Event Based Rumor Detection on Social Media for Digital Forensics and Information Security

Anamika Joshi
School Of Computer Science
Devi Ahilya University Indore, India
anajoshi4@gmail.com

Dr. D.S. Bhilare
School Of Computer Science
Devi Ahilya University Indore, India

***Abstract*—**Advancement in information technology such as social networking is on one side is powerful source of news and information and on other side have posed new challenges for those policing cybercrime. Cybercriminals and terrorists are spreading rumors that is unreal or even malicious information on social network which can bring massive panic and social unrest to our community. The rumor detection problem on social network has attracted considerable attention in recent years. A different type of rumors has different characteristics and need different techniques and approaches to detect. In this paper, we proposed an efficient approach to detect event based rumor on social media like Twitter. Experiment illustrates that our event based rumor detection method obtain significant improvement compared with the previous work.

Keywords- *Rumor detection,Digital forensics,Information security,Rumor debunking,Feature based model*

*****

I. INTRODUCTION

Emerging social media offer unprecedented opportunities to the evolution of civil disorders. In particular, Twitter, Facebook and WhattsApp, as both a social network and information-sharing medium, provides a real-time platform for disseminating news and opinions. Social media is being used to democratize today's civil rights movement and fuel citizen protest throughout the world. Now days, Social media plays an important role in social movements, its part of a larger toolbox for activists that can be used along with traditional organizing.

Using the Arab Spring as an example, one can hold that social media itself created the movement. Civil unrest can be triggered by any number of factors, either alone or in combination, including actions by the government (controversial legislation, police brutality) or a powerful third party (criminal gang activity), natural disasters that cause widespread human suffering (severe hurricane, major earthquake), and calls to action by "political entrepreneurs"—leaders with a sizable following— or real-world or online political organizations. But when the same tools and means are used by anti-social or anti national-elements, the results can be scary and grievous.

During the riots in Muzaffarnagar in 2013, government had said that social media was used extensively by anti-social elements to spread rumors, hatred and misinformation among communities [1]. Similarly in 2012, the mass exodus of people of northeast from south India had taken place allegedly due to the misinformation and rumor campaign carried out through Internet and social media [15, 16].

Recently in 2016 and 2017, the Government of Jammu and Kashmir has banned several social media platforms including Facebook, WhatsApp, Twitter and Snap Chat, in the Kashmir Valley "in the interest of maintenance of public order". The government order of 2017, said that they had observed that anti-social elements were "misusing" social media platforms to spread hatred and rumors among the public against the state government and security forces. They said that these platforms were also being used to incite people into committing offences [17].The order said that in 2016, "anti-national and subversive element, inter alia, extensively misused social media websites and instant messaging services for vitiating peace and instigating violence, spreading rumors which caused large-scale damage to life and property"[18].

Anonymity offered by the Internet technology has facilitated communication between the members of terror and anti-national/anti-social groups without much fear of being intercepted by law enforcement agencies.

To further enhance the dangers, as online social networking sites such as Facebook, Twitter and MySpace are becoming popular, they have become the latest targets of hackers to deliver computer viruses to steal critical information and pose a major source of risk to data and information security [21].

In recent years, users and businesses have seen hackers breach all kinds of targets, including individual desktops, mobile devices, POS systems and the corporate network at large. Recently, however, it appears that black hats have widened the scope of their attacks to include not only traditional vectors, but social media as well [22].

When looking at it from a hacker's point of view, it's easy to understand why social networks would be the next sensible target. Oftentimes, cybercriminals go where they have access to the largest pool of victims. With social media websites like Facebook, Twitter, LinkedIn and others steadily increasing their user numbers, it's not difficult to imagine hackers utilizing these platforms to the advantage of their malicious

activities. This poses a serious challenge for law enforcement and digital forensics.

The Koobface, Petya, Wannacry virus family types are an example of computer viruses attempted to gather sensitive information, such as credit card numbers and personal details, from Facebook and MySpace users [19, 20].

