Development of a Rumor and Spam Reporting and Removal Tool for Social Media

Shakil Ahmed
Department of Electrical and Computer Engineering
North South University
Dhaka,Bangladesh
shakil.ahmed02@northsouth.edu

Rifat Monzur
Department of Electrical and Computer Engineering
North South University
Dhaka,Bangladesh
rifat.monzur@northsouth.edu

Rajesh Palit
Department of Electrical and Computer Engineering
North South University
Dhaka,Bangladesh
rajesh.palit@northsouth.edu

***Abstract*—Spamming persists as a problem in Social Networking Sites (SNS). About 8.7% of the total accounts of Facebook are fake or duplicate. These fake accounts are spreading rumors. Facebook believes 14.3 million of these accounts were created for spamming. Sometimes these accounts are being used to fulfil personal vendetta. Using spam filters seems to be insufficient as Facebook is struggling to clean up the spam while users are falling in traps like clickbait. On the other hand rumors in SNS causing severe trouble on a large scale. In recent time a violent incident took place in Ramu, Bangladesh on the basis of a rumor spread through Facebook. Rumors like this affecting us as individuals, our society. The main challenge of a rumor detection tool is the accuracy of the system and the time it takes to give the results. An accuracy having 90% accuracy may not be enough in the time of natural disasters or war as remaining 10% can cause havoc. Again if the system takes too much time no matter what the result is, it may be too late. We designed a tool which can quickly detect if a post is a spam or rumor. This tool responds to a user query for a post and try to verify the authenticity of a post with the help of Facebook community. This tool is designed to give a quick verdict so that it can be useful to prevent a viral malcontent in Social Networking sites. We implemented this tool on Facebook. We used Facebook Graph API to select groups of users with relevant knowledge about the topic who can contribute to determine the authenticity of the post.**

***Index Terms*—Social Network; Spam; Rumor; Security; Tool; Privacy.**

I. Introduction

Through Internet we are more connected than ever. From 2005 to 2016, the number of Internet users increased from 16% of the world population to 40% [4]. Much of the credit goes to booming market of smartphone. Smartphones made Internet more accessible to users from all over the world. In developing countries the number of mobile Internet users doubling every year [5].Among Internet users, 2.34 billion people are social networking site users.

People using sites like Facebook, Twitter, Myspace, LinkedIn for connecting with others for social reasons. The impact of social networking sites has become tremendous in our real life. But it has also brought some problems like spamming, spreading rumor, cyberbullying, etc. While problems like spamming affect mostly on personal level, viral rumor on social media affect broader population. Some of the rumors spread in time of natural disaster or terrorist incidents like Boston bombing created confusion and chaos on social media while some caused violent incidents [6].

Spam can be manifested in many ways, including bulk messages, profanity, insults, hate speech, malicious links, fraudulent reviews, fake friends, and personally identifiable information. Experts estimate that as many as 40% of social network accounts are used for spam. In 2008, 83% users in Facebook received junk messages or unwanted friend request. Spammers, however, frequently change their address from one throw-away account to another, and are thus hard to track. These spams are usually clickbait which completely depends on user interaction with that spam. On the other hand detecting rumor remains a daunting task as rumors cannot be filtered like email spams. Since there is no existing system in social media which can verify authenticity of a post, stopping rumors from spreading is more difficult.

Social networking sites have been using email filter like systems to remove spams from their sites. Problem with this method is communication methods in social network sites are very different than emails. Currently these sites have reporting system for a user to report or hide a post. But a post can have multiple source and in social network as these posts tends to be shared, it becomes harder to remove them from sites. Moreover it takes too much time to remove post containing rumor. Again different methods are being used to identify fake accounts using criteria such as unsuccessful friend request sent or reports from others users. But these methods are preventive at some extent and cannot tackle with viral rumors. Some external apps are also available to users to block unwanted advertisement or adult content in form of extension. But these software are user focused. Some websites rely on advertisements to generate revenue but are losing money because of ad blockers as much as 20.3 billion USD in 2016 [7].

