Special issue of the 3rd International Conference on Computational and Experimental Science and Engineering (ICCESEN 2016)

Sentiment Analysis: an Application to Anadolu University

Z. Kamisli Ozturk*, Z.İ. Erzurum Cicek and Z. Ergul
Anadolu University, Industrial Engineering Department, Eskisehir, Turkey

Social media is a Web 2.0 platform that allows to share content and information without the limitations of time and space. Social media networks have managed to become a part of today's lifestyle and are increasingly gaining importance when viewed from a state perspective. Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. In this study, we focus on social media mining and sentiment analysis for students of an open and distance education system. Anadolu University which has approximately two million students and more than two million graduates, is a well-known institution in Turkey, that offers higher education through contemporary distance education model. Firstly, we have fetched Tweets related to Anadolu University open and distance education system. To perform sentiment analysis, these tweets were analysed by statistical and data mining techniques. Finally, results were visualized.

DOI: 10.12693/APhysPolA.132.753

PACS/topics: Sentiment Analysis, Twitter, Classification, Open and Distance Education System

1. Introduction

Social media is a Web 2.0 platform that allows to share content and information without the limitations of time and space. As one of the most known social media channels, Twitter is a free social networking microblogging service that allows registered members to broadcast short posts called tweets. Tweets are 140-charter little sentences and the users have to compress their feelings into this little space. Twitter members can broadcast tweets and follow other users' tweets by using multiple platforms and devices. The default settings for Twitter are public.

Sentiment analysis (SA) identifies the sentiment expressed in a text. The target of SA is to find opinions, identify the sentiments they express, and then classify their polarity [1]. Sentiment classification could be done on the word/phrase level, sentence level and document level [2].

The related studies on social media mining topic have been increasing since 2010. Asur and Huberman [3] used Twitter content to forecast the box-office revenues for movies and made sentiment analysis to investigate how efficient the attention can be for predicting opening weekend box-office values for movies. Karcı and Boy [4] studied the analysis of the social media using web mining techniques and presented an application for prediction of similarity with common attribution analysis using web structure mining. Choy et al. [5] used SA to estimate the votes in Singapore presidential election 2011. Wang et al. [6] used Naïve Bayes classifier (NBC) for SA in their study of 2012 U.S. presidential election cycle. Meral and Diri [7] used NBC, random forest and support vector machine to classify Twitter data through SA process and presented the comparison of the three classification methods. Ceron et al. [8] made a SA study on electoral campaigns and have shown that Twitter has a remarkable ability to forecast electoral results. Shang et al. [9] used Facebook to find regions of interest with respect to user's current geographic location. To understand the trends and public opinion toward massive online open courses, Shen and Kuo [10] conducted a SA study using a tool called OpinionFinder. A detailed survey about new methodologies for social big data, social data analysis, providing also social-based applications, are presented by Orgaz et al. [11]. Pandarachalil et al. [12] presented an unsupervised method for SA of Twitter data. To predict American presidential elections, Shei et al. [13] analyzed location-based Twitter data with SA in the first stage and proposed a feature model in the second stage. Besides these studies, Ofek et al. [14], Xu et al. [15] and Xia et al. [16] can be given as examples for other SA studies in different fields.

According to literature review, it can be said that there is not enough study about the SA in distance education systems. Literature about SA for distance education is contributed with this study. In the second section, basic model and process of SA and an application will be explained in detail. The conclusions and future work plans will be presented in the third section.

2. Problem definition and methodology

The basic SA model consists of the collection, analysis and pre-processing of data, sentiment classification and visualization. The SA model will be explained through the application to Anadolu University open and distance education (ODE) system.

Anadolu University ODE is the second mega-university in the world. Today, the total number of students of the three distance education faculties is over 2 million. Since it is very hard to service the millions of the students

*corresponding author; e-mail: zkamisli@anadolu.edu.tr

excellently, Anadolu University aims to overcome the problems immediately and provide ever-growing education. To identify the problems and present new education platforms to ODE students, ODE develops interactive systems based on student communication, like forums, online courses, etc. These systems are under control of the institution, so that the students can not share their feelings, opinions and complaints freely. In this situation, social media sharing of students can be the best source to reach students' real opinions.

2.1. Data source and data collection

To understand students' sentiments toward Anadolu University on social media, Twitter was considered. Tweets were collected using Twitter API called Twython between 1^{st} and 15^{th} of June. These dates involve examination days and the grades declaration day. We have used some query terms that refer to the ODE system such as "aof", "aöf", "acikogretim", "açıköğretim" to collect data.