Proliferation of rumors, misinformation, spams and malicious content on social media can often have adverse effect on information consumer and society. Misinformation and rumor can spread quickly and it may sensationalize it to wider audiences and even consequential in real world impact. Rumors carrying unreal or even malicious information can bring massive panic and social unrest to any community.

A rumor is a piece of information that has not yet been verified, and hence its truth value remains unresolved while it is circulating. Rumor could be classified as event based and long standing rumor. Event based rumor is a new rumor that emerge during breaking news. This is, for example in the Sydney Siege event a rumor stating that "Hostages are being forced to hold an ISIS flag at a Lindt cafe in Sydney's Martin Place, as police man the doors outside". Long-standing rumors that are discussed for long periods of time. This is, for example, the case of the rumor stating that "*Barack Obama is Muslim"*. Both type of rumors share different type of characteristics that is why need different ways to handle.

In this paper, we address the problem of detecting event based rumors in Twitter data. This paper is organized as follows. Section 2 presents the related work. Section 3 describes our method to rumor detection, especially the process of analyzing and extracting features. Section 4 presents our experiments. The last section draws a conclusion and future work.

II. RELATED WORK

Rumors and related phenomena have been studied from many different perspectives, ranging from psychological studies to computational analyses. The widespread adoption of the Internet gave rise to a new phase in the study of rumour in naturalistic settings and has taken on particular importance with the advent of social media, which not only provides powerful new tools for sharing information but also facilitates data collection from large numbers of participants. For instance, Takayasu et al. [5] used social media to study the diffusion of a rumour in the context of the 2011 Japan Earthquake, which stated that rain in the aftermath might include harmful chemical substances and led to people being warned to carry an umbrella. The authors looked at retweets of early tweets reporting the rumour, as well as later tweets reporting that it was false. Social media as a source for researching rumors has gained ground in recent years, both because it is an interesting source for gathering large datasets associated with rumors and also because it is a type of platform that gives rise to even more rumors from its many participants. Researchers have used social media, among others, to study how users orient to rumors. Castillo et al. [6] found that the ratio between tweets supporting and debunking false rumors was 1:1 (one supporting tweet per debunking tweet) in the case of a 2010 earthquake in Chile. Procter et al. [7] came to similar conclusions in their analysis of false rumors during the 2011 riots in England, but they noted that any self-correction can be slow to take effect. Many existing algorithms [8, 9] for debunking rumors followed the work of Castillo et al. in which they studied information credibility and proposed a set of features that are able to retrospectively predict if an event is credible [10]. The current research of rumor debunking on Twitter and its equivalent in China, SinaWeibo, mainly focused on identifying rumor microblogs [8, 9, 11]| without differentiating kind of rumours. While different types of rumours share different characteristics so they need different attention. In this paper we concentrate on detection event based rumor that emerge in the context of breaking news are generally rumours that have not been observed before. Therefore, rumours need to be automatically detected and a rumour classification system needs to be able to deal with new, unseen rumours, considering that the training data available to the system may differ from what will later be observed by the system.

III. PROPOSED METHOD

We formulate rumor detection as a classification problem. For a given message, features are extracted first from different aspect of view, then we will use a classifier to determine whether this message is rumor or not. In this section we will introduce our general process of rumor detection and some key features that contribute to rumor detection.

A. Rumor Detection Flow

Proposed work is done for identifying a message is rumor or not. Rumor detection method is a classification problem, and it mainly contains 3 parts which are data cleaning, feature extraction and model training as shown in figure 1.In the data cleaning process, we filter out some spam messages like message that contain only URL or punctuation. Feature extraction is a key step in our method. Content and users are two key factors of a message; patterns of these two factors for rumors are different from that of normal messages. We identified a set of features based on contents and users; those significantly contribute to detect rumors. After feature extraction, a classification model has trained using the extracted features.

B. Feature Extraction

Different types of rumors share different characteristics so they need different attention. We concentrate on detection event based rumor that emerge in the context of breaking news are generally rumors that have not been observed before. Therefore, rumors need to be automatically detected and a rumor classification system needs to be able to deal with new, unseen rumors, considering that the training data available to the system may differ from what will later be observed by the system. Content and users are two key factors of a message;

 ISSN: 2321-8169
1214 – 1218

patterns of these two factors for rumors are different from that of normal messages.