We think that without the help of social network community it is impossible to prevent spams and rumor. We think that at the current world of overflowing information it is impossible to prevent rumor from being created but it is possible that rumor can be detected at early stage and can be stopped from propagating further. We propose to involve the social network user base in detecting spam and rumor more actively. This

978-1-5090-5753-5/16 $31.00 © 2016 IEEE
DOI 10.1109/APWC.on.CSE.2016.33

method aims to use the first hand knowledge of the users who are more close to the incident that is trending or have relevant knowledge about the situation. Only giving correct information (information as knowledge) can reduce confusion [8] and thus prevent rumor. Our goal is to detect the rumor very early by finding the pattern of the asked questions of users. Second benefit of our proposed system is users question is private so the question itself cannot act as rumor spreading agent. Because rumor thrives in uncertainty and repetition of same question in public increases anxiety in public [9] which is a catalyst for rumor [10].Third benefit of our system is that once a community thinks a post a spam or rumor they can work together to report the post from the site which can effectively remove the post quickly.

Spam and rumors have great consequence on personal level, community level and society as a whole. Some problems include unwanted posts in their timeline, important posts are missed due to glut of spam posts, unwanted pop-up windows and loss of internet bandwidth. Some spam steals personal information which acts as click bait. Rumors creates misinformation and spread propaganda and try to manipulate social mob. In recent times some violent incident occurred in Bangladesh due to spreading of political rumors. Rumors have greater impacts than spam because it affects greater number of people and can induce people in such extent that simple conversation in social websites can turn into violent acts. People also get influenced by various terrorist propaganda like ISIS and get brainwashed and follow them in real life.

In this paper first we propose our solution to the spam and rumor problems in social network. Then we discuss the design of the spam and rumor removal tool in details. We discuss how it is going to affect social network users and how it can be used for wellbeing of the society. Then we discuss other notable related works. The section after that concludes the paper.

II. Related Works

Rumor in social network has been a problem for quite some time and there have been progressive work addressing that problem. At early stages the reason behind the spreading rumor and in what situation it is more likely to be propagated were discussed [11].Some work has been made about the properties of rumors in social media [12]. This paper discuss how fast rumors spread in social network from one node to another and travel from one cluster of network to another. Some paper discusses credibility of social microblogs and how to increase the credibility of posts [13].

After the explosion of social media new problem like spam and rumors began to emerge. Spam and rumors were very easy to spread compared to email. Social network had also brought new dynamics to the propagation of rumors. Questions were asked such as how social spam is different than email spam [14, 15].Gao et al. [16] presented an initial study to quantify and characterize spam campaigns launched using accounts on online social networks. They found that more than 70% of all malicious wall posts advertise phishing sites. They also studied the characteristics of malicious accounts, and saw that more than 97% are compromised accounts, rather than Fake accounts created solely for the purpose of spamming.

There have been some work focused in special situation like natural disaster. Oh et al. [17] explicated the conditions needed to enhance information quality of Twitter in the extreme event context. They analyzed the twitter data on Haiti earthquake using two key variable of rumor theory, anxiety and information uncertainty. They divided their database into four stages in chronological order. They showed in first two stages emotion were high than later stages and correlated emotion or anxiety with the number of rumor spread during that time period. They also showed that rising of sufficient authentic posts at first two stages helped suppress the reduction of rumor in later stages which is consistent with the rumor theory.

Other notable work focused on detecting rumor at its early stages based on the contents of posts. Zhao et al. [18] argues that there will be always some user who will question about the post in the comment. They described this inquire behavior as sensors. They identified a set of regular expression that define set of tweets. Then they compared clusters of signal data with the cluster that are independent of any particular topic to rank them by the likelihood of containing disputed factual claim.

