Automatic Detection of Rumor on Social Network

Qiao Zhang[1,2], Shuiyuan Zhang[1,2], Jian Dong[3], Jinhua Xiong[2(✉)], and Xueqi Cheng[2]

[1] University of Chinese Academy of Sciences, Beijing, China
[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
xjh@ict.ac.on
[3] The Third Research Institute of Ministry of Public Security, Shanghai, China

Abstract. The rumor detection problem on social network has attracted considerable attention in recent years. Most previous works focused on detecting rumors by shallow features of messages, including content and blogger features. But such shallow features cannot distinguish between rumor messages and normal messages in many cases. Therefore, in this paper we propose an automatic rumor detection method based on the combination of new proposed implicit features and shallow features of the messages. The proposed implicit features include popularity orientation, internal and external consistency, sentiment polarity and opinion of comments, social influence, opinion retweet influence, and match degree of messages. Experiments illustrate that our rumor detection method obtain significant improvement compared with the state-of-the-art approaches. The proposed implicit features are effective in rumor detection on social network.

1 Introduction

With the development of social network, the amount of information has been growing explosively. However, the quality of information does not become better. All kinds of false information, especially rumor information, have permeated almost every corner of social networks. Therefore, automatic assessment of information credibility has received considerable attention in recent years.

Rumor detection is one of the critical research topics of information credibility. It is often viewed as a tall tale of explanations of event circulating from person to person and pertaining to an object, event, or issue in public concern [6]. The diffusion of rumor is harmful to people's lives and the stability of the society, and it has become a serious concern of social network. Rumor detection is usually modeled as a classification problem based on shallow features of messages, including content and blogger features. But such shallow features cannot distinguish between rumor messages and normal messages in many cases.

In this paper, we also formulate rumor detection as a binary classification problem. and propose an automatic rumor detection classification method based on the combination of new proposed implicit features and shallow features of the

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 113–122, 2015.
DOI: 10.1007/978-3-319-25207-0_10

messages. Shallow features are usually extracted from basic attributes of user or content, while implicit features are generated by mining the deep information of user or content. The implicit features are the most innovative part of the paper. They are obtained by analyzing the popularity, sentiment or viewpoint of message contents and user historical information, including popularity orientation, internal and external consistency, sentiment polarity and opinion of comments, social influence, opinion retweet influence, and match degree of messages. Experiments illustrate that our rumor detection method obtain significant improvement, compared with the state-of-the-art approaches. The proposed implicit features are effective, and make more contribution to rumor detection on social network compared to shallow features.

The rest of this paper is organized as follows: Section 2 gives an overview of related works. Section 3 describes our method to rumor detection, especially the process of analyzing and extracting shallow and implicit features. Section 4 presents our experiments. The last section draws a conclusion.

2 Related Works

There are a large number of related studies on rumor detection. Recently, rumor detection method is mainly based on supervised learning. In other words, it formulates the problem of rumor detection as a classification problem. One key factor of classification model is features, for determine the upper bound of rumor detection performance. Therefore, feature extraction is a critical step of detecting rumors accurately.

Currently, related studies focus on extracting useful and efficiency features for rumor detection. Generally speaking, features for rumor detection can be divided into four types: (1) content-based features; (2) user-based features; (3) propagation-based features; (4) other-based features.

For content-based features, Ratkiewicz et al. (2010) [8] created the Truthy system, identifying misleading political memes on Twitter using content-based features, including hashtags, links and mention.Qazvinian et al. (2011) [7] applied unigrams, bigrams and pos-tagging results to detect rumors. Takahashi et al. (2012) [10] found that vocabulary distribution of rumor messages are different from non-rumor messages, so they computed the ratio of the number of rumor and non-rumor messages vocabulary words, as one of the features to detect rumors.

For the user-based features, Castillo et al. (2011) [2] used registration age, number of user posted messages, number of followers, number of friends and other attributes of users to detect rumors. Al-Khalifa et al. (2011) [1] and Gupta et al. (2012) [3] also used some attributes of users as features.