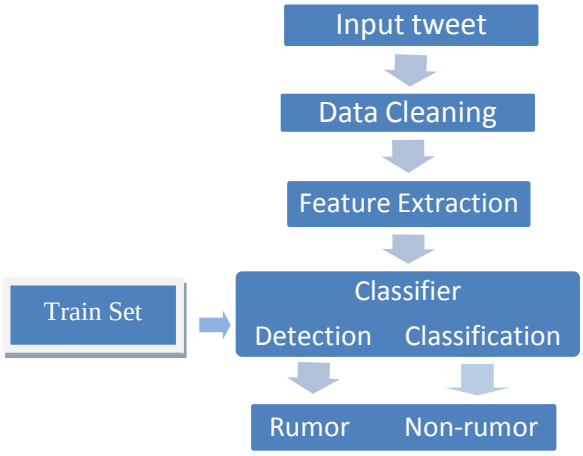


Figure 1: The Framework of proposed method

We identified and merged a set of features that have been proposed by previous work [8, 9, 12] based on contents and users; those significantly contribute to detect event based rumors as listed in table 1.

Table 1: Content and User Based Features

Category	**Name**	Description
User Based Features	Has_Profile_Img	Whether the user has profile image or not.
	Has_Description	Whether the user has personal description or not.
	Reliable	Whether the user's account has been verified by Twitter or not.
	Influence	Number of Followers.
	Has_Profile_URL	Whether the user has profile URL or not.
Content Based Features	Time Span	The time interval between the time of posting and user registration.
	Has URLs	Whether the message includes a URL pointing to an external source.
	Has Refer	Whether the messages include '@' referring to others.

IV. EXPERIMENT

A. *Dataset*

To build a generalized model so that the results can be generalized to other datasets (and events).we used data from two large-scale events. We use annotated dataset and the method proposed by Zubiaga et al. (2016) [13]. Annotated social media corpus collected from the Twitter platform, containing tweets in English related to Sydney siege event and German wings plane crash that had caught the interest of online crowds and were full of substantial amount of rumors:

- Sydney siege: A gunman held hostage ten customers and eight employees of a Lindt chocolate cafe located at Martin Place in Sydney, Australia, on December 15, 2014.
- German wings plane crash: A passenger plane from Barcelona to Dsseldorf crashed in the French Alps on March 24, 2015, killing all passengers and crew. The plane was ultimately found to have been deliberately crashed by the co-pilot of the plane.

Data collection is targeted and event based so annotated dataset of above two different newsworthy events has been used for the study. Annotated tweets, of which 760 were deemed rumors and 930, were deemed non-rumors as shown in Table 2.

Table 2 : Distribution of annotations of rumors and non-rumors

Event	Rumors	Non-rumors	Total
Sydney Siege	522 (42.8%)	699 (57.2%)	1,221
German wings Crash	238 (50.7%)	231 (49.3%)	469
Total	760	930	1,690

Data from one event was used to train a classifier and build a model, with data from another event being used to test the model's generalisability beyond a single event. The tweets collected from Sydney Siege are used as training set, and the tweets collected from German wings Crash are used as test set.

B. *Result Analysis and Evaluation*

There are three main goals to evaluate the event based rumor detection model:

1. Measure the accuracy at which our model predicts the rumor.
2. Measure the contribution of each feature.
3. Measure the contribution of the content based, user based and content-user based features.

We use the method of Yang et al. [14] as a baseline, and train a Logistic Regression classifier with our proposed features. We have used open source tool R and R studio to implement our model. Binomial logistic regression estimates the probability of an event (in this case, rumor) occurring. If the estimated probability of the event occurring is greater than or equal to 0.5 (better than even chance), it classifies the event as occurring (e.g., rumor being present). If the probability is less than 0.5, it classifies the event as not occurring (e.g., no rumor). It is very common to use binomial logistic regression

 ISSN: 2321-8169
1214 – 1218

to predict whether cases can be correctly classified (i.e., predicted) from the independent variables. The statistics summary is given in table 3.