In recent years some notable researches have been done to detect rumor automatically by developing classifiers [19, 20, 21]. In these papers authors trained classifier based on specific events. Yang et al. [22] added two new features (event location and client-program) along with old features (content-based features, account-based features and propagation-based features).They trained two classifiers. One without new features and another with new features. They showed that the accuracy of rumor detection increased by 5.43%, 4.73%, and 6.32%.

While there has been many progressive works like detecting rumors in extreme events [17] or verifying posts by experts these works can detect rumor with satisfactory accuracy but these methods takes too much time and thus rumors can do so much damage if not stopped early. Some interesting study has also been conduct like identifying the source of rumors [23] which may help us how propaganda are spread through the social network. Some study is also made for the feasibility study of conducting relief work in disaster aftermath tackling with rumors so that relief works can be conducted swiftly depending on the social networking sites [24].Some other works are done to predict credibility of the tweets [25].It follows a supervised learning approach for the task of automatic classification of credible news events. A first classifier decides if an information cascade corresponds to a newsworthy event. Then a second classifier decides if this cascade can be considered credible or not. The paper undertakes this effort training over a significant amount of labeled data, obtained using crowdsourcing tools. Then validates these classifiers under two settings: the first, a sample of automatically detected Twitter trends in English, and second, tests how well this model transfers to twitter topics in Spanish, automatically detected during a natural disaster.

III. Proposed Solution

Rumor is an important form of social communications, and spread of rumors plays a significant role in a variety of human affairs. Particularly, we can view rumor spread as a stochastic process in social networks. The rumor is propagated through the population by pairwise contacts between spreaders and others in the population. Any spreader involved in a pairwise meeting attempts to infect the other individual with the rumor. In the case this other individual is an ignorant, they become a spreader. In the other two cases, either one or both of those involved in the meeting learn that the rumor is known and decided not to tell the rumor anymore, thereby turning into stifles.

Social networking sites are becoming more and more popular day by day. Users are getting more and more involved in the concept of social networking. In recent times, some rumors were so misleading that some create huge chaos and violence. These rumors are usually spread by individual profiles or a page or an external links. These rumors gets importance because of herd behavior psychology of users. When a rumor becomes established it becomes very hard to discredit it. Most of the time rumor spread through the social network has a short lifetime. But often these short lived rumors can easily do severe damage to a person or a society as a whole.

Spam and rumors have great consequence on personal level, community level and society as a whole. Some problems include unwanted posts in their timeline, important posts are missed due to glut of spam posts, unwanted pop-up windows and loss of internet bandwidth. Some spam steals personal information which acts as click bait. Rumors creates misinformation and spread propaganda and try to manipulate social mob. In recent times some violent incident occurred in Bangladesh due to spreading of political rumors. Rumors can indenture people in such extent that simple conversation in social websites can turn into violent acts. People also get influenced by various terrorist propaganda like ISIS and get brainwashed and follow them in real life.

We aim to develop a social rumor removal tool which can help user from misleading information but also does not block information access in internet. Our vision is make people a part of the rumor removal process so that it cannot be used as a tool of bullying in social media. With a view of such philosophy we tried to develop a crowd sourced rumor detection and removal model which is fast to detect rumors within a very short period of time. The second part is the removal of the rumor contents by the participation of the people. The moderators who conclude if a post is spam or rumor are consist of general people and experts. Different weights are given based on the expertise of a person and also previous judgment. A person gets positive point if their verdict turns out to be correct later on and negative points if the verdict turns out to be incorrect.

