Advances in Computer Science Research, volume 80
International Conference on Computer Science, Electronics and Communication Engineering (CSECE 2018)

Sentiment Analysis of Emergencies Based on Microblogging

Changjin Liu[1,2], Xin Ye[1,*], Hongxia Dai[1], Fan Chen[3] and Luan Dong[1]
[1]Faculty of Management and Economics, Dalian University of Technology, No.2 Linggong Road, Dalian 116024, China
[2]Tsinghua-Berkeley Shenzhen Institute, University Town of Shenzhen, Nanshan District, Shenzhen,518055,China
[3]Transportation Management College, Dalian Maritime University, NO.1 LingHai Road, Dalian 116026, China
[*]Corresponding author

***Abstract*—With wide use of microblogging, the sentiment analysis of emergencies based on microblogging is helpful to analyze the trend of public opinion, and is beneficial to monitor and guide the public opinion correctly for government. Firstly, word2vec is used to transform microblogging text into the feature vector with high-dimensional space. Then, a classification algorithm based on random forest optimized by genetic algorithm is proposed. Finally, an experiment is performed. The result shows that the accuracy of microblogging sentiment classification based on proposed classification algorithm is improved greatly compared with classical single classifiers.**

Keywords—emergencies; sentiment analysis; Word2vec; random forest; genetic algorithm

I. INTRODUCTION

With the rapid development of Web technology, people use Internet forums, blogs, microblogging and other ways to express their views and feelings about social emergency. However, when the public opinions pose a threat to social order and have a bad influence on people's life, Internet emergency happens [1]. The platform like microblogging contains a large amount of opinion-rich data, which reflects the viewpoint of people targeting various types of Internet emergencies in real-time [2]. The emotion of people often affects the development of the whole event. Therefore, it is of great value to do research on public opinion of emergencies network, especially the sentiment analysis of emergencies based on microblogging. And it is helpful to monitor and guide the public opinion correctly for government.

In recent years, sentiment analysis in the field of emergencies has attracted a large number of researchers to study. In this part, we give a brief introduction to the previous work on the methods for sentiment analysis in the field of emergencies. Tong LI proposed a sentiment analysis and prediction model based on multiple model integration, and use this model to achieve the purpose of sentiment classification and trend prediction of public opinion [3]. Jing WANG considered the difference of evaluation objects, used a variety of feature selection methods and machine learning method to analyze the sentiment [4]. Guolan CHEN proposed a sentiment analysis method based on the combination of sentiment lexicon and semantic rules, and calculated the sentiment tendencies of the whole topic based on the influence of microblogging users [5]. Yuanyuan LI proposed a multi-feature combination method based on emotion dictionary and CRF model to analyze the sentiment [6]. Shihai TIAN proposed to improve the potential semantic analysis and support vector machine algorithm for sentiment classification, based on mining public opinion information through meta-search technology, and increasing the baseline offset value to optimize the emotional feature orientation weight [7].

At present, scholars put forward a variety of methods in the study of sentiment analysis of emergencies [8-10]. The commonly methods are based on the traditional lexicon [11-13] or based on machine learning such as SVM [14-16], and random forest algorithm is rarely used. Random forest algorithm is the integration of multiple single decision tree classifiers, which can improve the generalization ability of classifiers [17]. Furthermore, the combination of random forest and genetic algorithm can greatly improve the accuracy of the sentiment classification [18]. Therefore, aiming at the sentiment analysis of emergencies, the classification algorithm based on random forest optimized by genetic algorithm(GA-RF) is proposed in this paper.

The rest of this paper is organized as follows. Section 1 details our techniques and proposes the classification algorithm based on random forest optimized by genetic algorithm(GA-RF). Then, Section 2 performs an experiment and analyzes the results. Finally, Section 3 concludes this paper.

II. PROPOSED METHOD

A. Overview

The general framework of this paper is displayed in Figure 1. Firstly, the comments data are extracted from microblogging. Secondly, these data are preprocessed. And next, the comments data are transformed into vectors based on word2vec model. Finally the classification algorithm based on random forest optimized by genetic algorithm (GA-RF) is used to classify comments data into two sentiment polarities (positive or negative).

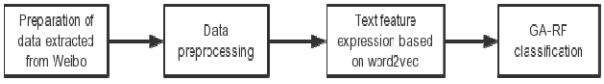

FIGURE I. THE GENERAL FRAMEWORK OF PAPER

B. Preprocessing

1) Data collection: There are more than 5000 comments data are extracted from the search API of SinaWeibo.

