Published in final edited form as:

Soc Comput Behav Cult Model Predict (2015). 2015 ; 9021: 121–130. doi: 10.1007/978-3-319-16268-3_13.

Detecting Rumors Through Modeling Information Propagation Networks in a Social Media Environment

Yang Liu[1], **Songhua Xu**[1], and **Georgia Tourassi**[2]

[1] New Jersey Insititute of Technology, University Heights, Newark, NJ 07102, USA

[2] Health Data Sciences Institute, Oak Ridge National Laboratory, Biomedical Science and Engineering Center, 1 Bethel Valley Road, Oak Ridge, TN 37830, USA

Abstract

In the midst of today's pervasive influence of social media content and activities, information credibility has increasingly become a major issue. Accordingly, identifying false information, e.g. rumors circulated in social media environments, attracts expanding research attention and growing interests. Many previous studies have exploited user-independent features for rumor detection. These prior investigations uniformly treat all users relevant to the propagation of a social media message as instances of a generic entity. Such a modeling approach usually adopts a homogeneous network to represent all users, the practice of which ignores the variety across an entire user population in a social media environment. Recognizing this limitation of modeling methodologies, this study explores user-specific features in a social media environment for rumor detection. The new approach hypothesizes that whether a user tends to spread a rumor is dependent upon specific attributes of the user in addition to content characteristics of the message itself. Under this hypothesis, information propagation patterns of rumors versus those of credible messages in a social media environment are systematically differentiable. To explore and exploit this hypothesis, we develop a new information propagation model based on a heterogeneous user representation for rumor recognition. The new approach is capable of differentiating rumors from credible messages through observing distinctions in their respective propagation patterns in social media. Experimental results show that the new information propagation model based on heterogeneous user representation can effectively distinguish rumors from credible social media content.

Keywords

Rumor detection; Heterogeneous user representation and modeling; Information propagation model; Information credibility in social media

yl558@njit.edu.

1 Introduction

With the rapid development of social media sites and proliferation of social media content, user-generated messages can easily reach a large audience. Such a potential for rapid and far-reaching information propagation in social media brings unprecedented challenges in information quality assurance and management. Most social media sites that care about information quality currently gauge information credibility and detect rumors manually. Such practice cannot efficiently and affordably scale up to handle a large volume of messages typically seen in a popular social media environment. The incurred deficiency undesirably facilitates an easy spread of rumors to a large population at a fast pace, leading to elevated societal harm and potential damages. Consequently, automatic detection of rumors in social media is highly desirable and socially beneficial. This paper introduces a new method for automatically detecting rumors circulated in social media through observing their propagation patterns as distinct from those of credible messages. We formulate the social media rumor detection task as a microblog classification problem by introducing an information propagation model built upon a heterogeneous user representation. The derived microblog classifier follows a hypothesis that rumors and credible messages tend to propagate in a social media environment following quantitatively differentiable patterns among a heterogeneous population of human users. By extracting user context sensitive features from individual users' profiles, the new method estimates the retweeting probability of a message when it is a rumor versus a credible message respectively. Given this estimate, the method is able to classify a message as being a rumor or a truthful message by comparing the likelihood that the observed information propagation network of the message is generated by the proposed information propagation model under the rumor and credible message mode respectively.

Although many previous studies focused on Twitter, in this study, we explore the popular Chinese social media site–Sina Weibo [1]–as the experimental social media environment due to three main reasons. First, Sina Weibo is the largest microblog service site and the most popular and influential social media site in China. Like Twitter, on Sina Weibo users can set up their personal profiles and broadcast microblogs to the public. Those broadcast microblogs will appear in the timeline of a user's followers. As of today, it has 1.4 billion users registered under four general user categories, including ordinary individuals, celebrities, enterprises, and government organizations. The number of active users on the site reaches 66 million. Over 125 million microblogs are posted daily. Second, Sina Weibo users tend to share more personal information, such as their ages and education levels, than Twitter users. Being able to access such personally descriptive information is very beneficial for the proposed method. Third, Sina Weibo offers an official rumor busting service, which is not provided by many of today's social media sites, including Twitter. This service gives us a reliable collection of rumors as ground-truth data in this study. The official rumor busting expert team of Sina Weibo assesses the information credibility of suspicious microblogs reported by users through referring to third-party credible information resources and optionally consulting with external domain experts. All detected rumors are posted in a dedicated area on the homepage of Sina Weibo. According to self-released data from Sina Weibo, at least two messages are announced as rumors everyday. Those rumors usually had