As to the propagation-base features, Men-doza et al. (2010) [5] analyzed the retweet network topology and found that the diffusion patterns of rumors are different from news, and found that rumors tend to be questioned more than news by the Twitter community. Kwon et al. (2013) [4] discovered that rumor tweets had more cycle volatility, compared with non-rumor tweets. And

they proposed PES (Periodic External Shocks) model to detect rumors. Wu et al. (2015) [11] introduced the propagation tree, and used random walk kernel algorithm to build rumor detection classifiers.

Finally, some other-based features are put forward. Yang et al. (2012) [12] proposed the client program that user has used to post a microblog and the actual place where the event mentioned by the messages has happened. Sun et al. (2013) [9] used the multimedia features of pictures in messages to detect event rumors on Sina microblog.

However, none of these work considered the implicit features of contents and users. In this paper, We propose some innovative features by analyzing the popularity, sentiment or viewpoint of message contents and user historical information, and build a effective classifier to detect rumors.

3 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 a rumor. Features are critical in our method, and we focus on implicit features of contents and users. In this section, we will introduce our general process of rumor detection and some key features that contribute to rumor detection a lot.

3.1 Rumor Detection Flow

Our proposed rumor detection method is a typical classification problems, and it mainly contains 3 parts which are data cleaning, feature extraction and model training.

There are a lot of spam message, these message will cause interference to our approach. In the data cleaning process, we filter out some spam message such as message which only contains URL or punctuation.

Feature extraction is a key step in our method, and we focus on features extracted from message contents and users. Contents and users are two key factors of a message, patterns of these two factors for rumors are obviously different from that of normal messages. We identify a set of implicit features based on contents and users, and these features make a great contribution to detect rumors. Also, we combine some features which have been studied in previous work, including shallow text features of contents and basic attribute features of users.

After feature extraction, a classifier model will be trained using the extracted features. A large amount of supervised model can be used such as Support Vector Machine, Random Forest.

3.2 Content-Based Implicit Features

Popularity Orientation refers to the relevance of message content and the current social hot topics or events. Since many contents of rumors are associated

with the current hot topics, popularity is a valuable feature for rumor detection. Popularity orientation of content is defined as Equation 1.

$$Popularity\_Orientation = max\left(simi(W, T_1), simi(W, T_2), ..., simi(W, T_m)\right) \quad (1)$$

Where W means the keywords set of message, and T_i means a certain category of popular topic words. $simi(W, T_i)$ means the Jaccard similarity between W and T_i .

Internal and External Consistency refers to the correlation between the message content and the content of the corresponding external page. The more relevant they are, the less likely the message is rumor. Internal and external consistency is defined as Equation 2.

$$Internal\_External\_Consistency = \begin{cases} 0, TnotcontainURL \\ max(Rel(T, title), Rel(T, description), Rel(T, keywords), TcontainURL) \end{cases} \quad (2)$$

Where $Rel(T, title)$ means the Jaccard similarity between message T and title of external page.

Sentiment Polarity refers to the sentiment polarity of messages. Subjective information is a key factor of one message, contents of rumors are usually exaggerated, extreme words such as "disfigure", "poisonous" often appear in rumors. In order to get the sentiment polarity of messages, we use classification method together with some dictionary corpus to classify the messages.

In the process of classifying sentiment polarity of messages, we modify the traditional TF-IDF, and propose the TF-FW to compute item weight based on different types of dictionary. FW is defined as Equation 3.

$$FW_w = log_2(level_w + 1) \quad (3)$$

Where $level_w$ means the level of item w and its definition can be found in Table 1.

