Automatic Detection of Rumor on Sina Weibo

Fan Yang[1,3]
hkyang8@gmail.com

Xiaohui Yu[1,2,3]*
xhyu@yorku.ca

Yang Liu[1,3]
yliu@sdu.edu.cn

Min Yang[1,3]
yangm1022@gmail.com

[1] School of Computer Science and Technology, Shandong University, Jinan, China
[2] School of Information Technology York University, Toronto, Canada
[3] Shandong Provincial Key Laboratory of Software Engineering

ABSTRACT

The problem of gauging information credibility on social networks has received considerable attention in recent years. Most previous work has chosen Twitter, the world's largest micro-blogging platform, as the premise of research. In this work, we shift the premise and study the problem of information credibility on Sina Weibo, China's leading micro-blogging service provider. With eight times more users than Twitter, Sina Weibo is more of a Facebook-Twitter hybrid than a pure Twitter clone, and exhibits several important characteristics that distinguish it from Twitter. We collect an extensive set of microblogs which have been confirmed to be false rumors based on information from the official rumor-busting service provided by Sina Weibo. Unlike previous studies on Twitter where the labeling of rumors is done manually by the participants of the experiments, the official nature of this service ensures the high quality of the dataset. We then examine an extensive set of features that can be extracted from the microblogs, and train a classifier to automatically detect the rumors from a mixed set of true information and false information. The experiments show that some of the new features we propose are indeed effective in the classification, and even the features considered in previous studies have different implications with Sina Weibo than with Twitter. To the best of our knowledge, this is the first study on rumor analysis and detection on Sina Weibo.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-*Data Mining*

General Terms

Algorithm

*Corresponding author.

Keywords

Rumor Detection, Sina Weibo, Classification

1. INTRODUCTION

With the rise of micro-blogging platforms, information is generated and propagated at an unprecedented rate. The automatic assessment of information credibility therefore becomes a critical problem, because there is often not enough resource to manually identify the misinformation about a controversial and large scale spreading news from the huge volume of fast evolving data.

Whereas most previous work has used Twitter as the premise of study, we in this work choose to study the problem of automatic rumor detection on Sina Weibo, due to its wide popularity and unique characteristics. Sina Weibo is China's largest micro-blogging service. Launched by Sina Corporation in late 2009, Sina Weibo now has more than 300 million registered users (eight times more than Twitter as of May 2011), generating 100 million microblogs per day[1]. Sina Weibo is used by more than 30% of the Internet, and is the most of the most popular websites in China.

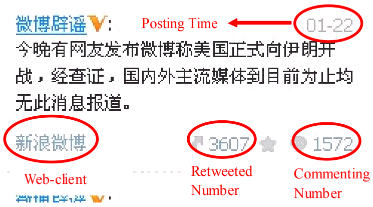
Rumors present a serious concern for Sina Weibo. Statistics show that there is at least one rumor that is widely spread on Sina Weibo every day. For example, at the end of April 2011 a rumor that stated that "the National Statistics Bureau announced that China's urban per capita income has reached 9000 RMB mark" [2] caused a large scale of forwarding. There are about 200 thousands microblogs about that rumor.

There are some major differences between Sina Weibo and Twitter with respect to rumor analysis and detection, which must be taken into consideration: (1) Some linguistic features that are studied in previous work for English tweets, such as the case sensitivity of English words, repeated letters, and word lengthening, do not apply to the Chinese language that dominate Sina Weibo. (2) The types of trending microblogs retweeted (forwarded) are different in Sina Weibo than in Twitter. In Sina Weibo, most trends are created due to retweets of media content such as jokes, images and videos, whereas on Twitter, the trends tend to have more to do with current global events and news stories. (3) Sina Weibo has an official service for rumor busting (with

[1] The Sina corporation annual report 2011 is available (in Chinese) at http://news.sina.com.cn/m/2012-02-29/102024034137.shtml

[2] (In Chinese) http://www.my1510.cn/article.php?id=58593.