been retweeted thousands of times before they were officially denounced. Such a high number of retweets involved indicates that rumor spreading remains as a major problem in Sina Weibo despite the site's serious endeavor to combat the phenomenon.

The main contribution of this paper lies in its proposal of a novel information propagation model for social media environments based upon a heterogeneous user representation. The new information propagation model is capable of characterizing information propagation patterns across different user sub-populations in a social media environment. A direct application of the model is to detect and separate rumors apart from credible messages in a social media environment. We conducted comprehensive experiments using the Sina Weibo platform to explore the usefulness and advantages of applying the new information propagation model for rumor detection.

The rest of the paper is organized as follows: Section 2 briefly discusses previous studies most related to this work. Section 3 describes the proposed algorithmic method for rumor detection in a social media environment. Section 4 reports experimental results obtained using the proposed method for rumor detection. Finally, Section 5 concludes this work.

2 Related Work

Detecting rumors in social media environments has been richly studied and reported in the literature. For example, a variety of microblog classification methods has been attempted to derive features for rumor detection. In a comprehensive study [2], Castillo et al. summarized three main categories of such features, including: (1) message-based features, which focus on characterizing the content of a microblog, (2) user-based features, which consider attributes of site users, (3) propagation-based features, which encode network-based characteristics of a social media environment. In a related study, Qazvinian et al. [3] proposed three categories of features for rumor detection, including: (1) content-based features, which characterize both lexical and part-of-speech patterns of a microblog; (2) network-based features, which are extracted from a user network consisting of both users who spread rumors and those who do not participate in rumor dissemination; and (3) Twitter specific features, including Twitter hash-tags and shared URLs.

It is noted that traditional content-based features generally ignore the rich information regarding human users involved in rumor spreading in a social media environment. In contrast, propagation or network-based features are extracted from a rumor propagation network. Such a network consists of users directly involved in rumor propagation. To extract propagation-based features, rumor propagation models are used. To establish these propagation models, one key task is to estimate the probability of an arbitrary user to spread a rumor. In [4], Moreno et al. modeled rumor propagation by treating a microblog propagation network as a homogeneous scale-free network. They assumed that the probability of a user to forward a microblog is affected by the number of online friends of the user who spread the microblog. Many propagation-based methods alike treat all users in a network as homogeneous instances of a common type of nodes. Their practice ignores any variance among a user population. In reality, however, even when two users have the same number of online friends, they may still respond differently to a rumor because of their

personal abilities in discerning rumors. Xia and Huang argued that a person would believe a rumor if the cumulative influence from his or her online friends regarding the rumor is larger than the person's internal rumor resistance [5]. Their method only observes the number of followers to determine a user's rumor resistance threshold while ignoring many other aspects of user differentiation. Sun et al. found that besides content features, the numbers of followers and followees as well as the age of a user account all affect a user's likelihood to retweet a message [6]. Their method considers the above user-specific features in rumor detection. We recognize that in reality, factors that affect rumor spreading are far richer than the aforementioned ones. To overcome the limitations of all existing methods in comprehensively modeling and leveraging user context for rumor detection, our new method utilizes a wider spectrum of user-specific features available in Sina Weibo when constructing its information propagation model. The design of this new method is also inspired by the work of Jin and Dougherty et al. [7], which employed epidemio-logical models to characterize information cascades of both news and rumors in twitter for rumor detection. Their work reveals that information diffusion patterns for rumors and normal news are different. Learning from their discovery, our information propagation model captures content dissemination patterns in social media environments for rumors and credible messages respectively through two dedicated modes.