Table 1. Item level distribution table

Item	Level
Items in Sentiment Dictionary	5
Items in Emotional Dictionary	4
Items in Sensitive Dictionary	3
Items in Punctuation Dictionary	2
Other Items	1

The weight of k-th term can be calculated by Equation 4.

$$weight_k = \frac{(\log(f_k) + 1.0) \times \log_2(level_k + 1)}{\sqrt{\sum_{k=1}^{l} [(\log(f_k) + 1.0) \times \log_2(level_k + 1)]^2}} \tag{4}$$

Where f_k means the frequency of current term in message, l means the number of terms in message, and $level_k$ means the level of current term, which can be computed by Equation 3.

Then the sentiment polarity can be obtained by text classification, we classify the sentiment polarity of a message into three types including positive, negative and neutral.

Opinion of Comments refers to the degree of acceptance of the comments in message. A large number of doubtful and inquiring comments, such as "Gab", "Fake", always appear in rumor messages. Therefore, opinion analysis carried out on the comments can be used to obtain the credibility of the message. In this paper, we first get the opinion polarity of comment via classification method, the classification process is the same with that of sentiment polarity. Then the opinion of comments is defined as Equation 5.

$$Opinion\_Of\_Comments = \log \frac{N_{pos}}{N_{neg}} \tag{5}$$

Where N_{pos} means the number of comments supported, and N_{neg} means the number of comments nonsupport.

3.3 User-Based Implicit Features

Users are the core of social network. In the process of rumor propagation, user is not only the producer of rumors, but also is the disseminator of rumors. We use social influence to measure the impact of one user in the information diffusion process.

Social Influence refers to the communicative influence of one user on social network. It has a great relationship with the number of followers and friends of this and is defined as Equation 6.

$$Social\_Influence = \log\left(\frac{fol\_num - bi\_fol\_num}{fri\_num + 1}\right) \tag{6}$$

Where fol_num means the number of followers and fri_num means the number of friends, and bi_fol_num means the number of both followers.

Opinion Retweet Influence refers to the degree of acceptance by other users of the user's opinion. Generally speaking, the users with higher degree of

acceptance by others, the lower the probability of publishing rumors. It can be calculated by Equation 7.

$$Opinion\_Retweet\_Influence = \frac{retweets\_num}{statuses\_num} \tag{7}$$

Where *retweets_num* means the total retweet number of user's all messages, and *statuses_num* means the total number of user's messages.

Match Degree of Messages refers to the match degree of the user's professional orientation and his message contents. The subject of rumors often serious discrepancies with the theme of the users historical content. We use topic model to get the distribution of user's historical contents and current message content. These distributions are viewed as a mixture of various topics, we use the match degree of historical and current messages' topic distribution to measure the the feature. It is defined by Equation 8 as follows.

$$\begin{aligned} Match\_Degree\_of\_Messages &= cosin\_simi(his\_topic, cur\_topic) \\ &= \frac{his\_topic \times cur\_topic}{|his\_topic| \times |cur\_topic|} \end{aligned} \tag{8}$$

Where *his_topic* and *cur_topic* are the topic distributions of the user's historical and current messages respectively.

3.4 Feature Fusion

In this part, we also use some shallow features that have been proposed in previous work to strengthen our model, and merge these features with implicit features proposed in this paper.

For content-based feature,features we extract are listed in Table 2.

For user-based feature, we choose some attributes of user profile, these are listed in Table 3.

4 Experiments

4.1 Dataset

We use the method proposed by Yang et al. (2012) [12] to collect Sina Weibo rumor messages from Sina Weibo Community Management Center. The crawled microblogs are used as out dataset, and it contains 3229 rumor microblogs, and 12534 non-rumor microblogs from the same time span. Two-thirds of them are used as training set, and the rest as test set.

