Mining Important Comments of Micro-Blog
Based on Feature Weighting

Jianfeng Fu*, Xujie Zhang and Ce Shi
College of Information Management, Shanghai Lixin University of Accounting and Finance, No. 995, Shangchuang Road, 201209 Shanghai, China

ABSTRACT

Important comments of micro-blog not only reflect the views of users but also can influence the public’s opinion towards a particular topic. This paper presents a method of feature weighting based k-Nearest Neighbors for mining important comments of hot topics on Sina Weibo. By using feature weighting method, each selected feature is assigned to a corresponding weight. The comprehensive experimental results demonstrate that the presented method can better predict important comments than traditional k-Nearest Neighbors method. Furthermore, we show the presented method significantly outperforms the state-of-the-art classifiers.

INTRODUCTION

Microblogging has become a very popular way for people to share thoughts, information, links, news and express their opinions. On Sina Weibo, one of the most popular microblogging services in China, more than 100 million messages are posted per day. The massive amount of microposts turned the microblogging website into a vast repository of electronic information resource for social science research. Sina Weibo provides reply and retweet functions for users commenting on a topic and expressing their sentiments or opinions. A comment can be not only just a word, an abbreviation or an emoticon, but also a detail, a discussion or an opinion to the topic. An important comment can even influence public opinions towards a hot topic of microblog [1].

Many research efforts have been conducted on mining comments of Microblog recently. Much of the work is mainly focused on sentiment analysis in microblogs [2-5]. However, comments are important supplementary information of microblogs. More and more researchers devote their attention to it. L. Zhang, et al. [6] develop a framework based on a heuristic–systematic model for comparing the effects of content and context of microblogging posts. As the result, the purpose of retweet is to disseminate information, whereas the comment emphasizes social interaction and conversation. A. Kothari, et al. [7] apply machine learning based classification approaches for identifying comments on specific news articles from twitter. Such comments are provided with news articles to improve reader experience.

In this paper, we focus on Sina Weibo and propose the task of mining important comments of hot topic. Such comments could be summarized and provided to readers for an overview of the public reaction towards the topic as quickly as possible. Given a set of hot topics and their comments, we apply text mining technique to identify which comments are important. Specifically, we present a method of feature weighting based
k-Nearest Neighbors (FWKNN for short) for mining important comments of hot topics on Sina Weibo. By using feature weighting method, each selected feature is assigned to a corresponding weight. Comparing with traditional k-Nearest Neighbors method, the FWKNN significantly improves the classification performance of the comments. Furthermore, it achieves better accuracy on predicting important comments than the state-of-the-art classifiers.

COMMENTS OF HOT TOPICS ON SINA WEIBO

On Sina Weibo, topic is preceded and followed by “#” since there is no delimiter to mark Chinese word or phrase boundaries. A comment is a message about a user’s response towards a micropost by using reply and retweet (Unlike Twitter, users are allowed to retweet messages with additional information on Sina Weibo) functions. If a topic is frequently discussed in a period of time, it will become hot and its comments will increase correspondingly.

TABLE I. AN EXAMPLE OF A HOT TOPIC AND THE COMMENTS.

| Topic | Comment | Comment type |
|-------|---------|--------------|
| #AirAsia# shares fall 11 percent after plane goes missing | @Reuters We are worrying over missing human beings, not wall street. | important |
| http://reut.rs/1AY8udN | @Reuters I imagine the relatives of the missing flight passengers will be crestfallen to hear that news... | important |
| “#AirAsia# shares fall 11 percent after plane goes missing” | // which one? “#AirAsia# shares fall 11 percent after plane goes missing” repost | unimportant |
| @Reuters My name is Sara. I live in China. I only know a few phrases in Chinese. I am American from the United States. I will be in China for the summer (May-August). I am looking for a Language-exchange partner. | unimportant |

The reactions of people toward a hot topic are different. This is also applied to their comments. A comment may express an opinion, a sentiment, a question, a wishing, an experience sharing or a call to action [7]. Besides, it may be a spam for advertising purpose, and has nothing to do with the topic. Here, we are concerned about whether a comment is important, as well as whether it can reflect public opinion or arouse public sympathy. Therefore, we divide the comments into two categories: important and unimportant. TABLE I shows an example of a hot topic along with its comments and the comment types on Sina Weibo.

IDENTIFYING IMPORTANT COMMENTS

In this part, we extract multiple features from Sina Weibo and propose the FWKNN method for identifying important comments.
Feature Description

To identify the important comments, the following features are selected in our method.

1) Bag-of-words features. We segment the Chinese microposts and its comments by ICTCLAS [8] system, remove stop words, and weight TF-IDF values of the terms. Intuitively, a repeating comment cannot be important because it has no an independent opinion. We can calculate the cosine similarity between the micropost and each comment. The greater the similarity of the two messages, the larger the possibility the comment is a repeat of the micropost.