Table 3: Summary Statistics

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.6006	0.9916	8.673	<2e-16 ***
Has_Profile_Img	-4.9072	0.5823	-8.427	< 2e-16 ***
Has_Description	-2.4259	0.4158	-5.834	5.41e-09 ***
Reliable	-1.6253	0.5108	-3.182	0.001463 **
Influence	-2.0741	0.4207	-4.930	8.21e-07 ***
Has_Profile_URL	-3.1173	0.4717	-6.608	3.89e-11 ***
Has_refer	3.4479	0.4901	7.036	1.98e-12 ***
Has_URL	-1.6305	0.4259	-3.828	0.000129 ***
Time.Span	-3.0920	0.5012	-6.169	6.87e-10 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

The output shows the coefficients, their standard errors, the z-statistic (sometimes called a Wald z-statistic), and the associated p-values. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable. In the next column, we see the standard error associated with these estimates. That is, they are an estimate of how much, on average, these estimates would bounce around if the study were re-run identically, but with new data, over and over. If we were to divide the estimate by the standard error, we would get a quotient which is assumed to be normally distributed with large enough samples. This value is listed in under z value. Below Pr(>|z|) are listed the p-values that correspond to those z-values in a standard normal distribution. Lastly, there are the traditional significance stars it determine statistical significance for each of the independent variables. We can see in the summary that all the features are significantly contributing in rumor detection.

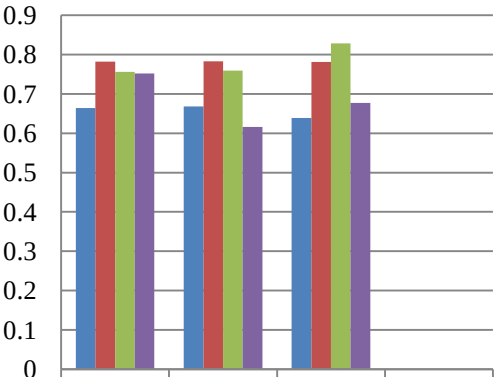
To evaluate the performance of our model, we use the standard information retrieval metrics of precision or accuracy, recall and F1. The accuracy is the ratio of the number of rumors classified corrected to the total number of messages predicted as rumors. The recall is the ratio of the number of rumors classified correctly to the total number of true rumors. The F1 is a comprehensive assessment of accuracy and recall rate, and it is defined as Equation 1.

$$F1 = (2 \times precision \times recall)/precision + recall$$

We conduct experiment in three groups Content Based, User Based and Content-User Based to better understand the impact of features on model. Then we compare the result with baseline. The experimental results in Figure 2 show that there is significant improvement in the performance.

Interpretation: A logistic regression was performed to ascertain the effects of all the selected features on the likelihood that result is a rumor. The logistic regression model was statistically significant and the model predicts the result with the 0.828 F1 value.

Figure 2: Rumor Detection Results



	Accuracy	Recall	F1	
User Based Features	0.664	0.668	0.639	
Content Based Features	0.782	0.783	0.781	
User-Content Based Features	0.756	0.759	0.828	
Baseline	0.752	0.616	0.677	

V. CONCLUSION AND FUTURE WORK

In this paper we focus on detecting event based rumor on social network. To distinguish normal messages from rumors, we propose a rumor detection method based on generalized features of contents and users of messages. In the feature engineering process, we identified and merge some features from the previous work that are significantly contribute in detecting event based rumors.

Current system only works on the data from twitter. So in future data from heterogeneous sources (such as Facebook, Google+ and many more) can be collected for analysis.

REFERENCES

[1] Social media being used to instigate communal riots, says HM Rajnath Singh - http://www.dnaindia.com/india/report-social-media-being-used-to-instigate-communal-riots-rajnath-singh-2032368

[2] AlKhalifa, H.S., AlEidan, R.M.," An experimental system for measuring the credibility of news content in Twitter". International Journal of Web Information Systems **7**(2), 130–151 (2011).