Table 2. Fusion Features of Contents

Category	Features	Description
Shallow Text 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_Multimedia	Whether the messages contains picture, videos, or audios
	Has_Refer	Whether the messages include '@' referring to others
implicit Content Features	Popularity Orientation	The popularity orientation of the message
	Internal and External Consistency	The internal and external consistency of the message
	Sentiment Polarity	The sentiment polarity of the message
	Opinion of Comments	The opinion of comments of the message

Table 3. Fusion Features of Users

Category	Features	Description
Basic Attribute Features	Verify_Type	The verify type of the user
	Gender	The gender of the user
	Has_Description	Whether the user has personal description
	Has_Profile_URL	Whether the user has profile URL
	Has_Domain	Whether the has domain
	Post_Num	The number of messages posted by the user
	User_Activity	The activity of the user
	Favorite_Num	The number of favorite messages
	Fans_Rate	The number of user's followers
	Friends_Rate	The number of user's friends
	Bi_Follow_Num	The number of users' binary followers
implicit User Features	Social_Influence	The user's social influence
	Opinion Retweet Influence	The user's opinion retweet influence
	Match Degree of Messages	The user's match degree of messages

4.2 Evaluation

To evaluate the performance of our methods, we use the standard information retrieval metrics of precision, recall and F1 [13]. The precision is the ratio of the number of rumors classified corrected to the total number of microblogs 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 precision and recall rate, and it is defined as Equation 9.

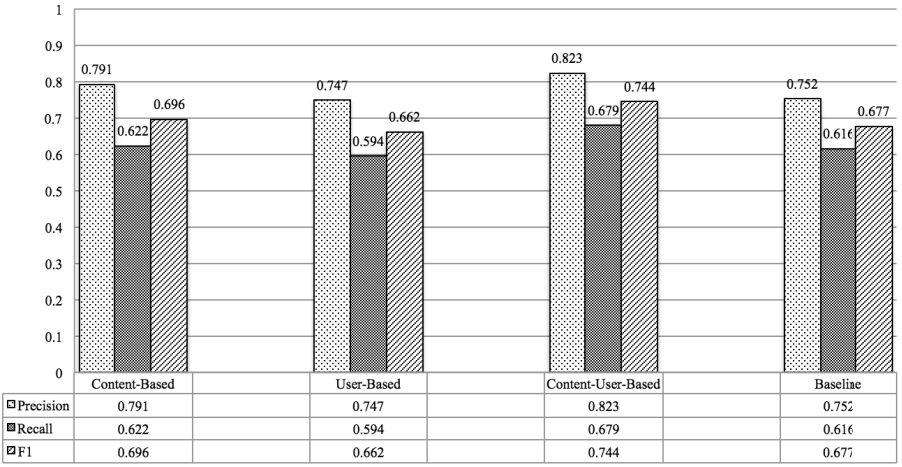


Fig. 1. Rumor detection results with SVM

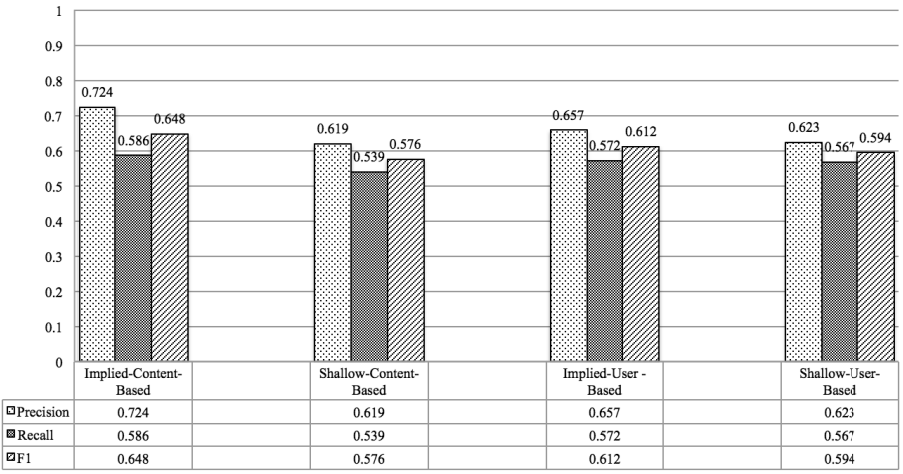
$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{9}$$

We use the method of Yang et al. (2012) [12] as baseline, and train a Support Vector Machine classifier with our proposed features. We conduct three sets of experiments to better understand the impact of different classification method and implicit features on rumor detection. We use Content-Based, User-Based, Content-User-Based and Baseline to indicate different features used in rumor detection.