2) Data preprocessing: This step mainly consists of noise reducing, text segmenting, and stopwords deleting three parts.

*a) Noise reducing: R*egular expressions are used to remove URL links, microblogging topic tags, location information, microblogging forward signs, and microblogging emoticons.

b) Text segmenting: Jieba word segment tool is adopted to segment the comments text into words and text.

c) *Stopwords deleting:* The natural language processing stopwords provided on CSDN.is used to remove useless words.

C. Vector Representation Based on Word2vec

Word2vec is a tool released by Google in 2013. This tool can find the semantic relationships between words in the document and vectorize text [19-20]. Word2vec adopts two main model architectures, continuous bag-of-words(CBOW) model and continuous skip-gram model [19].

In this paper, skip-gram model is used to train the training file. Firstly, Chinese Wikipedia corpus are trained to get the word2vec model. Then, the trained model is applied to represent the comments data as 400 dimensional vectors, which are marked with sentiment polarities.

D. Classification algorithm based on GA-RF

Integrated learning is a methodology that combines multiple classifiers, and can achieve better performance than a single classifier. In order to improve the accuracy of microblogging sentiment classification, a new integrated learning algorithm based on random forest is proposed. This algorithm use genetic algorithm to optimize random forest for selecting the best performing decision tree.

The process of classification algorithm based on random forest optimized by genetic algorithm is represented in Figure 2.

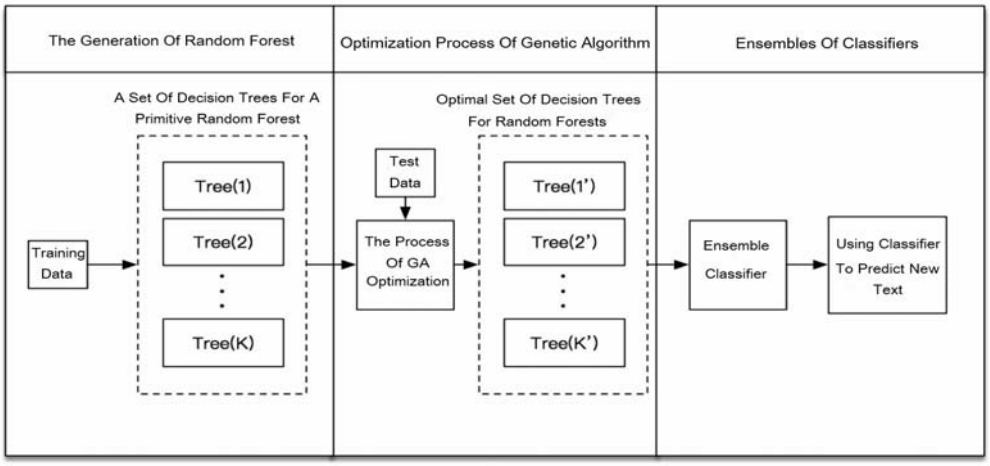

FIGURE II. THE PROCESS OF CLASSIFICATION ALGORITHM BASED ON GENETIC ALGORITHM FOR RANDOM FOREST OPTIMIZATION

Where, K is the total number of decision trees in trained random forest, and K' is the number of new decision trees after the selection of genetic algorithm, and $K' < K$.

In genetic algorithm, fitness function directly reflects the target optimization function of algorithm. In this paper, fitness function is defined as the accuracy of classifiers. In order to make the crossover and mutation directly affect the independent variables of the objective function, 0 and 1 are used to indicate whether the N-th tree of a trained random forest is selected. For example, there is a random forest that has been trained with 10 decision trees, 0 indicates that the N-th tree is not selected, while 1 is selected. Assuming that the initial selection is 2-th, 3-th, 5-th and 8-th trees. Table 1 gives an example of chromosome coding.

TABLE I. THE EXAMPLE OF CHROMOSOME CODING

Tree number	1	2	3	4	5	6	7	8	9	10
Binary coding	0	1	1	0	1	0	0	1	0	0

In genetic algorithm, there are three kinds of operators that control the genetic and replication process of chromosomes. In this paper, binary tournament selection is selected as the selection operator, and single-point crossover is selected as the crossover operator. A variation gene is selected randomly in a chromosome, and the value is transformed between 0 and 1.