[3] Gang Liang, Wenbo He, Chun Xu, Liangyin Chen, and Jinquan Zeng." Rumor Identification in Microblogging Systems Based on Users Behavior". *IEEE Transactions on Computational Social Systems* 2,3 (2015), 99–108.

[4] Michal Lukasik, PK Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn."Hawkes Processes for continuous time sequence classification: an application to rumor stance classification in Twitter". In *Proceedings of 54th Annual Meeting of the Association for Computational Linguistics*.Association for Computational Linguistics, 393–39,.(2016).

[5] Misako Takayasu, Kazuya Sato, Yukie Sano, Kenta Yamada, Wataru Miura, and Hideki Takayasu. "Rumor diffusion and convergence during the 3.11 earthquake: a Twitter case study." *PLoS one* 10, 4 (2015), e0121443.

[6] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. "Predicting information credibility in timesensitive social media". *Internet Research* 23, 5 (2013), 560–588.

[7] Rob Procter, Farida Vis, and Alex Voss. 2013b." Reading the riots on Twitter: methodological innovation for the analysis of big data". *International journal of social research methodology* 16, 3 (2013), 197–214.

[8] K. Wu, S. Yang, and K. Q. Zhu. "False rumors detection on sina weibo by propagation structures". In IEEE International Conference of Data Engineering,2015.

[9] F. Yang, Y. Liu, X. Yu, and M. Yang."Automatic detection of rumor on sina weibo". In Proc. of the ACM SIGKDD Workshop on Mining Data Semantics, page 13, 2012..

[10] C. Castillo, M. Mendoza, and B. Poblete. "Information credibility on twitter". In Proc. International Conference on World Wide Web, pages 675-684, 2011.

[11] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: "Identifying misinformation in microblogs". In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1589{1599. ACL, 2011.

[12] Laura Tolosi, Andrey Tagarev and Georgi Georgiev. "An Analysis of Event-Agnostic Features for Rumour Classification in Twitter", The Workshops of the Tenth International AAAI Conference on Web and Social Media, Social Media in the Newroom: Technical Report WS-16-19,2016

[13] https://figshare.com/articles/PHEME rumour scheme dataset journalism use case/2068650

[14] Sun, S., Liu, H., He, J., Du, X.: "Detecting event rumors on sina weibo automatically".In: Ishikawa, Y., Li, J., Wang, W., Zhang, R., Zhang, W. (eds.) APWeb 2013.LNCS, vol. 7808, pp. 120–131. Springer, Heidelberg (2013)

[15] Northeasterners' exodus in India underlines power of social media - http://articles.latimes.com/2012/aug/18/world/la-fg-india-social-media-20120819

[16] Social media and the India exodus - http://www.bbc.com/news/world-asia-india-19292572

[17] J&K Bans Facebook, Whatsapp And Most Social Media From Kashmir Valley Indefinitely - http://www.huffingtonpost.in/2017/04/26/jandk-bans-facebook-whatsapp-and-most-social-media-from-kashmir-v_a_22056525/

[18] Pakistan-backed misinformation campaign drawing Kashmiri youth to violence: Army chief Bipin Rawat - http://www.financialexpress.com/india-news/pakistan-backed-misinformation-campaign-drawing-kashmiri-youth-to-violence-army-chief-bipin-rawat/711692/

[19] Petya cyber attack: Ransomware virus hits computer servers across globe, Australian office affected – http://www.abc.net.au/news/2017-06-28/ransomware-virus-hits-computer-servers-across-the-globe/8657626

[20] Ransomware 'Nyetya' behind new global cyber attack: Cisco - http://economictimes.indiatimes.com/articleshow/59349471.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

[21] The Facebook Virus Spreads: No Social Network is Safe - https://readwrite.com/2008/12/10/the_facebook_virus_spreads_no_social_network_is_safe/

[22] Social media malware on the rise – http://blog.trendmicro.com/social-media-malware-on-the-rise

.