The experimental results in Figure 1 show that the method combined with content and user features is better than others, with 7.1% improvement in precision and 6.3% improvement in recall rate combined with baseline. The reason is that in our method, we merge the implicit content and user features to improve the effectiveness on detecting rumors.

In order to further assess the effectiveness of implicit features proposed by this paper, we use shallow text features, implicit content features, user attribute features and implicit user features alone to detect rumors, and denote as Shallow-Content-Based, Implicit-Content-Based, Shallow-User-Based and implicit-User-Based respectively. Figure 2 is the result of using different types of features above.

The experimental results show that Implicit-Content-Based method have significant improvement compared with Shallow-Content-Based method, with 10.5% improvement in precision and 4.7% in recall rate. The main reason is that the implicit content features have better identification of rumor detection. As to the implicit-User-Based method, it also achieves better performance than Shallow-User-Based, and fully verified the effectiveness of the implicit features we proposed.



	Implied-Content-Based	Shallow-Content-Based	Implied-User-Based	Shallow-User-Based
Precision	0.724	0.619	0.657	0.623
Recall	0.586	0.539	0.572	0.567
F1	0.648	0.576	0.612	0.594

Fig. 2. RumorRumor Detection with Different Types of Features

5 Conclusion

In this paper we focus on detecting rumor on social network. To distinguish rumors from normal messages, we propose a rumor detection method based on implicit features of contents and users.

In the feature engineering process, we introduce some implicit features based on the characteristic of rumors. These features focus on popularity orientation, internal and external consistency, sentiment polarity and opinion of comments, social influence, opinion retweet influence, and match degree of messages.

User credibility is an important factor that impact information credibility, so the analysis of user credibility help detect message credibility. In the future, we would like to do some work on user credibility, and use this to improve the performance of rumor detection.

Acknowledgments. This research was supported by the National High Technology Research and Development Program of China (Grant No. 2014AA015204), the National Basic Research Program of China (Grant No. 2014CB340406), the NSFC for the Youth (Grant No. 61402442) and the Technology Innovation and Transformation Program of Shandong (Grant No.2014CGZH1103).

References

1. 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)
2. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th, pp. 675–684 (2011)

3. Gupta, A., Kumaraguru, P.: Credibility ranking of tweets during high impact events. In: Proceedings of Workshop on Privacy and Security in Online Social Media Ser Psosm (2012)
4. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y.: Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th International Conference on Data Mining (ICDM), pp. 1103–1108 (2013)
5. Mendoza, M., Poblete, B., Castillo, C.: Twitter under crisis: Can we trust what we rt? In: Proceedings of the First Workshop on Social Media Analytics (2010)
6. Peterson, W.A., Gist, N.P.: Rumor and public opinion. American Journal of Sociology **57**(2), 159–167 (1951)
7. Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q.: Rumor has it: Identifying misinformation in microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1589–1599. Association for Computational Linguistics (2011)
8. Ratkiewicz, J., Conover, M.D., Meiss, M., Goncalves, B., Flammini, A., Menczer, F.M.: Detecting and tracking political abuse in social media. In: Proceedings of Icwsm (2011)
9. 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)
10. Takahashi, T., Igata, N.: Rumor detection on twitter. In: 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS), pp. 452–457 (2012)
11. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on sina weibo by propagation structures. In: IEEE International Conference on Data Engineering, ICDE (2015)
12. Yang, F., Liu, Y., Yu, X., Yang, M.: Automatic detection of rumor on sina weibo. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics (2012)
13. Yang, Y.: An evaluation of statistical approaches to text categorization. Information Retrieval **1**(1–2), 69–90 (1999)