Researchers have also explored the particular problem of rumor detection on Sina Weibo. Compared with Twitter, Sina Weibo offers fewer functions to protect user privacy and anonymity yet provides richer services to encourage versatile and open-minded information sharing among users. As a result, users on Sina Weibo generally share more personal information with peers than their counterparts on Twitter. Similar to our proposed method, many studies on Sina Weibo utilize the rich spectrum of user-specific information to classify rumors from credible messages. For example, Yang et. al [1] proposed two features for rumor detection, which are respectively based on the type of computing device used for microblog posting, e.g. a mobile or PC, and the geographical location where a microblog is posted. Sun et. al [8] classified rumors on Sina Weibo that relate to social events into four categories—purely fictitious rumors, time-sensitive event rumors, rumors due to fabricated details, and rumors engineered from mismatched text and pictures. Their work particularly explored the problem of picture misuse as a source of rumors. For readers interested in more studies regarding rumor propagation and detection on Sina Weibo, they are referred to [9–12].

3 Our Method

3.1 Problem Formulation

In this study, we approach the rumor detection task as a classification problem. That is, two classes of microblogs are considered—rumors and credible messages. We consider rumors as popular microblogs primarily conveying misinformation about an event or a fact. In contrast, credible messages are microblogs not carrying such information. Under this problem formulation, the task of rumor detection is reduced to constructing a classifier capable of reliably assigning an appropriate binary class label to any given microblog.

For simplicity, we borrow the term “retweet” to refer to the message forwarding action on Sina Weibo. For each microblog, the proposed method first extracts its propagation network by tracking all its retweets and identifying all users involved in tweeting, receiving, or retweeting the microblog. We also borrow some terminologies from a previous study on rumor propagation network [13] to describe our new algorithm. We define a user's followee as a person whom the user follows. In a microblog's propagation network, we define a *spreader* as a user who receives the microblog from his/her followee and subsequently retweets the microblog. We define a *stifler* as a user who receives a microblog but doesn't retweet it further.

We hypothesize that when a user receives a microblog, the probability that the user will retweet it depends not only upon content features regarding the microblog, but also upon specific attributes of that user, such as the person's age and education background. Based on this hypothesis, we propose our new information propagation model using a heterogeneous user representation for rumor detection.

3.2 Information Propagation Model under Credible Message Mode

When constructing the new information propagation model under the credible message mode, we consider two key influence factors for a user to retweet a credible message as follows.

The first factor is the user's general tendency to retweet a microblog, For a specific user, we select two features to estimate this factor, including: (1) x_1: the number of microblogs retweeted by the user in the past three months; and (2) x_2: the number of the user's retweeted microblogs divided by the total number of microblogs tweeted or retweeted by his/her in the past three months. Such selection of features is based upon both our empirical observations and the experimental finding in Suh's study [6], which points out that a user's past retweeting behaviors affect his or her present behaviors. In the selection process, the time window we examine spans over three months, which is of the same length as the one used in Suh's original study.

The second factor characterizes the influence of the user's followees. In the propagation network for a specific microblog, we define the influence of followees to a user u_i as *INF* (u_i), which is abbreviated as x_3, as follows:

$$INF(u_i) = \sum_{u_j \in In(u_i)} Ret(u_j, u_i)\, IsF(u_j, u_i)\, log|Out(u_j)|/log|In(u_i)|; \quad (1)$$

$$Ret(u_j, u_i) = \begin{cases} 1 & \text{if } u_j \text{ retweets that microblog to } u_i; \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$IsF(u_j, u_i) = \begin{cases} \alpha & \text{if } u_j \text{ and } u_i \text{ follow each other.} \\ 1 & \text{otherwise.} \end{cases} \quad (3)$$

In (1), $In(u_i)$ is the set of followees of u_i; $|In(u_i)$ is the number of followees of u_i; $Out(u_j)$ is the set of followers of u_j; $|Out(u_j)$ is the number of followers of u_j. This formula is designed under the inspiration of Liu's study [13].