Web App: In Facebook based reporting system users have to become group member of spam and rumor reporting community. Moderators will have to be members of both the public group to communicate with users and a secret group to communicate with themselves. Moderators will have to install Facebook app to review the reported posts and engage in voting procedure. Facebook user can easily use this feature. User can use this app from all platforms and devices. The only thing they need to do is share the post in our group that they want to report. The moderators also can use its full features on any platforms and any devices. But they will have to depend on desktop in case of direct notification send to them from Facebook app as Facebook app notification is allowed only for desktops. Chrome Extension: Once a user installs extension they can enjoy its full feature like removing a post from his feed forever, reporting a post as spam to our group. When user report a post on Facebook he/she will never see that post again on their browser regardless of the Facebooks decision to keep or remove that post. Extension is browser dependent. So people who use mobile phone and only want to use mobile apps will not be able to use this feature from the app but they can always use this extension using browser from their phone. This system depends on the crowd response on a situation. Therefore developing a user friendly system is a priority. A post is chosen as spam if it received votes as spam given that at least 70% moderators voted.

IV. Design and Implementation

The gist of the system can be obtained by looking through the life cycle of a post. From the point it is regarded as possible spam or rumor by the users to the stage it is judged by the moderator. Out main focus of the system was to identify rumor at its early stages so we tried to shorten the steps Fig. 1 a post taken to its judgment. From Fig. 1 we can see that a post upon investigation it goes through a modicum steps to be verified.

In next subsection we discuss the system from point of view of a user. Then we talk about the moderator and their selection for specific topics. After that we have a detailed discussion about two implementations of the design, chrome extension and Facebook App. We conclude our system design discussion by observing a query which gives an insight of the quality of the implementations.

A. Total System Overview

User logging into Facebook from desktop can use this app in two ways. One, user can use Facebook app. Using Facebook app is very easy but it provides less option for the user. User has the option to query for a supposed rumor as they see fit. Upon querying they just have to wait for the result. If the moderators were able to reach a decision, users are notified.

Two, user can use chrome extension. Chrome extension have extra features like an integrated notify button, identified post and its shared copies removal, etc. We can see from Fig. 2 that Chrome extension gave us more freedom to work with as it had less restriction than working on Facebook app. Also, we did not have to depend on the graph API to interact with the user in case of extension. Which enabled us to provide more flexible option for the users. But after user gives a response the query is stored to database and then the process of identifying rumor

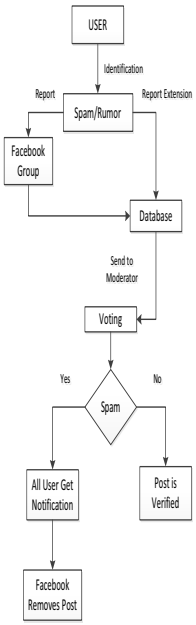

Fig. 1. Diagram of post life cycle

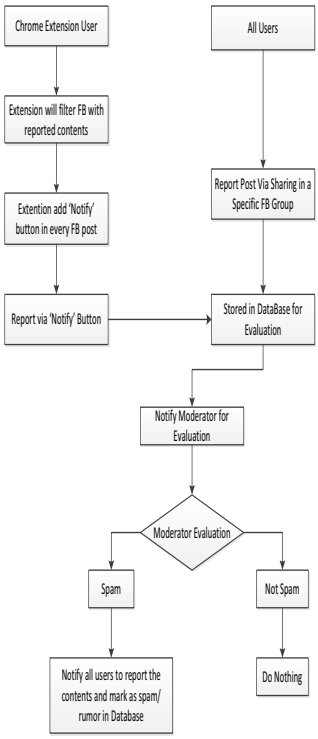

Fig. 2. Diagram of Total System

is same. User logging from mobile can only use Facebook app option.

B. Moderator Selection

In social networks sites different rumors emerges every day. These rumors are different in terms of origin, type, celerity, etc. It is obvious that different moderators will have different skill set and knowledge to determine the authenticity of a rumor. Keeping in mind that its voluntary work for moderators we provided a notification system with Facebook app as shown in Fig. 3 so that they can easily navigate through the voting process.

Bearing that in mind we created moderator profile based on their origin, academic background, expertise, knowledge, experience, response time and availability. To make sure all votes carries almost equal weight we pull moderators from our moderator pool based on their profile.