Figure 1: Instance of Sina Weibo Rumor-Busting



微博辟谣：

今晚有网友发布微博称美国正式向伊朗开战，经查证，国内外主流媒体到目前为止均无此消息报道。

新浪微博

the user name of “Weibo Rumor-Busting” if translated into English), which focuses on busting those wide spread rumors. While Twitter does not have this type of service. For instance (see Figure 1), this is a rumor-related microblog about United States officially declaring war to Iran at January 23 2012. The original message caused 3607 reforward times and 1572 commented times. Its left bottom shows the function of microblog posting program client, and that represent the web-program-used client. As Sina Weibo provides this authoritative source for verifying information, the datasets we collected are almost referred to widely spread rumor. We classify these rumor-related microblogs in two sets and label them as whether the microblog is true information (the orientation of the microblog is not in accordance with the rumor) or false information (the orientation of the microblog is in accordance with the rumor).

In this paper, we formulate the problem of rumor detection as a classification problem, and build classifiers based on a set of features related to the specific characteristics of Sina Weibo micro-bloging service. The corpus is built by collecting the rumors that are announced by Sina Weibo's official rumor-busting service, along with the microblogs related to those rumors. In total, 19 features are extracted from each microblog, including the content, the micro-blogging client program used, the user account, the location, the number of replies and retweets, etc. We find that the client program used for microblogging and the event location, two features that have not been previously studied, are particular useful in classifying rumors on Sina Weibo. Our experiments also show some interesting results with respect to the effectiveness of various features.

The rest of the paper is organized as follows: in Section 2 we give an overview of related work. In Section 3 we describe how we collect and annotate data. In section 4 we show how to analyze and extract features based on those rumor-related topics announced by Sina Weibo's rumor-busting account, and provide a description of two new features, the client program used and event location. In Section 5 we present the experimental results. Section 6 concludes this paper.

2. RELATED WORK

There is an extensive body of related work on misinformation detection. In this section, we focus on providing a brief review of the work most closely related to our study. We outline related work in three main areas: rumor analysis, features for classification, and data collection and annotatation.

2.1 Analyzing Rumors

Rumor has been a research subject in psychology and social cognition for a long time. It is often viewed as *an unverified account or explanation of events circulating from person to person and pertaining to an object, event, or issue in public concern* [10]. Bordia et al. [1] propose that transmission of rumor is probably reflective of a “collective explanation process”. In the past, the spread of rumors can only be diffused by mouth to mouth. The rise of social media provides an even better platform for spreading rumors.

There have appeared some recent studies on analyzing rumors and information credibility on Twitter, the world's largest micro-blogging platform. Castillo et al. [3] focus on automatically assessing the credibility of a given set of tweets. They analyze the collected microblogs that are related to “trending topics”, and use a supervised learning method (J48 decision tree) to classify them as credible or not credible. Qazvinian et al. [11] focus on two tasks: The first task is classifying those rumor-related tweets that match the regular expression of the keyword query used to collect tweets on Twitter Monitor. The second task is analyzing the users' believing behaviour about those rumor-related tweets. They build different Bayesian classifiers on various subsets of features and then learn a linear function of these classifiers for retrieval of those two sets. Mendoza et al. [8] use tweets to analyze the behavior of Twitter users under bombshell events such as the Chile earthquake in 2010. They analyze users' retweeting topology network and find the difference in the rumor diffusion pattern on Twitter environment than on traditional news platforms.

2.2 Features for Classification

Feature extraction is an important step in a classification task. Generally speaking, various sets of feature are extracted from different corpora. Castillo et al. [3] use four types of features: (1) message-based features, which consider characteristics of the tweet content, which can be categorized as Twitter-independent and Twitter-dependent; (2) user-based features, which consider characteristics of Twitter users, such as registration age, number of followers, number of friends, and number of user posted tweets; (3) topic-based features, which are aggregates computed from message-based features and user-based features; and (4) propagation-based features, which consider attributes related to the propagation tree that can be built from the retweets of a specific tweet.

Qazvinian et al. [11] use three sets of features, which are content-based features, network-based features, and Twitter-specific memes. For content-based features, they follow Hassan et al. [6], and classify tweets with two different patterns: lexical patterns and part-of-speech patterns. For network-based features, they build two features to capture four types of network-based properties. One is the log-likelihood that $user_i$ is under a positive user model, and another feature is the log-likelihood ratio that the tweet is retweeted from a $user_j$ who is under a positive user model than a negative user model. Finally, the Twitter-specific memes features that have been studied in [12] are extracted from memes which are particular to twitter: hash-tags and URLs.

For our work, we consider some features that have been proposed in previous work, such as the number of posted microblogs or retweeted microblogs. We also propose two new features, the location of event, and the client program used for posting the microblog, which have not been studied in previous work.