We assume that when the user u_i reads a microblog that is retweeted by his/her followees, the probability for u_i to retweet the message, assuming the message is a credible one, can be estimated by the following logistic function:

$$P_c(u_i|\boldsymbol{\beta_1}, \alpha) = \frac{1}{1+e^{-\boldsymbol{\beta_1}\boldsymbol{X_1}}}, \quad (4)$$

where $\boldsymbol{X_1} = [x_1\ x_2\ x_3]$ is the feature vector for u_i and $\beta_1\boldsymbol{X}_1 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_0$. In the above equation, α and β are model parameters.

3.3 Information Propagation Model under Rumor Mode

Besides the above general influence factors, we also assume that rumor propagation is conditioned on a user's ability to verify a message. That is, users who have a high level of verification capability, such as those who command a higher level of capability in independent thinking, a higher knowledge level and/or education level, are much less likely to believe and spread rumors than those who don't possess the above qualifications. We characterize this influence factor through the following user context sensitive features listed in Table 1. Below, we explain some features unfamiliar to Twitter users. Credit score is related to a user's behaviors of tweeting and retweeting rumors, for which Sina Weibo has an official reporting mechanism. Users reported as spreading rumors or other problematic information will be penalized in their credit scores. Personal hashtags are used to describe a user's interests, such as “travel” or “movie.” Sina Weibo experts are users who hold expertise in a certain area and post a certain amount of authoritative tweets. All the above user context sensitive features are available from a user's profile in Sina Weibo. For user context sensitive features listed in Table 1, features f_1 to f_9 are integer variables; features f_{10} to f_{13} are binary variables; the last two features–f_{14}, f_{15}–regarding job and personal descriptions are represented as bags of words. We organize the aforementioned user context sensitive features as a feature vector $\boldsymbol{X_1'}$.

The final form of feature vector for characterizing the propagation of rumorous microblogs is denoted as $\boldsymbol{X_2} = \left[\boldsymbol{X_1}|\boldsymbol{X_1'}\right]$, which is the concatenation of the two feature vectors $\boldsymbol{X_1}$ and $\boldsymbol{X_1'}$ introduced in the above. The probability for a user u_i to retweet a rumorous microblog is estimated by:

$$P_r(u_i|\boldsymbol{\beta_2}, \alpha) = \frac{1}{1+e^{-\boldsymbol{\beta_2}\boldsymbol{X_2}}}, \quad (5)$$

in which β_2 is another set of model parameters, whose dimensionality is higher than that of the feature vector of $\boldsymbol{X_2}$ by one, i.e. 1 + 3 + 15 = 19.

3.4 Microblog Classification

Given a microblog, we first extract its propagation network in which each user functions either as a spreader or a stifler. We then compute the likelihood that this propagation network is generated by the proposed propagation model under the rumor mode (P_r) and non-rumor mode (P_c) respectively. After that, we label the microblog as a rumor if $P_r > P_c$ and a credible message otherwise. The free parameters, α, β_1, and β_2, in our proposed model are optimized using the Maximum Likelihood Estimation (MAE) method.

4 Experiments

4.1 Data Collection

We collect rumors from Sina Weibo's official rumor busting account [14]. The obtained collection contains 400 rumorous microblogs published in 2013 and 2014. Those microblogs are crawled directly from the timeline of the rumor busting account, which announces all confirmed rumorous microblogs following a temporal order. We download all such rumor microblogs as our ground-truth rumor data. For the credible messages, we collect them from Sina Weibo's official hot topic recommendation service [15]. The service publishes up-to-date microblogs about hot topics retweeted most frequently within the previous 12 months. All those microblogs must have been reviewed by Sina Weibo's official censorship team to ensure that they don't contain any misinformation. In our experiments, we randomly select 3600 credible messages published in 2013 and 2014 from this service and label them as “credible.” Table 2 shows some high-level statistics for propagation networks of individual microblogs examined in this study.