C. Chrome Extension

Extension was designed so that user can get rid of a post and shared version of that post. Because current hiding option of Facebook is not enough because it spoil the purpose if user still has to see almost same kind of post. Fig. 4 demonstrate that when a user logs into account using extension, it search for the

Fig. 3. Notification to review a post

contents that user reported as rumor or spam. It removes them for home page. When the page loads it also adds a notification button with every post so that when user wants to remove a post from their news feed, they can do it very easily by one click.

To delete a post from the news feed we used four features. The content of the post, the source, related trending topic and number of times post was reported by other users. The main challenge was to identify meme posts, images or videos without any description. Because very important feature content could not be used in that matter.

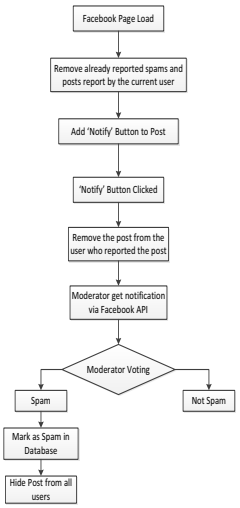

Fig. 4. Diagram of Chrome Extension Functions

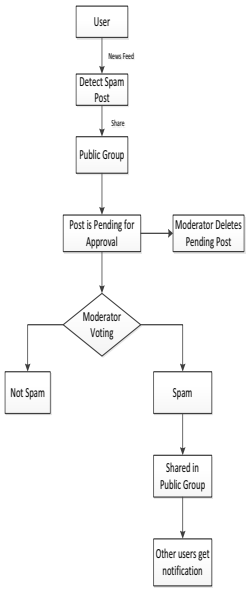

Fig. 5. Diagram of Web using FB graph API

D. Facebook App

Facebook app was designed keeping in mind that Facebook tends to change graph API within a very short period of time. So we tried to use as less API as possible. It is very difficult to integrate additional tool on Facebook because of so many restriction keeping in mind that tool must be user friendly. That is why we wanted to use Facebook share button for a report as this is very familiar to the user and can be easily done. When sharing the post we request users to provide original link as attachment as it is very helpful to determine rumor if the source is known.

Fig. 5 describes the different Facebook App stages. We created a Facebook group so that user could share the post which they think of spam or rumor with the group. After sharing the post, the post is not shared in public with rest of the member as it would create more rumors. So the post is sent to approved section where an admin sees the post and takes a decision whether to send this post to database or delete (if similar kind of post already had reached before admin can give a quick verdict) it. We could not automate this part because Facebook API do not grant access to the approved section of a group.

E. Inquiry life cycle

After the user inquiry, the necessary information are stored in database using Facebook API as shown in Fig. 6 Then the systems determines which moderators to send this query. Moderators with same origin or place of incident, expertise on the topic, experience on the topic and availability at the time gets priority. After selecting the required number of moderators, query is sent for verification. For each moderator takes a decision as in Fig. 7, database updates.

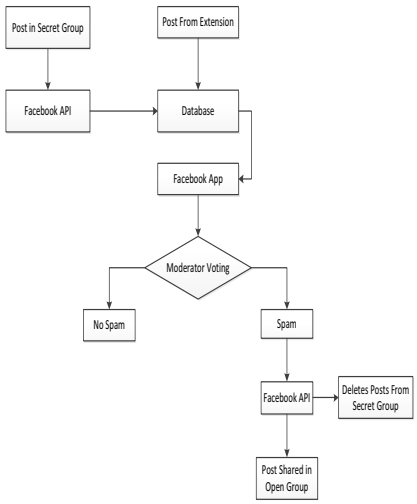

Fig. 6. Diagram of inquiry life cycle

Fig. 7. Moderator's voting window

The final decision awaits for the necessary percentage of the moderators to take a decision. When it happens decision is taken to identify the post as a spam or not spam and the result is sent to user. The result is also sent to other users who had subscribed the option to get notified for rumor posts. The decision is also posted in the public group using Facebook API so that the all group members know the correct information. That also helps to reduce the recurrence of same inquiry.