2.3 Methods For Data Collection and Annotation

Qazvinian et al. [11] use Twitter's search API with regular expression queries, and collect data from the period of 2009 to 2010. Each query corresponds to a popular rumor that is listed as "false" or only "partly true" on About.com's Urban Legends reference site[3]. During the annotation process, they let two annotators scan the dataset and label each tweet with a "1" if it is related to any of the rumors, and with a "0" otherwise. They use this annotation in analyzing which tweets match the regular expression query posed to the API, but are not related to the rumor. And then they asked the annotators to mark each tweet with "11" if the user believes the rumor and with "12" if the user does not believe or remains neutral in the previous annotated rumor-related dataset. They use the second annotated dataset to detect users' beliefs in rumors.

Castillo et al. [3] use keyword-based query interface provided by Twitter Monitor to collect data. They separate the collected topics into two broad types: news and conversation. For annotation, they use Amazon Mechanical Turk[4], a crowdsourcing website that enables netizens to co-ordinate the use of human intelligence to perform tasks that computers are unable to do yet.

3. DATA COLLECTION AND ANNOTATATION

As of February 2011, Sina Weibo reports that its registered users post more than 100 million microblogs per day. This makes Sina Weibo an excellent case to analyze disinformation in online social network. We first build a high quality dataset by using Sina Weibo's official rumor busting service. Those microblogs we collected consist of true information and false information for some specific and happened events, and almost of them are relevant to the rumor topics announced by the rumor busting service, and also the work of labeling the dataset. Therefore, in this work, the labeling is done by an authoritative source, avoiding the errors in judgment when human participants annotate. This section describes how we collected a set of messages related to rumor events from Sina Weibo.

3.1 Data Collection

As Sina Weibo has an official rumor busting account, an unique function of this service that other microblogging services do not have. Topics it announces as rumors are all confirmed false information that is related to controversial events and has been widely spread. For every event considered, we use the form of keyword-based query defined by Twitter Monitor [7]. The form of query is A $\wedge$ B where A is a conjunction of event participants and B is a disjunction of some descriptive information about the event. For example, one querying form as (US $\wedge$ Iran) $\wedge$ (declare $\vee$ war) refers to the rumor about U.S. officially declaring war on Iran on January 23, 2012.

We collect microblogs matching the keywords in the topics published by the rumor busting account from March 1, 2010 to February 2, 2012. The dataset thus collected can be divided into two subsets, including one that contains microblogs related to the rumors and the other that contains those microblogs that match the querying keywords but are directly related to the specific rumor. As the querying keywords are based on the topics announced by the official account, the number of rumor-related microblogs in the collected dataset is quite high.

3.2 Data Annotation

We ask two annotators to go through all microblogs in the dataset independently and eliminate microblogs that are not related to any rumor topics published by Sina Weibo's official rumor-busting account. We also ask annotators to label each microblog kept with "1" if the orientation of the microblog is in accordance with the rumor, and with "-1" otherwise.

We manually processed 5,144 microblogs, only 7 of which match the querying keywords but are not related to the rumor topics. Moreover, among those microblogs that are related to rumors, about 18.3 % are labeled with "1".

We calculate the κ statistic to measure the inter-rater agreement. The κ statistic is defined as

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where $\Pr(a)$ is the relative observed agreement among annotators, and $\Pr(e)$ is the probability of chance agreement [2] [4]. In our case, we have $\kappa = 0.95$ with confidence interval $C.I. = 95\%$, demonstrating that the two annotators can reach a high level of agreement in identifying rumors.

4. FEATURES

We identify a set of features that can be extracted from the microblogs for the classification purpose. These include several features that are specific to the Sina Weibo platform, but most of them are quite general and can be applied to other platforms. Some of the features have been studied in previous works [3] [11] [9]. In addition, we propose two new features that have not been studied in previous works. The set of features are listed in Table 1. We divide these features into five types: content-based features, client-based features, account-based features, propagation-based features, and location-based features.

In what follows, we first describe the features that have been proposed in the previous work and are adopted in our study, and then provide a detailed description of the newly proposed features.

4.1 Previously Proposed Features

Content-based features consider attributes related to the microblog content, which include whether it contains a picture or URL, the sentiment of a microblog (measured by the number of positive/negative emoticons used), and the time interval between the microblog's time of posting and the user's registration time.