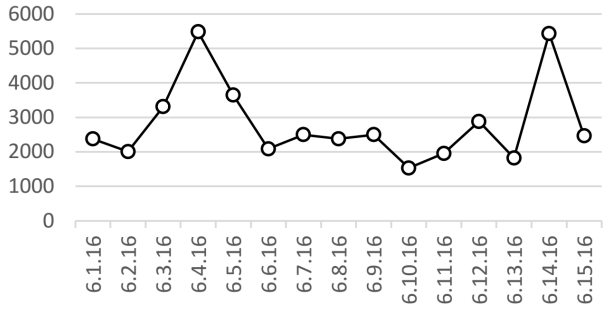

Fig. 1. Number of tweets.

After data collection, the distribution of the tweets day by day was investigated. It is obviously seen that the first day of the final examination and the declaration day are the most active days in this time interval. Figure 1 gives the number of the tweets on a daily basis.

2.2. Pre-processing phase

Pre-processing is necessary in order to achieve good-quality results [17]. The raw data collected from Twitter contains a lot of data types that are not related to our issue, like duplicated tweets, tweets in foreign languages, unrelated links and also advertisements. To get rid of these kinds of data, a two-stage pre-processing operation was performed.

At the first stage, tweets which were duplicated or include unrelated links or advertisements were removed. After that, to determine the languages of tweets, a language detection web service, named "Language Detection API", was used. Language detection API is a free API that allows up to 5000 requests per day and detects 160 different languages. After detection operation with Python, 103 foreign languages were detected, such as Thai, English, Korean, etc. For the second phase of pre-processing, all tweets, which were not including some Turkish words were removed permanently. At the end of this phase, the number of tweets was reduced from 63.699 to 4.652 tweets. Given the decrease in data count, the importance of data cleaning has arisen.

2.3. Classification phase

Actually, SA is a classification process. In the study, the collected tweets were aimed to be classified into three sentiment classes: positive, negative and neutral.

To make the classification, sentence-level SA methodology was followed. The sentence-level SA takes into account the whole sentence and tries to classify it. For the classification operation NBC was used, because of the common usage in text mining operations. NBC was also used in [7, 18, 19] as the method of SA studies.

The NBCs are known as a simple Bayesian classification algorithm. They provide a more straightforward interpretation of predictions and have been employed in many areas and have been proven very effective for text categorization [20, 21]. NBC classifies the sentences using the frequencies of the words in each sentence.

To code NBC, Python Natural Language Toolkit (NLTK) was used. NLTK is a leading platform for building Python programs to work with human language data. NLTK is a free, open source, community-driven project [22].

The processed data were divided into three parts, the training, the test and the validation datasets, for classical classification process. As given by Pandarachalil et al. [12], machine learning approaches require a huge amount of labeled training data to achieve desired accuracy. In twitter domain, this is not always practical. So, in this study a hundred tweets for each sentiment, which were collected during the first week, have formed the training datasets. For the test set, 3.952 tweets were collected from the second week tweets.

Some common words are determined as taboo words. Those are prepositions, pronouns, conjunctions and query terms, which were not included into classification. The result of the test phase is summarized in Table I.

TABLE I

Results of classification of test dataset.

Class	Classified data
Negative	1628
Neutral	1847
Positive	477
Total	3952

TABLE II

Results of classification for validation dataset.

Class	Actual	Correctly classified	Incorrectly classified	Success rate [%]
Negative	180	116	85	64.44
Neutral	202	97	38	48.2
Positive	18	11	53	61.11
Total	400	224	176	56.00

Unfortunately, this procedure had not provided an expressive result about the classifying success. We have

conducted a validation phase to confirm the classification. Four hundred tweets were used for validation and the classification results of validation dataset are given in Table II. As shown in Table II, the classification success rate for whole validation dataset is at an average value.

2.4. Visualization

In the last step of SA process, it is necessary to visualize data and the results. We use the word clouds for visualization. Word cloud is a visualization tool that provides words used more frequently in a dataset to appear bigger and sometimes bolder. The word clouds also visualize the text-based data in an esthetical and plain way. There are a lot of word cloud free generators available in the Internet. The word clouds can also be generated by R programming. We have generated three word clouds for each sentiment class with validation dataset. Frequently used words can be seen for each word cloud, given in Figure 2.