V. Impact

In social networks, we often see misleading information which creates confusion among users. Users do not know what to believe and what not to believe. Misleading information affect user thinking process and guide people for harm. With the help of our project people can know which information is true and which is not. Which will reduce the confusion among the users and create more certainty among the users. Since our project also contain the process of hiding misleading information, it will create a trust within the user about what information user witness in social networks.

The social networks have a great influence in our life now a day. The influence is so great that it is affecting our thinking process. The influence has both bad and good effect in our social life. In recent years, we have even seen misleading information created chaos and violence. To avoid such rumors, our system could come in handy. Users could use our system as media to detect the validity of any information and avoid any misleading.

In recent years, there has been many terrorist activities around the world. Terrorist use social media to make threat and spread activity of their terrorism. In many cases they recruit through social networks. ISIS, also known as ISIL recruit through SNS. Innocent young people are being captured and misguided via social media. They are using false propaganda to hire their recruit. Many online news portal and various Facebook group sharing their activities. Both adult and children get to know about their barbaric activity. However, with the help our project, we could remove those inappropriate posts. Our filtering system could keep a vital role in the fight against terrorism. Our reporting system can detect spam and report to Facebook more quicker than traditional way.

VI. Limitations

We divided the whole system into two main parts. One is reporting of inappropriate contents such as spam and rumors. Other is to removal of the reporting contents. So, the first step was to build a reporting system. The first challenge was to develop a reporting system that will work in all the versions of Facebook. We build a web app which will interact with users and moderators via Facebook graph API used both graph API of Facebook and current available notification system of groups to make the system. After all, we wanted to develop a web application using Facebook Graph API which will be accessible to both desktop and mobile app users.

The rumor part of the application is fully crowdsourced. We rely heavily on the opinion of the moderators when it comes to report the posts to Facebook. If the majority of the moderators does not think that post as a rumor or enough number of users does not participate in reporting process then the post may not be removed as a result. Again flagging a post as a rumor (getting a post as a rumor and verify request send to us) is totally dependent on the moderator participation. We can only verify a post when it is send to us.

Success of this app depends on users and skilled moderators. We build an automated system where moderator can work properly. But still without proper moderators system will remain flawed. The moderators need to be skilled and fair. One moderator fault can be overcome in our system by multiple moderator voting system. However, if a big portion of the moderator panel make wrong decision which is very unlikely, we cannot do anything about it. There is still considerable room to improve the effectiveness of the rumor detection method. We can introduce machine learning along with moderator voting. This can make the tool more accurate. Other improvement can be done on selecting moderator. We can add new features to pull moderators from moderator pool like previous posts, previous voting accuracy.

VII. Conclusion

Social Networking Sites like Facebook use email spam filter to identify spams like clickbait [2], but are facing serious problem to tackle rumor. In the time of natural calamity like earthquake in Haiti or terrorist attack in Boston or Mumbai, we saw a lot of rumors going viral in social media. Some

of these rumor were harmless but some of the rumors had malignant affect like hindering relief process, and some of the rumors were creating anxiety and distress among people of the affected area. The goal of this project was to build a system to prevent spams and harmful rumors from spreading in social media. Since Facebook is the most popular social network site, we targeted Facebook as our initial platform to build the system. So, the scope of this project was strictly within Facebook. Our project was to build filter system for the reported contents. There is no Facebook API to stop same post (shared posts too) from appearing in Facebook of a specific user. For that we used browser based solution. Since Google chrome is the most popular browser, we build the system for chrome. We made a browser extension for chrome which work as a spam and rumor removal tool. And, for the ease of user experience, we also added a button so that user can report the content with just one click.