Account-based features consider the characteristics of users, which can be personal dependent or personal independent. Personal dependent features include whether the user's identity is verified, whether the user has a personal description, the gender of user, the age of the user, the type of user name and user's logo. We found that among the confirmed rumor topics, the proportion of microblogs posted by non-organizational users that have the default or a cartoon logo is particular high. Personal independent features include the number of followers, the number of friends, and

[3] http://urbanlegends.about.com

[4] https://www.mturk.com/mturk/welcome

Table 1: Description of features

Category	Features	Description
CONTENT	HAS MULTIMEDIA	Whether the microblog contains pictures, videos, or audios
	SENTIMENT	The numbers of positive and negative emoticons used in the microblog
	HAS URL	Whether the microblog includes a URL pointing to an external source
	TIME SPAN	The time interval between the time of posting and user registration
CLIENT	CLIENT PROGRAM USED	The type of client program used to post a microblog: web-client or mobile-client
ACCOUNT	IS VERIFIED	Whether the user's identity is verified by Sina Weibo
	HAS DESCRIPTION	Whether the user has personal descriptions
	GENDER OF USER	The user's gender
	USER AVATAR TYPE	Personal, organization, and others
	NUMBER OF FOLLOWERS	The number of user's followers
	NUMBER OF FRIENDS	The number of users who have a mutual following relationship with this user
	NUMBER OF MICROBLOGS POSTED	The number of microblogs posted by this user
	REGISTRATION TIME	The actual time of user registration
	USER NAME TYPE	Personal real name, organization name, and others
	REGISTERING PLACE	The location information taken at user's registration
LOCATION	EVENT LOCATION	The location where the event mentioned by rumor-related microblogs happened
PROPAGATION	IS RETWEETD	Whether the microblog is original or is a retweet of another microblog
	NUMBER OF COMMENTS	The number of comments on the microblog
	NUMBER OF RETWEETS	The number of retweets of the microblog

the number of microblogs which have been posted by the user.

Propagation-based features consider attributes related to propagation of the rumor, such as whether the microblog is an original post or a retweet from another microblog, the number of comments, and the number of retweets it has received.

4.2 New Features

Client-based feature refers to the client program that user has used to post a microblog. It contains non mobile-client program and mobile-client program two types. The non mobile-client program includes Sina Weibo web-app, timed-posting tools and embedded Sina Weibo's third party applications. The mobile-client program type includes mobile phone based client and Tablet PC based client.

Location-based feature refers to the actual place where the event mentioned by the rumor-related microblogs has happened. We distinguish between two types of locations, domestic (in China) and foreign.

For the aforementioned microblog dataset, the distributions of values of the two features, the client program used and event location, are shown in Figure 2 and Figure 3 respectively. As shown in Figure 2, about 71.8% of false information is posted by non-mobile client programs. In our collected rumor-related microblogs, there is a significant difference in the proportion between domestic and foreign events for true and false information, as shown in Figure 3. For microblogs containing false information, about 56.1% of the events occurred abroad. For those containing true information, on the other hand, the majority of the events (82.3%) are domestic.

In addition, we find that if a microblog describes an event that happened abroad and the client program used is non-mobile (such as Web-based or timed posting tools), then it is a rumor with high probability. For example, on January 22, 2012 (the Chinese New Year), there appeared a microblog about The United States formally declaring war against Iran. It was forwarded (retweeted) 949 times in less than 12 hours, among which 77.77% were done by Webbaesd or other timed-posting clients, much higher a percentage than the average usage frequency of those clients.

Table 2: Hypothesis Test of the Independence between the Client Program Feature and Microblogs' Truthfulness

H_0	The client program used feature is independent of the truthfulness of a microblog
H_1	The client program used feature is not independent of the truthfulness of a microblog

As the content-based features, account-based features, and propagation-based features have been studied in the previous works [3] [11], we here just identify the effectiveness of the two new features that we proposed. In order to test whether the two proposed features are significant indicators of the truthfulness of microblogs, we use Pearson's chi-squared test (χ^2) to perform the *test of independence* between the client program feature and the truthfulness; the same is done for the event location feature as well. For the client program feature, we make the null hypothesis and alternative hypothesis about the independence between the client program feature and the microblogs' truthfulness in Table 2.