4.2 Experimental Results

To leverage the propagation network-based features for rumor detection, we train a classifier according to the description in Sec. 3. We also explored a set of popular features frequently used in previous studies for rumor detection. In the Twitter information credibility study [2], the authors summarize three categories of features used to identify rumors, which are respectively based on messages, users, and propagation patterns. Using these features, we construct a baseline method by training a SVM classifier coupled with a RBF kernel function. We also combine those three categories of features as a combined feature set. We further compare the performance of the proposed method with that of two peer methods for rumor detection on Sina Weibo, i.e. [1,8]. We use the precision, recall, F-rate, and the Area under the ROC Curve (AUC) as the evaluation metrics. Table 3 shows that the proposed method outperforms the SVM classifiers and the two peer methods. The results demonstrate the effectiveness of the proposed user context sensitive information propagation model for rumor detection.

5 Conclusion

In this paper we propose a new information propagation model based on a heterogeneous user representation for automatically differentiating rumors from credible messages in social media. We conducted a series of experiments using the popular Chinese social media site Sina Weibo to explore the effectiveness of the new method. The experimental results

confirm our hypothesis that propagation patterns of rumors and credible messages in social media indeed differ and distinct from each other. Based on the captured distinction between information propagation patterns for rumors versus credible messages, the proposed method can successfully and reliably detect rumors in social media at their early stage of spreading. In principle, the proposed method can be generically applied onto other social network platforms and social media sites similar to Sina Weibo. In particular, for sites like Facebook, which support detailed user profiles, the method would work more effectively than for another group of sites, such as Twitter, where little personal information is available. In the future, we will explore the above potential of this work for application to other social media sites.

Acknowledgments

This study was performed under the Protocol, F 186-14, approved by the Institutional Review Board (HHS FWA #00003246). The study was funded in part by the National Cancer Institute (Grant number: 1R01CA170508) and the Natural Science Foundation of China (NSFC) (Grant number: 61320106008).

References

1. Yang, F.; Liu, Y.; Yu, X.; Yang, M. Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. ACM; 2012. Automatic detection of rumor on sina weibo.; p. 13:1-13:7.
2. Castillo, C.; Mendoza, M.; Poblete, B. Proceedings of the 20th International Conference on World Wide Web. ACM; 2011. Information credibility on twitter.; p. 675-684.
3. Qazvinian, V.; Rosengren, E.; Radev, DR.; Mei, QZ. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2011. Rumor has it: Identifying misinformation in microblogs.; p. 1589-1599.
4. Moreno Y, Nekovee M, Pacheco AF. Dynamics of rumor spreading in complex networks. Physical Review E. 2004; 69(6):066130.
5. Xia, Z.; Huang, LL. Proceedings of the 7th International Conference on Computational Science. Springer; 2007. Emergence of social rumor: Modeling, analysis, and simulations.; p. 90-97.
6. Suh, B.; Hong, L.; Pirolli, P.; Chi, EH. Proceedings of the 2010 IEEE Second International Conference on Social Computing (SocialCom). IEEE; 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network.; p. 177-184.
7. Jin, F.; Dougherty, E.; Saraf, P.; Cao, Y.; Ramakrishnan, N. Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM; 2013. Epidemiological modeling of news and rumors on twitter.; p. 8:1-8:9.
8. 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., editors. APWeb 2013. LNCS. Vol. 7808. Springer; Heidelberg: 2013. p. 120-131.
9. Liao, QY.; Shi, L. Proceedings of the 2013 Conference on Computer Supported Cooperative Work. ACM; 2013. She gets a sports car from our donation: rumor transmission in a chinese microblogging community.; p. 587-598.
10. Lei, K.; Zhang, K.; Xu, K. Proceedings of the 2013 IEEE Global Communications Conference (GLOBECOM). IEEE; 2013. Understanding sina weibo online social network: A community approach.; p. 3114-3119.
11. Cai G, Wu H, Lv R. Rumors detection in chinese via crowd responses. Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2014:912–917.
12. Bao YY, Yi CQ, Xue YB, Dong YF. A new rumor propagation model and control strategy on social networks. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2013:1472–1473.