2) Weibo-specific features. We use four features related to the Sina Weibo comment itself: Hashtag, Mention, Favorite and URL. As is mentioned before, hashtag is used to indicate the topic of Sina Weibo. If a comment contains hashtag, the corresponding feature will be assigned a value of 1; otherwise, a value of 0. Mention (@some_user) is used for replying other user on microblog. It may contain personal opinion towards a hot topic. Therefore, we use a binary feature to indicate whether the comment contains a mention or not. Favorite means other users agree with or like the comment. We use the number of favorites as a feature. URL in comment offers a link to an external resource. It may be an objective introduction about the hot topic. However, it also could be a spam. Hence, we use a feature to indicate whether the comment contains an URL or not.

3) Author features. Author’s information is also applied in our method to classify the important comments. The author features include the numbers of followers, friends and posts. The number of followers reflects the popularity of a comment author. We take the numbers of friends and posts of author into consideration because a spammer often has many friends and post many unsolicited or undesired messages.

4) Mark features. We use three binary features to indicate whether the comment contains non-lexical marks or not. The marks are (1) question mark, (2) exclamation mark, and (3) emoticon.

Feature Weighting

The k-Nearest Neighbors (KNN) algorithm is one of the most fundamental classification methods [9]. In traditional KNN, a hypothesis is implied that all the features contribute equally to the classifier while computing the similarity of two instances. However, the qualities of features are different [10]. An appropriate way is to assign a weight to each feature. Given a dataset \( X = \{ x_1, x_2, \ldots, x_n \} \), each instance is a point in a multidimensional space defined by a feature vector \( F = \{ f_1, f_2, \ldots, f_m \} \), and estimated by a feature weight vector \( W = \{ w_1, w_2, \ldots, w_m \} \). For two instances \( x_a \) and \( x_b \), their distance is defined by:

\[
d(x_a, x_b) = \left( \frac{\sum_{i=1}^{m} w_i \cdot \text{value}(f_i, x_a) - \text{value}(f_i, x_b))^2}{\sum_{i=1}^{m} w_i} \right)^{1/2}
\] (1)
where \( r = 2 \), \( \text{value}(f, x) \) is the \( i \)th feature of \( x \). We normalize the distance by dividing it with \( \sum_{i=1}^{m} w_i \) to ensure \( 0 \leq d(x, x_0) \leq 1 \).

Using heuristic measure to estimate the quality of the features has been widely studied. Previous work has demonstrated the outperformance of ReliefF algorithm [10]. We employ ReliefF to weight the selected features in this study. Suppose \( x \) is an instance randomly selected from dataset \( X \). We first find \( k \) nearest neighbors \( p = 1, \ldots, k \) from \( \text{class}(x) \), and also \( k \) nearest neighbors \( q = 1, \ldots, k \) from each class \( C \neq \text{class}(x) \). For each feature \( f \in F \), the corresponding weight \( w \) is defined by:

\[
w = w - \sum_{i=1}^{k} \text{diff} (f, x, p_i) + \sum_{C \neq \text{class}(x)} \left[ \frac{P(C)}{1 - P(\text{class}(x))} \sum_{i=1}^{k} \text{diff} (f, x, q_i) \right]
\]

(2)

where \( P(\text{class}(x)) \) is the probability of \( \text{class}(x) \). It can be estimated from training set. Function \( \text{diff} (f, x, p) \) is defined as:

\[
\text{diff} (f, x, p) = \frac{\text{value}(f, x) - \text{value}(f, p)}{\max(f) - \min(f)}
\]

(3)

For each feature \( f \), its corresponding weight \( w \) (initialize to 0) is calculated by using equation (2). The whole process is repeated until convergence.

**EXPERIMENTS AND RESULTS**

In this section, we first introduce the dataset collected from Sina Weibo. We then compare our proposed method with baseline and other state-of-the-art classifiers.

**Dataset**

We designed a web crawler to collect the hot topics and their comments from Sina Weibo. The collected messages were posted from May 1, 2014 to May 1, 2017. We select 200 hot topics and their 41708 comments from six categories (Emergency, Entertainment, Sports, Economy, Health and Education) of the collection for our experiments. Each comment is attached a label, important or unimportant, by annotators manually. The detail information of the annotated dataset is shown in TABLE II.

| Category   | Hot topics | Important comments | Unimportant comments |
|------------|------------|--------------------|---------------------|
| Emergency  | 40         | 3769               | 5245                |
| Entertainment | 40       | 3526               | 5014                |
| Sports     | 35         | 3112               | 4238                |
| Economy    | 30         | 2556               | 3471                |
| Health     | 30         | 2370               | 3289                |
| Education  | 25         | 2208               | 2910                |
Performance Comparison

We first conduct the experiments to evaluate the contribution of feature weighting by comparing the proposed FWKNN with traditional KNN (baseline). We then compare our method with two state-of-the-art classifiers, SVM and Maximum Entropy (ME). We use the libSVM [11] and Stanford ME classifiers with default settings. Several widely-used performance metrics are utilized to evaluate the classification tasks: precision (P), recall (R), F1 score and accuracy. We use 5-fold cross validation to evaluate the classification performance.