The formulas and notation used for the test are summarized in Table 3. The null hypothesis is that the client program used is statistically independent of the truthfulness of a microblog. The observed frequency, $O_{i,j}$, is the frequency of a microblog taking the i-th value of the client program

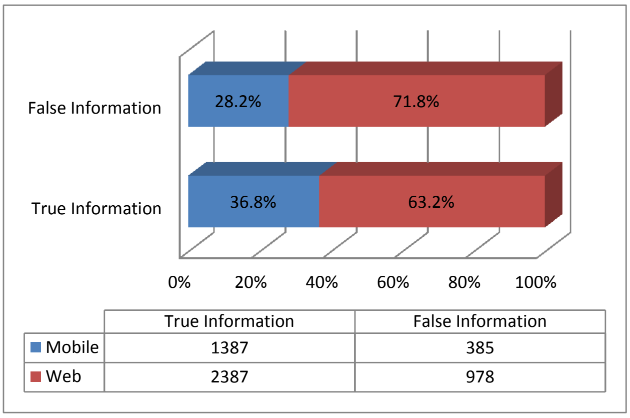


Figure 2: The distribution of client program used on Sina Weibo

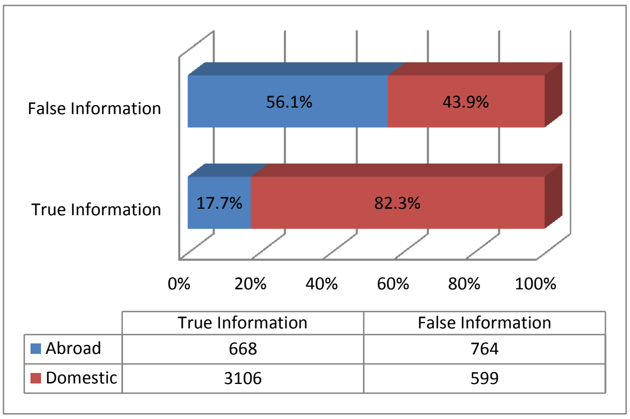


Figure 3: The distribution of event location in rumor-related microblogs

Table 3: Summary of Notations Used In the Independence Test

$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$	
$E_{i,j} = \frac{(\sum_{n_c=1}^{c} O_{i,n_c}) \cdot (\sum_{n_r=1}^{r} O_{j,n_r})}{N}$	
d	The degrees of freedom which value is equal to the number of $(r-1) \cdot (c-1)$
α	The degree of confidence
r	The number of table's rows
c	The number of table's column
O_{ij}	An observed frequency
E_{ij}	An expected frequency, asserted by the null hypothesis
n	The number of cells in the table
N	The total sample size (the sum of all cells in the table)

Table 4: Test of Independence between Client Program Used and Truthfulness

Observed Frequency			
	True Info.	False Info.	Total
Web	2387	978	3365
Mobile	1387	385	1772
Total	3774	1363	5137
Expected Frequency			
Web	2472.164688	892.8353124	
Mobile	1301.835312	470.1646876	
$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$			
Web	2.933875742	8.12358551	
Mobile	5.571383675	15.42656052	
Test-Statistical Value			
$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$	32.05540545		
$\chi^2_{\alpha=0.05,d=1}$	3.841458821		

and the j-th value of truthfulness. $E_{i,j}$, is the expected frequency of this combination assuming they are independent. As shown in the Table 4, the degrees of freedom is $d = 1$. For the test of independence, a chi-squared probability of less than or equal to 0.05 is commonly interpreted as ground for rejecting the null hypothesis [5]. For our case, the χ^2 value is calculated by the inverse function with $\alpha = 0.05$ and $d = 1$, which results in 3.841458821. We calculate the expected frequency of each cell, as shown in the Table 4. The test statistic (chi-square value) is greater than the threshold $(\chi^2 = 32.05540545) > (\chi^2_{\alpha=0.05,d=1} = 3.841458821)$. Therefore, we reject the null hypothesis H_0: *The client program used feature is independent to the truthfulness of a microblog.* This clearly indicates that the client program feature has a nontrivial relationship with the truthfulness, and can be used as a feature in the rumor classification task.

We can similarly perform the independence test between the event location feature and the microblog's truthfulness, and result is shown in Table 5. The test also confirms that the event location is not independent of the truthfulness, and can be used as a good indicator for classification.

5. EXPERIMENT

In order to better understand the impact of various categories of features on identifying the truthfulness of rumor-related microblogs, we conduct two sets of experiments at the feature level, in which we systematically include/exclude the features mentioned above to measure their effect. In the first set of experiment, we train a classifier using specific subsets of the previously proposed features to study how well those subsets of features perform in rumor detection on Sina Weibo. In the second set of experiments, we study the impact of incorporating the two newly proposed features.