III. EXPERIMENT AND DISCUSSION

A. Data Sets

The comments data of four hot topics in 2016 are selected. And then, the AHP method is used to label 3000 comments data manually. A brief summary of data sets is represented in Table 2. To conduct experiments, data sets is divided into two parts, one part with 1350 positive comments and 1350 negative, is used as train data. The other part is used as test data.

TABLE II. A BRIEF SUMMARY OF DATA SETS

Data Set	**Positive**	**Negative**	**Total**
Train Data	1350	1350	2700
Test Data	150	150	300

B. Experimental Parameters

In this section, the value of experimental parameters for three classic classification algorithms and GA-RF are set. The value of parameters are mainly depended on the best results of some experiments and the experience. The value of specific parameters are listed in Table 3.

TABLE III. THE VALUE OF PARAMETERS

Method	Parameters set
SVM	SVM: C-SVC; Kernel function: RBF;
Random Forest	Random forest scale: 100;
KNN	Number of neighbours: 5
Genetic Algorithm	Population number: 260; Crossover and Mutation rate: 0.8; Gene mutation rate: 0.1; Number of reproduction: 500;

Results of experiments are evaluated with accuracy, precision and recall. The accuracy is used as the core evaluation index, and the precision and recall is used as the secondary evaluation index.

C. Performance of Sentiment Classification

The performance of three classification algorithm (SVM, random forest and KNN) is listed in Table 4. A comparative analysis based on three classification algorithm is given below. Firstly, the precision of SVM is very low, while the recall rate is high. This performance indicates that the classification accuracy of the samples, which divided into positive emotions by the SVM algorithm is very high, but not all positive emotion samples are correctly classified. On the contrary, the SVM algorithm is very likely to ignore the positive emotion samples. Secondly, compared with SVM, the precision of random forest is higher and recall is lower. This shows that random forest performs better for text marked as positive emotion. Thirdly, KNN shows the worst performance among three classification algorithm, indicating that KNN is unsuitable for emotion classification in this particular issue.

Then, the algorithm based on random forest optimized by genetic algorithm model is adopted. Similarly, 10-fold cross is used and calculate the average of accuracy. The experimental results are listed in Table 5. As expected, GA-RF is significantly better than other algorithms.

TABLE IV. THE PERFORMANCE OF THREE CLASSIFICATION ALGORITHM (SVM, RANDOM FOREST AND KNN)

	SVM			**Random Forest**			**KNN**		
	Accuracy	***Precision***	***Recall***	***Accuracy***	***Precision***	***Recall***	***Accuracy***	***Precision***	***Recall***
1	0.6367	0.3985	0.6647	0.6133	0.4060	0.6367	0.5567	0.3008	0.4431
2	0.59	0.4311	0.6165	0.6067	0.4251	0.6466	0.5867	0.2994	0.4436
3	0.59	0.4236	0.6603	0.6267	0.4091	0.6644	0.6167	0.2597	0.4863
4	0.62	0.4236	0.6603	0.6367	0.3611	0.6346	0.5367	0.3333	0.4167
5	0.6367	0.3958	0.6667	0.5833	0.4792	0.6410	0.5967	0.2778	0.4808
6	0.65	0.3846	0.6815	0.6033	0.4266	0.6306	0.5867	0.3217	0.5032
7	0.6667	0.3241	0.6581	0.6333	0.3793	0.6452	0.57	0.2897	0.4387
8	0.6267	0.4028	0.6538	0.6167	0.3889	0.6218	0.56	0.2708	0.4038
9	0.5933	0.4765	0.6846	0.5833	0.5	0.6923	0.5933	0.3059	0.4615
10	0.6333	0.4286	0.6928	0.6133	0.4014	0.6275	0.5667	0.3537	0.4902
AVG	0.6243	0.4089	0.6639	0.6133	0.4177	0.6433	0.5667	0.3013	0.4568

TABLE V. THE EXPERIMENTAL RESULT OF GA-RF

Fold	1	2	3	4	5	6	7	8	9	10	AVG
Accuracy	0.7667	0.7612	0.7600	0.7846	0.7652	0.7824	0.7935	0.7658	0.7846	0.7741	0.7738

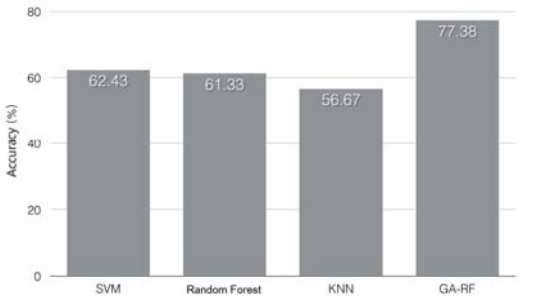

FIGURE III. THE COMPARISONS OF LASSIFICATION AVERAGE ACCURACY

Figure 3 shows the comparisons of average accuracy from different classification algorithm. The classification accuracy of GA-RF is the highest, and is 23.9%, 26.2% and 36.5% higher than SVM, random forest and KNN respectively. SVM has a second high accuracy, which is 1.8% and 10.2% higher than the random forest and KNN.

