International Journal of Engineering & Technology, 7 (4.4) (2018) 40-41

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Application of Machine Learning Techniques to Tweet Polarity Classification with News Topic Analysis

Hoyeon Park1, Hyeonjeong Seo1, Kyoung-jae Kim2*, Gundoo Moon1

[1]*Department of MIS, Graduate School, DonggukUniversity_Seoul, South Korea*
[2]*Department of MIS, Business School, DonggukUniversity_Seoul, South Korea*

Abstract

The exponential growth of online community provides the tremendous amount of textual information in terms of human behavioral reaction. Thus, online social media platforms such as Twitters, Facebook and YouTube are reflected as an essential part of human relationship networks. Especially, Twitter is widely applied to the disaster situation as a text and it provides critical insights into emergency management. In this study, we propose a topic analysis and sentiment polarity classification with machine learning techniques for emergency management. In this study, we compared the polarity classification models using three machine learning methods and found that the model with random forests showed the best classification performance.

Keywords:Polarity classification, Topic analysis, Machine learning.

1. Introduction

Nowadays, Twitter is a powerful information media when a disaster occurs because many people's local groups are trying to communicate by using tweets for natural disasters. However, Twitter is basically a social networks services, do not necessarily include disaster information. This study aims to predict words related to disaster situations using online news and tweets data. In this study, we propose a model that uses topic analysis and polarity analysis in the process of information utilization in disaster situations such as a hurricane and test which machine learning techniques are suitable for predicting which terms are useful for polarity analysis. We collected and analyzed the dataset of Hurricane Harvey, 2017, and the experimental results are promising.

2. Related works

2.1. Topic Analysis

Text mining is a useful tool for dealing with a large amount of text data for making patterns. Among text mining techniques, topic analysis has the advantage of being able to analyze the subject and trend when text-based documents exist. In this paper, we will use LDA (Latent Dirichlet Allocation), which is the most widely used topic modeling algorithm (Blei, 2003). This model is possible to understand the whole topic by distribution frequency(O'Connor et al., 2010).

2.2. Sentiment Analysis

Sentiment analysis refers to systematically identifying, extracting, quantifying, and studying emotional states and subjective information using natural language processing and text analysis. This paper focuses on the polarity classification of sentiment analysis. The polarity classification of opinions is represented as an object features of opinion texts (Liu et al., 2003), and classifythem as positive and negative values (Bollen et al., 2012).

3. Experiments & Results

3.1. Experiments

We conduct this research through the research process shown in Figure 1 in this study.

This study uses R and API development tool for collecting news articles about Hurricane Harvey from August 17 to August 31 using comScore's "Top 100 Online News 2017" news site. Experimental process of cleaning contents is carried out by moving unnecessary data from news data and tweets such as tags, emoticons, whitespace, numbers, URLs, stop words, punctuations. After preprocessing, there are 20,900 tweets and 9,107 news articles. Topic models analyze news articles to create DTM (Document-Term Matrix). Through the DTM, we identify the positive and negative parts of the news, where K = 30. Figure 2 shows a positive and negative word-cloud of the news article related toHurricane Harvey.

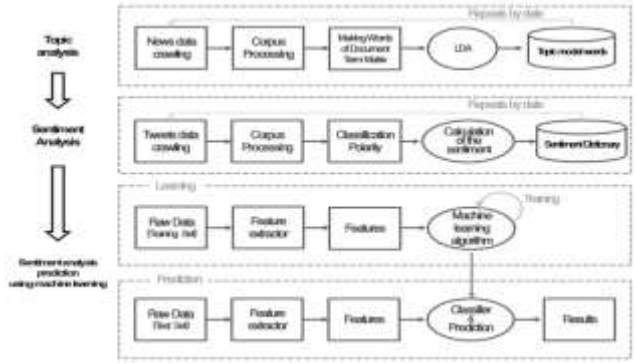

Figure 1: - Research Process

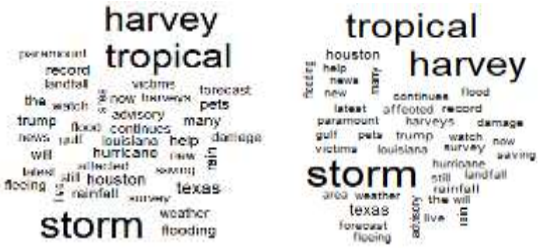

Figure 2: – positive & negative word-cloud of Hurricane Harvey

Topic modeling is implemented using news articles and a dictionary is constructed. Then, polarity classification is performed using the constructed dictionary. The polarity of the information attribute (prediction information) of the tweetsis judged to be 0 and 1.

In this research, three machine learning methods including SVM(Support Vector Machine), Random Forest and Naïve Bayes are used for classifying sentiment analysis outcomes. This experimental data set is divided into 70% of the training set and 30% of the test set.

3.2. Results

According to Table 1, we can see that the experimental results. It comes from a result of polarity classification based on the total DTM. The number in this table shows the classification accuracy of each model.

Table 1:. Classification Accuracies of three models

Random Forest		Actual	
		Negative	Positive
Predicted	Negative	87.8%	14.9%
	Positive	12.2%	85.1%
Average classification accuracy: 86.5%			
Naïve Bayes		Actual	
		Negative	Positive
Predicted	Negative	82.16%	14.9%
	Positive	17.84%	85.1%
Average classification accuracy: 83.6%			
SVM		Actual	
		Negative	Positive
Predicted	Negative	88.23%	19.9%
	Positive	11.77%	80.1%
Average classification accuracy: 84.2%			

The results show that random forests have the best performance for the three classification techniques in polarity classification.

4. Conclusion

Most of the social network services are based on everyday life, and thus various pattern analysis using texts of social network services is possible. This study aimed to construct information through SNS and to test which methods should be used according to the emergency situation to obtain the best classification accuracy. As a result, this study showed the best predictive performance when the Random Forests algorithm are employed. Since there is no way to perform real-time analysis in polarity classification, a method through real-time machine learning techniques is needed for future research issues.

References

[1] F. A. Pozzi, E. Fersini, E. Messina, & B. Liu, Sentiment Analysis in Social Networks, Morgan Kaufmann, 2016.
[2] D. M. Blei, A. Y. Ng, & M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research. 3 (2003), 993-1022.
[3] J. Bollen, H. Mao, & X. Zeng, Twitter mood predicts the stock market, Journal of Computational Science. 2 (2011), 1-8.
[4] B. Liu, Y. Dai, X. Li, W. S. Lee, & P. S. Yu, Building text classifiers using positive and unlabeled examples, Proceedings of Third IEEE International Conference on Data Mining. (2003), 179–186.
[5] B. O'Connor, R. Balasubramanyan, B. R. Routledge, & N. A. Smith, From tweets to polls: Linking text sentiment to public opinion time series. Proceedings of the Fourth International Conference on Weblogs and Social Media, (2010), 122–129.