5.1 Effect of Previously Proposed Features

We first consider the three subsets of features that have been proposed in the literature: content-based features, account-based features, and propagation-based features. We train a SVM classifier with RBF kernel function ($\gamma = 0.313$, obtained through 10-fold cross validation strategy) using the above mentioned three subsets of features respectively to measure the impact of those features on the classification performance for the rumor related corpus. For example, in the first experiment, we only use the content-based features;

Table 5: Test of Independence between Event Location and Truthfulness

Observed Frequency			
	True Info.	False Info.	Total
Abroad	668	764	1432
Domestic	3106	599	3665
Total	3774	1363	5137
Expected Frequency			
Abroad	1052.047499	379.9525015	
Domestic	2692.5657	972.4343002	
$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$			
Abroad	140.1956483	388.1866301	
Domestic	63.48142984	143.4062708	
Test-Statistical Value			
$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$	735.269979		
$\chi^2_{\alpha=0.05,d=1}$	3.841458821		

Table 6: The Notation of Evaluation Measure Used In the Classification

Predicted Class	**Actual Class**	
	AC_{ti}	AC_{fi}
PC_{ti}	T_{ti}	F_{ti}
PC_{fi}	F_{fi}	T_{fi}

	Precision	**Recall**
ti	$\frac{T_{ti}}{T_{ti}+F_{ti}}$	$\frac{T_{ti}}{T_{ti}+F_{fi}}$
fi	$\frac{T_{fi}}{T_{fi}+F_{fi}}$	$\frac{T_{fi}}{T_{fi}+F_{ti}}$

in the second, we only use the account-based features, and so on. We use Precision, Recall, and F-score as evaluated metrics in our rumor identifying task. In general, the above mentioned three measurements are used as metrics in information retrieval's performance evaluation. For instance, in the filed of information retrieval, the Precision is the proportion of retrieved documents which are relevant to user searched, the Recall is defined as number of relevant documents retrieved by a search divided by the total number of existing relevant documents, and the F-score is a trade-off between Precision and Recall.

While in the classification tasks, the terms true positives, true negatives, false positives, and false negatives compare the results of the SVM-classifier under test with Sina Weibo official judgements. The terms positive and negative refer to the SVM-classifier's classification (i.e. positive represents that the rumor-related microblog is classified into the non-rumor category, negative represents that the rumor-related microblog is classified into the rumor category), and the true and false terms refer to whether that classification corresponds to the judgement made by the Sina Weibo's rumor-busting service. The detail explanation of the measure methods are showed in Table 6.

In the table, "ti" represents the class of true information namely the orientation of the rumor related microblog is not consistent with the rumor, and "fi" represents the class of false information namely the orientation of the rumor related microblog is consistent with the rumor by which Sina Weibo rumor busting service is identified. In order to facilitate understanding, we use T_{ti} instead of the term true positive, F_{ti} instead of the term false positive, T_{fi} instead of term true negative, and use F_{fi} instead of term false negative. In addition, the *Actual Class* represents the Sina Weibo rumor-busting service's judgements, in which AC_{ti} (AC_{fi}) means that the microblog is verified as true (false) information by Sina or other already known facts. The *Predicted Class* represents the SVM-classifier's classification of the rumor-related microblogs, in which the PC_{ti} (PC_{fi}) represents that the microblog is predicted to be a non-rumor (rumor) by the SVM-Classifier. For example, T_{fi} means that the SVM classified one rumor-related microblog into the false information category as well as the Sina Weibo rumor-busting service makes the same judgement.

Table 7: The Evaluation Measure of Different Subsets of Features on the Classification Performance

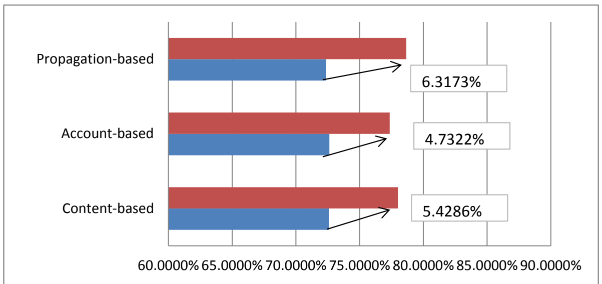
Content-based Feature			
Class	**Precision**	**Recall**	**F-score**
fi	0.5024	0.1697	0.2537
ti	0.7449	0.9660	0.9059
Account-based Feature			
fi	0.5000	0.3355	0.4016
ti	0.8783	0.9351	0.8293
Propagation-based Feature			
fi	0.5000	0.2059	0.2917
ti	0.7631	0.9254	0.8364