IV. CONCLUSION

In this paper, the sentiment analysis of emergencies based on microblogging is studied. After word2vec is used to transform text into the feature vector, the classification algorithm based on random forest optimized by genetic algorithm is proposed. The experimental results show that the

method gives better results than single classifiers. However, compared with other studies, the accuracy is still need to be improved.

In the future, there is a need to apply the same method to explore the change of public opinion during the life cycle of emergency. Try to find the changing characteristics of public opinion during the evolution of events, and find the impact of government measures on public opinions.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 71533001 and in part by the Liaoning province economic and social development project of Liaoning Provincial Federation Social Science Circles of China under Grant 2017lslktyb-044.

REFERENCES

[1] Dehai LIU, Weiguo WANG, Hongyi LI. Evolutionary Mechanism and Information Supervision of Public Opinions in Internet Emergency [J]. Procedia Computer Science, 2013, 17:973-980.

[2] Faliang HUAN, Shichao ZHANG, Jilian ZHANG, Ge YU. Multimodal Learning for Topic Sentiment Analysis in Microblogging[J]. Neurocomputing, 2017, 253(C):144-153.

[3] Tong LI, Zhijie SONG. Public sentiment of emergent events based on model integration and its trend prediction[J]. Systems Engineering — Theory & Practice, 2015, 35(10):2582-2587.

[4] Jing WANG. Analysis on Sentiment Orientation and Its Evolution of Emergency Network Public Opinion, 2012.

[5] Guolan CHEN. Monitoring Based on The Micro-blog Burst Incidents Word And Sentiment Analysis [J]. Nanjing University Of Posts And Telecommunications, 2015.

[6] Yuanyuan LI. Research on Hot Event Sentiment Analysis for Topic Microblog [J]. Anhui University, 2016.

[7] Shihai TIAN, Deli LYU. An Early Warning Algorithm for Public Opinion of Safety Emergency[J]. Data Analysis and Knowledge Discovery, 2017, 1(2):11-18.

[8] Huifeng TANG, Songbo TAN, Xueqi CHENG. A survey on sentiment detection of reviews[J]. Expert Systems with Applications, 2009, 36(7):10760-10773.

[9] Bing LIU. Sentiment Analysis and Opinion Mining[M]. Morgan & Claypool Publishers, 2012.

[10] Tsytsarau M, Palpanas T. Survey on mining subjective data on the web[M]. Kluwer Academic Publishers, 2012.

[11] Minqing HU, Bing LIU. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 168-177.

[12] Kim S M, Hovy E. Identifying and analyzing judgment opinions[C]//Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics, 2006: 200-207.

[13] Augustyniak L, Kajdanowicz T, Szymanski P, et al. Simpler is better? Lexicon-based ensemble sentiment classification beats supervised methods[C]//Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. IEEE, 2014: 924-929.

[14] Shoushan LI, Zhongqing WANG, Guodong ZHOU, Sophia Yat Mei Lee. Semi-Supervised Learning for Imbalanced Sentiment Classification.[C]// Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Springer-Verlag, 2009:588-595.

[15] Rui XIA, Tao WANG, Xuelei HU, Chengqing ZONG. Dual Training and Dual Prediction for Polarity Classification[C]// Meeting of the Association for Computational Linguistics. 2013:521-525.

[16] Hongwei WANG, Lijuan ZHENG. Sentiment classification of Chinese online reviews: a comparison of factors influencing performances[J]. Enterprise Information Systems, 2016, 10(2):228-244..

[17] Abellán J, Mantas C J, Castellano J G. A Random Forest approach using imprecise probabilities[J]. Knowledge-Based Systems, 2017.

[18] Dao S D, Abhary K, Marian R. An innovative framework for designing genetic algorithm structures[J]. Expert Systems with Applications, 2017, 90:196-208.

[19] Tomas Mikolov, Kai CHEN, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[20] Tomas Mikolov, Ilya Sutskever, Kai CHEN, Greg Corrado, Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.