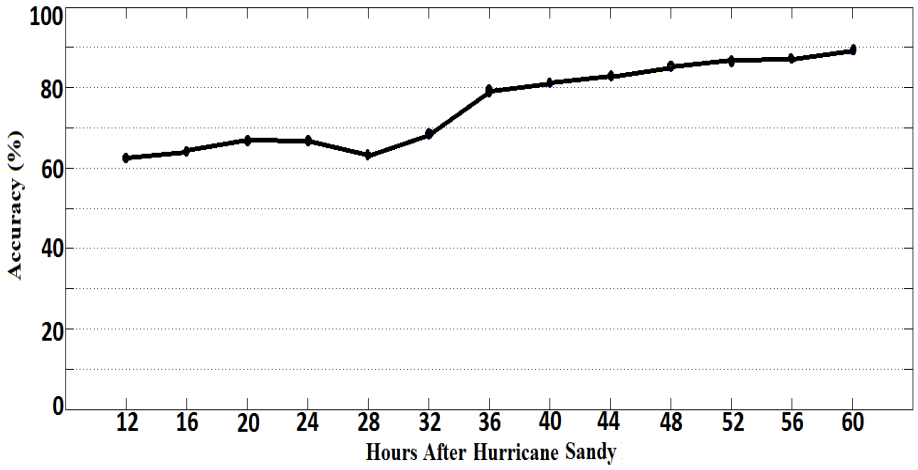

Figure 7.3: Flat classification results of FT classifier with 4 hour interval in Hurricane Sandy dataset.

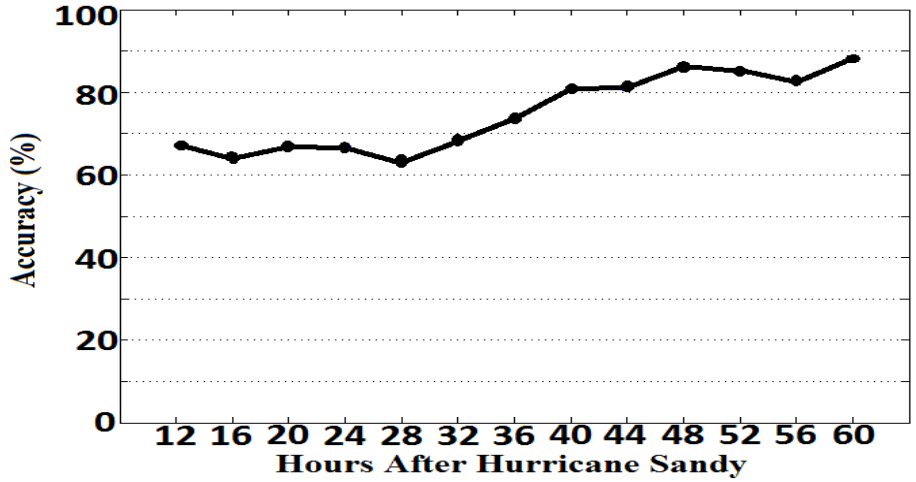

Figure 7.4: Hierarchical classification results of FT classifier with 4 hour interval in Hurricane Sandy dataset.

CHAPTER 8

CONCLUSIONS

In this paper, we have conducted a case study of two natural disasters – 2013 Moore Tornado and Hurricane Sandy. 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) and the combined dataset.

In 2013 Moore Tornado dataset, a flat classifier based on FT algorithm achieved 88.67% accuracy, while a hierarchical classifier achieved 91.71% accuracy in step 1 and 94.9% accuracy and in step 2. In Hurricane Sandy dataset, a flat classifier based on NBTree algorithm achieved 86.2% accuracy while a hierarchical classifier achieved 86.04% accuracy in step 1 and 96.43% accuracy in step 2. In the combined dataset, our classifiers consistently worked well by achieving up to 94.92% accuracy. Our experimental results show that detecting fake and spam tweets during natural disasters is possible with a high accuracy.

CHAPTER 9

FUTURE WORK

Future work may include the implementation of the model in real time, and assessing the performance. Accordingly, there can be changes made in the definitions of fake and spam content, thus resulting in better recall values for the categories.

Following our procedures, steps can be taken to reduce the amount of non-legitimate data online and block the users who spread misinformation in the network.

REFERENCES

[1] G. Wang, T. Wang, B. Wang, D. Sambasivan, Z. Zhang, H. Zheng, and B. Y. Zhao, "Crowds on wall street: Extracting value from collaborative investing platforms," in *CSCW*, 2015.

[2] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*. Springer, 2012, pp. 415–463.

[3] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: A content-based approach to geo-locating twitter users," in *CIKM*, 2010.

[4] 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.

[5] 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.

[6] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@ spam: the underground on 140 characters or less," in *CCS*, 2010.

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

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

[9] S. Ghosh, G. Korlam, and N. Ganguly, "Spammers' networks within online social networks: a case-study on twitter," in *WWW companion*, 2011.

[10] M. S. P. G. Kelley and J. Cranshaw, "Don't follow me: Spam detection in twitter," in *SECRYPT*, 2010.

[11] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *ACSAC*, 2010.

[12] S. Yardi, D. Romero, G. Schoenebeck *et al.*, "Detecting spam in a twitter network," *First Monday*, vol. 15, no. 1, 2009.

[13] K. Lee, B. D. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on twitter." in *ICWSM*, 2011.

[14] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee, "How opinions are received by online communities: A case study on Amazon. com helpfulness votes," in *WWW*, 2009.

[15] S. Feng, L. Xing, A. Gogar, and Y. Choi, "Distributional footprints of deceptive product reviews," in *ICWSM*, 2012.

[16] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via temporal pattern discovery," in *KDD*, 2012.

[17] M. McCord and M. Chuah, "Spam detection on twitter using traditional classifiers," in *Autonomic and trusted computing*. Springer, 2011, pp. 175–186.

[18] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal, "Detecting group review spam," in *WWW*, 2011, pp. 93–94.

[19] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media," in *ICWSM*, 2011.

[20] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *IMC*, 2010.

[21] K. Lee, J. Caverlee, Z. Cheng, and D. Z. Sui, "Campaign extraction from social media," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 1, pp. 9:1–9:28, Jan. 2014.

[22] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *WWW*, 2011.

[23] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "Tweetcred: A real-time web-based system for assessing credibility of content on twitter," *arXiv preprint arXiv:1405.5490*, 2014.

[24] X. Xia, X. Yang, C. Wu, S. Li, and L. Bao, "Information credibility on twitter in emergency situation," in *PAISI*, 2012.

[25] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," in *SIGKDD Workshop on Mining Data Semantics*, 2012.

[26] M. Giatsoglou, D. Chatzakou, N. Shah, C. Faloutsos, and A. Vakali, "Retweeting activity on twitter: Signs of deception," in *PAKDD*, 2015.