The Precision and Recall are defined as shown in Table 6. Because we do not consider the user's interest degree with Precision and Recall, therefore we use the traditional F-score namely the harmonic mean of precision and recall.

$$F = 2.\frac{precision.recall}{precision + recall}$$

The experimental results are shown in Table 7. The results indicate that among those features, the account-based features are good at detecting false information, and content-based features play an important role in detecting true information. We observe that using propagation-based features alone does not perform as well as using the other two subsets of features. This is because the corpus we crawled by the Sina Weibo API, the reforward relationship just contains two levels which are the original posted and the last retweeted.

As the features in the subset of content-based are related to the microblog content, hence is not effective to identify whether one microblog's message is false information just through analyzing the content. Most in the account-based features are user's attributes, so it is effective to detect the false rumors by microblog-account's features, like whether the user's account is verified, the number of its friends, the time span between its registering time and the posting time. For instance, if one who is verified by Sina Weibo and has a large number of friends (fans), then the microblogs posted by this account are rumors with a small probability. Contrary to the scenario, if one is just registering, with little friends (fans), default or fake avatar, and not verified by the official service, then the message posted by this account is false



	Content-based	Account-based	Propagation-based
🟥(+)Client (+)Location	78.0066%	77.3574%	78.6617%
🟦(-)Client (-)Location	72.5780%	72.6252%	72.3444%

Figure 4: The Effectiveness of New Features

rumor with high probability if this microblog related to a controversial event.

5.2 Impact of New Features

Before adding the two new features, the classifying accuracies of using content-based, account-based, and propagation-based features alone are 72.5780%, 72.6252%, and 72.3444% respectively.

We introduce the client program used feature and the event location feature into the features used for classification (which already consist of content-based features, account-based features, and propagation-based features) respectively to study their effectiveness. To illustrate the impact on classification accuracy, we show the results of adding the client program used feature and the event location feature. As shown in Figure 4, the classification accuracy is improved to varying degrees which are 5.4286%, 4.7322%, and 6.3173% with using the same SVM classifier and the same RBF kernel function ($\gamma = 0.313$), demonstrating the clear advantage of incorporating those two newly proposed features into the task of classification.

6. CONCLUSIONS

The vast volume of microblogs and the rapid propagation nature of the microblogging platforms make it critical to provide tools to automatically assess the credibility of microblogs. In this paper, we collect and annotate a set of rumor-related microblogs from Sina Weibo based on the information provided by Weibo's rumor-busting service. We propose two new features, namely the client program used and the event location, which can be extracted from the microblogs and used the classification of rumors. We show the effectiveness of those two features through extensive experiments.

7. ACKNOWLEDGEMENTS

This work was supported in part by National Natural Science Foundation of China Grants (No. 61070018, No. 60903108), the Program for New Century Excellent Talents in University (NCET-10-0532), an NSERC Discovery Grant, the Independent Innovation Foundation of Shandong University (2012ZD012, 2009TB016), and the SAICT Experts Program.

8. REFERENCES

[1] P. Bordia and N. DiFonzo. Problem solving in social interactions on the internet: Rumor as social cognition. *Social Psychology Quarterly*, 67(1):33–49, 2004.

[2] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[3] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW*, pages 675–684, 2011.

[4] B. D. Eugenio and M. Glass. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101, 2004.

[5] T. S. Ferguson. *A Course in Large Sample Theory*. Chapman and Hall, London, 1996.

[6] A. Hassan, V. Qazvinian, and D. R. Radev. What's with the attitude? identifying sentences with attitude in online discussions. In *EMNLP*, pages 1245–1255, 2010.

[7] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 international conference on Management of data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[8] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 71–79, New York, NY, USA, 2010. ACM.

[9] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, New York, NY, USA, 2012. ACM.

[10] W. Peterson and N. Gist. Rumor and public opinion. *American Journal of Sociology*, pages 159–167, 1951.

[11] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *EMNLP*, pages 1589–1599, 2011.

[12] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. *CoRR*, abs/1011.3768, 2010.