| Category   | FWKNN    | KNN(baseline) |
|------------|----------|---------------|
|            | P  | R  | F1 | P  | R  | F1  |
| Emergency  | 0.7616 | 0.7905 | **0.7758** | 0.6708 | 0.6915 | 0.6810 |
| Entertainment | 0.7547 | 0.7932 | **0.7735** | 0.6772 | 0.6938 | 0.6854 |
| Sports     | 0.7421 | 0.7846 | **0.7628** | 0.6415 | 0.6732 | 0.6570 |
| Economy    | 0.7437 | 0.7824 | **0.7626** | 0.6544 | 0.6689 | 0.6616 |
| Health     | 0.7302 | 0.7911 | **0.7594** | 0.6387 | 0.6786 | 0.6580 |
| Education  | 0.7398 | 0.7885 | **0.7634** | 0.6408 | 0.6812 | 0.6604 |

TABLE III. PERFORMANCE COMPARISON OF FWKNN AND KNN.

| Category   | FWKNN | SVM | ME |
|------------|-------|-----|----|
| Emergency  | 0.7913 | 0.7882 | 0.7835 |
| Entertainment | 0.7876 | 0.7807 | 0.7729 |
| Sports     | 0.7869 | 0.7783 | 0.7752 |
| Economy    | 0.7795 | 0.7605 | 0.7664 |
| Health     | 0.7583 | 0.7502 | 0.7443 |
| Education  | 0.7697 | 0.7581 | 0.7555 |
| Overall    | **0.7809** | 0.7720 | 0.7686 |

TABLE IV. CLASSIFICATION ACCURACY.

TABLE III presents the comparison results of FWKNN and KNN on mining important comments of hot topics on Sina Weibo. It can be seen that our proposed method achieves better classification performance on each category and make a significant improvement on F1 score. It demonstrates that the feature weighting efficiently enhances the classification performance of the comments. TABLE IV displays the comparison results of classification accuracy of FWKNN, SVM and ME. It can be seen that the presented FWKNN has the best overall accuracy at 0.7809, which achieves 0.0089 improvement than SVM and 0.0123 improvement than ME. Our method significantly outperforms SVM and ME on predicting important comments.

CONCLUSIONS

In this paper, we proposed a method of FWKNN for mining important comments on Sina Weibo. The approaches of feature selection and feature weighting were described in detail. Experimental results not only demonstrated the feature weighting improves classification performance on mining important comments of hot topics on Sina Weibo, but also showed that the FWKNN significantly outperforms the
state-of-the-art classifiers. Future work involves optimizing the feature weighting algorithm and summarizing hot topics based on important comments.

ACKNOWLEDGEMENT

This work is sponsored by Innovation Program of Shanghai Municipal Education Commission (14YZ151) and Natural Science Foundation of Shanghai (17ZR1419900). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

1. D. Li, X. Shuai, G. Sun, J. Tang, Y. Ding and Z. Luo. 2012. “Mining topic-level opinion influence in microblog,” Proc. of the 21st ACM international conference on Information and knowledge management, pp.1562-1566.
2. A. Bermingham, and A. F. Smeaton. 2010. “Classifying sentiment in microblogs: is brevity an advantage?” Proc. of the 19th ACM international conference on Information and knowledge management, pp.1833-1836.
3. E. Kouloumpis, T. Wilson, and J. Moore. 2011. “Twitter sentiment analysis: The good the bad and the omg!” Fifth International AAAI Conference on Weblogs and Social Media, pp.538-541.
4. A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau. 2011. “Sentiment analysis of twitter data,” Proc. of the Workshop on Languages in Social Media, pp. 30-38.
5. K. Ravi, and V. Ravi. 2015. “A survey on opinion mining and sentiment analysis: tasks, approaches and applications,” Knowledge-Based Systems, 89, pp.14-46.
6. L. Zhang, T. Peng, Y. Zhang, and X. Wang. 2015. “Content or context: Which carries more weight in predicting popularity of tweets in china,” Proc. of the 65th Annual Conference of World Association for Public Opinion Research, pp. 1-1.
7. A. Kothari, W. Magdy, K. Darwish, A. Mourad and A. Taei. 2013. “Detecting Comments on News Articles in Microblogs,” Proc. of the Seventh International AAAI Conference on Weblogs and Social Media, pp. 293-302.
8. H. P. Zhang, H. K. Yu, D. Y. Xiong and Q. Liu. 2003. “HHMM-based Chinese lexical analyzer ICTCLAS,” Proc. of the second SIGHAN workshop on Chinese language processing, pp. 184-187.
9. J. M. Keller, M. R. Gray, and J. A. Givens. 1985. “A fuzzy k-nearest neighbor algorithm,” IEEE Transactions on Systems, Man and Cybernetics, 15(4), pp. 580-585.
10. Y. Sun. 2007. “Iterative RELIEF for feature weighting: Algorithms, theories, and applications,” IEEE transactions on pattern analysis and machine intelligence, 29(6), pp.1035-1051.
11. C. Chih-Chung, and L. Chih-Jen. 2011. “LIBSVM: a library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, 2(3), pp. 1-27.