REFERENCES

[1] 83 million Facebook accounts are fakes and dupes. http://edition.cnn.com/2012/08/02/tech/social-media/facebook-fake-accounts/

[2] News Feed FYI: Further Reducing Clickbait in Feed http://newsroom.fb.com/news/2016/08/news-feed-fyi-further-reducing-clickbait-in-feed/

[3] Ramu violence 2012. http://www.thedailystar.net/city/ramu-mayhems-prime-accused-detained-dhaka-200293

[4] Internet Users. http://www.internetlivestats.com/internet-users/

[5] ICT Facts and Figures 2016 http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx

[6] What Twitter Got Wrong During the Week Following Last Years Boston Marathon. http://archive.boston.com/news/local/massachusetts/2014/04/18/what-twitter-got-wrong-during-the-boston-marathon-bombing-week/ZOYLJpEydYgJ8UYNUT674H/story.html

[7] The cost of ad blocking. https://downloads.pagefair.com/reports/2015\textunderscorereport\textunderscorethe\textunderscorecost\textunderscoreof\textunderscoread\textunderscoreblocking.pdf

[8] Buckland, Michael K. "Information as thing." Journal of the American Society for Information Science (1986-1998) 42.5 (1991): 351.

[9] Fine, Gary Alan, Vronique Campion-Vincent, and Chip Heath, eds. Rumor mills: The social impact of rumor and legend. Transaction Publishers, 2005.

[10] Weeks, Brian Edward. The roles of personal relevance, anxiety, and source medium in understanding belief and transmission of rumors in the News. Diss. University of Minnesota, 2010.

[11] Rosnow, Ralph L. "Inside rumor: A personal journey." American Psychologist 46.5 (1991): 484.

[12] Nekovee, Maziar, et al. "Theory of rumour spreading in complex social networks." Physica A: Statistical Mechanics and its Applications 374.1 (2007): 457-470.

[13] Morris, Meredith Ringel, et al. "Tweeting is believing?: understanding microblog credibility perceptions." Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. ACM, 2012.

[14] Heymann, Paul, Georgia Koutrika, and Hector Garcia-Molina. "Fighting spam on social web sites: A survey of approaches and future challenges." IEEE Internet Computing 11.6 (2007): 36-45.

[15] Mathur, Amrita, and Prachi Gharpure. "Spam Detection Techniques: Issues and Challenges."

[16] Gao, Hongyu, et al. "Detecting and characterizing social spam campaigns." Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. ACM, 2010.

[17] Oh, Onook, Kyounghee Hazel Kwon, and H. Raghav Rao. "An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter during the Haiti Earthquake 2010." ICIS. 2010.

[18] Zhao, Zhe, Paul Resnick, and Qiaozhu Mei. "Enquiring minds: Early detection of rumors in social media from enquiry posts." Proceedings of the 24th International Conference on World Wide Web. ACM, 2015.

[19] Vosoughi, Soroush. Automatic detection and verification of rumors on Twitter. Diss. Massachusetts Institute of Technology, 2015.

[20] Zhang, Qiao, et al. "Automatic Detection of Rumor on Social Network." National CCF Conference on Natural Language Processing and Chinese Computing. Springer International Publishing, 2015.

[21] Ma, Jing, et al. "Detecting Rumors from Microblogs with Recurrent Neural Networks."

[22] Yang, Fan, et al. "Automatic detection of rumor on Sina Weibo." Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. ACM, 2012.

[23] Seo, Eunsoo, Prasant Mohapatra, and Tarek Abdelzaher. "Identifying rumors and their sources in social networks." SPIE defense, security, and sensing. International Society for Optics and Photonics, 2012.

[24] Abbasi, Mohammad-Ali, et al. "Lessons learned in using social media for disaster relief-ASU crisis response game." International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer Berlin Heidelberg, 2012.

[25] Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. "Predicting information credibility in time-sensitive social media." Internet Research 23.5 (2013): 560-588.

[26] Oh, Onook, Manish Agrawal, and H. Raghav Rao. "Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises." Mis Quarterly 37.2 (2013): 407-426.