Fig. 2. Word clouds of validation.

3. Conclusions and future plans

In open education systems, the core issue should be the satisfying of student preferences and needs. With this study, the obtained negative feelings of the students will steer the education system of ODE. The managers of the institution can concentrate on the shortcomings of the system and student complaints by using the outcome of this study.

Using the sentence-level SA and NBC with limited data provides an average success rate. To increase the success of the analysis, lexicon-based sentiment model will be developed. Because Turkish sentiment lexicon does not exist, as the first step Turkish sentiment lexicon for ODE system will be formed. The classification method and parameters are some of the factors influencing the SA success rate. With the lexicon based sentiment model, different classification methods will be experienced. Moreover, the sizes of the training, test and validation datasets should be increased. With an automated collection system, the data collection operation will be made continuous. This system will also provide increase in dataset size.

Acknowledgments

This study is supported by Anadolu University Scientific Research Projects Committee (AUBAP-1604E173).

References

[1] W. Medhat, A. Hassan, H. Korasy, *Ain Shams Eng. J.* **5**, 1093 (2014).

[2] V. Jagtap, H. Pawar, *Int. J. Scientif. Eng. Technol.* **2**, 164 (2013).

[3] S. Asur, B.A. Huberman, in: *WI-IAT '10 Proc. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010, p. 492.

[4] A. Karcı, O. Boy, *Sosyal Ağlarin Web Madenciliği Teknikleri ile Analizive Ortak Atıf Analizi ile Benzerlik Tahmini*, in: *Elektrik-Elektronik ve Bilgisayar Sempozyumu*, 2011.

[5] M.J. Choy, M.L.F. Cheong, N.M. Ma, P.S. Koo, arXiv:1108.5520, 2011.

[6] H. Wang, D. Can, A. Kazemzadeh, F. Bar, S. Narayanan, in: *ACL '12 Proc. ACL 2012 System Demonstrations*, 2012, p. 115.

[7] M. Meral, B. Diri, *Sentiment Analysis on Twitter*, in: *2014 IEEE 22nd Signal Processing and Communications Applications Conf*, 2014.

[8] A. Ceron, L. Curini, S.M. Iacus, *Soc. Sci. Comput. Rev.* **33**, 3 (2015).

[9] S. Shang, D. Guo, J. Liu, K. Zheng, J. Wen, *Neurocomputing* **173**, 118 (2016).

[10] C. Shen, C. Kuo, *Comput. Human Behav.* **51**, 568 (2015).

[11] G.B. Orgaz, J.J. Jung, D. Camacho, *Inform. Fusion* **28**, 45 (2016).

[12] R. Pandarachalil, S. Sendhilkumar, G.S. Mahalakshmi, *Cognitive Computat.* **7**, 254 (2015).

[13] L. Shi, N. Agarwal, A. Agarwal, R. Garg, J. Spoelstra, *Predicting US Primary Elections with Twitter (2012)*, paper presented at *The workshop social network and social media analysis: methods, models and applications (NIPS)*.

[14] N. Ofek, S. Poria, L. Rokach, E. Cambria, A. Hussain, A. Shabtai, *Cognitive Computat.* **8**, 467 (2016).

[15] R. Xu, T. Chen, Y. Xia, Q. Lu, B. Liu, X. Wang, *Cognitive Computat.* **7**, 226 (2015).

[16] D.R. Recupero, V. Presutti, S. Consoli, A. Gangemi, A.G. Nuzzolese, *Cognitive Computat.* **7**, 211 (2015).

[17] F.H. Khan, U. Qamar, S. Bashir, *Cognitive Computat.* **8**, 614 (2016).

[18] A. Tripathy, A. Agrawal, S.K. Rath, *Procedia Comput. Sci.* **57**, 821 (2015).

[19] A. Go, R. Bhayani, L. Huang, *Twitter Sentiment Classification using Distant Supervision, Technical Report*, Stanford 2009.

[20] W. Dai, G. Xue, Q. Yang, Y. Yu, in: *AAAI'07 Proc. 22nd National Conference on Artificial Intelligence*, 2007, p. 540.

[21] M. Cevri, D. Üstündağ, *Acta Phys. Pol. A* **130**, 45 (2016).

[22] *NLTK 3.0*, 14 September, 2016, http://www.nltk.org/.