13. Liu, DC.; Chen, X. Proceedings of the Third International Conference on Multimedia Information Networking and Security (MINES). IEEE; 2011. Rumor propagation in online social networks like twitter-a simulation study.; p. 278-282.
14. Weibo Rumor Busting. http://weibo.com/weibopiyao
15. Weibo Hot Topics. http://d.weibo.com

Table 1

Description of user context sensitive features used in this study

No.	Feature	No.	Feature	No.	Feature
f_1	Age	f_2	Registration time	f_3	Number of followers
f_4	Number of followees	f_5	Number of friends	f_6	Weibo level
f_7	Active days	f_8	Credit score	f_9	Number of hashtags
f_{10}	Is VIP member	f_{11}	Is Weibo expert	f_{12}	Has college degree
f_{13}	Has personal website	f_{14}	Job description	f_{15}	Personal description

Table 2

High-level statistics of microblog propagation networks analyzed in our experimental study

Attributes	Rumor		Credible messages		Both	
	Mean	Std.	Mean	Std.	Mean	Std.
Total number of users in the propagation network	39466.1	1527.3	70961.4	1685.2	61860.9	1634.2
Number of spreaders	3586.7	537.5	5677.1	768.3	5258.5.8	688.4
Number of stiflers	35418.8	11369.5	64384.2	1096.7	61721.3	1329.1
Percentage of spreaders (%)	9.8	1.5	8.1	1.4	8.5	1.5
Average out degree	159.9	14.3	188.6	16.7	171.3	16.5
Depth of the network	9.5	2.1	11.3	2.9	10.1	2.7
Time span in days	7.8	1.2	10.6	1.4	9.2	1.3

Table 3

Comparing the performance of the proposed method with that of SVM classifiers using different sets of features [2], including (1) message-based, (2) user-based, (3) propagation-based, and (4) combined, as well as two peer methods—(5) Yang's [1] and (6) Sun's [8]. “R” represents rumors; “C” represents credible messages; “W. Avg” represents weighted average of rumors and credible messages

No.	Class	Precision	Recall	F-rate	AUC	No.	Class	Precision	Recall	F-rate	AUC
(1)	R	0.708	0.636	0.672	0.659	(4)	R	0.781	0.767	0.772	0.818
	C	0.700	0.648	0.677	0.684		C	0.802	0.730	0.775	0.805
	W.Avg	0.704	0.640	0.669	0.687		W.Avg	0.761	0.782	0.776	0.812
(2)	R	0.719	0.667	0.71	0.728	(5)	R	0.732	0.741	0.735	0.749
	C	0.703	0.728	0.713	0.708		C	0.718	0.729	0.721	0.758
	W.Avg	0.711	0.697	0.710	0.713		W.Avg	0.725	0.733	0.728	0.754
(3)	R	0.687	0.606	0.654	0.685	(6)	R	0.712	0.683	0.703	0.701
	C	0.698	0.769	0.738	0.742		C	0.677	0.679	0.677	0.705
	W.Avg	0.691	0.718	0.702	0.718		W.Avg	0.705	0.680	0.688	0.702
(7)	R	0.829	0.803	0.813	0.839						
	C	0.797	0.789	0.791	0.828						
	W.Avg	**0.812**	**0.793**	**0.799**	**0.831**						

Soc Comput Behav Cult Model Predict (2015). Author manuscript; available in PMC 2016 March 17.