Mugla Journal of Science and Technology

USER OPINION ANALYSIS ON TURKISH TWEETS

Kadir Kebabcı[1*], *Banu Diri*[1]

[1]Computer Engineering Department, Yildiz Technical University, 34220, Istanbul, Turkey
kkebabci@std.yildiz.edu.tr, banuce@yildiz.edu.tr

Received: 17.01.2017, Accepted: 16.05.2017 doi: 10.22531/muglajsci.286291
*Corresponding author

Abstract

Nowadays, the use of micro-blogging services such as Twitter seems to increase exponentially with the popularity of social media. Through these services, users share their opinions, complaints, requests and suggestions about the subjects or institutions and organizations they wish. In this study, it is aimed to develop a system which can detect general opinion s and determine what they are by classifying and summarizing the Turkish tweets sent by users via Twitter. At classification step of this system, SVM and Naive Bayes were used together. The Hybrid TF-IDF method was preferred for the purpose of summarizing classified tweets. According to the results obtained, this developed system has been found to be able to successfully determine the opinions of the users and to get a general idea of what is going on.

Keywords: Turkish Tweets, Opinion Detection, Machine Learning, Summarization

TÜRKÇE TWEETLERDE KULLANICI GÖRÜŞ ANALİZİ

Öz

Günümüzde sosyal medyanın popülerlik kazanmasıyla birlikte, Twitter gibi mikro-blog servislerinin kullanımının katlanarak arttığı görülmektedir. Bu servisler aracılığı ile kullanıcılar, diledikleri konu veya kurum ve kuruluşlar hakkında görüşlerini, şikâyetlerini, istek ve önerilerini paylaşmaktadır. Bu çalışmada, Twitter üzerinden kullanıcıların gönderdikleri Türkçe tweetlerin sınıflandırılmasını ve her bir sınıfa ait tweetlerin özetlenmesini otomatik hale getirip raporlayan sistem geliştirilerek, genel görüşün tespiti ve neler olduğunun belirlenmesi hedeflenmiştir. Tweetlerin sınıflandırılması için SVM ve Naïve Bayes yöntemleri birlikte kullanılmıştır. Sınıflandırılan tweetlerin özetlenmesi amacı ile Hibrit TF-IDF yöntemi tercih edilmiştir. Elde edilen sonuçlara göre, geliştirilen bu sistemin, kullanıcı görüşlerinin tespiti ve neler olduğu hakkında genel fikir edinmeye başarılı bir şekilde olanak sağladığı görülmüştür.

Anahtar Kelimeler: Türkçe Tweetler, Görüş Tespiti, Makine Öğrenmesi, Özetleme

1 Introduction

Twitter is at the top of its growing use of micro-blog services. This environment, where millions of active users can share their views, is a valuable asset in terms of information resources [1]. For this reason, Twitter has gained popularity in scientific studies on the identification and evaluation of user opinions. It is almost impossible for the institutions or individuals who value the opinions of the users to manually evaluate tweets without scientific solutions in this platform where there are millions of sharings daily. The aim of this study is to develop a system that automates and reports the evaluation process for Turkish tweets. Detection of user opinions in Twitter is done by classifying tweets. Tweet classification is a sub-field of text classification [2], in which many studies are carried out using machine learning and natural language processing methods. Due to the 140-character constraint and improper language structure, it includes additional difficulties that traditional text classification is insufficient. Approaches developed on classification of tweets can be examined under three headings. These are dictionary based methods [3], machine learning methods [4], and hybrid methods in which both are used together [5].Since a large number of tweets belonging to each class are obtained after classification, tweet summarization is done for each class to be able to get a general idea of what the user opinions are. There are two main methods used in the literature for tweet summarization studies. First one is the automatic creation of the summary tweet by bringing together the important parts from the tweets. The second method is to select the tweets best represent the class to which it belongs [6]. In this study, Hybrid TF-IDF method [7-9], which is more successful than other methods based on the selection of summarytweets, was used.

2 System Design

In this study, it is aimed to develop a system that contains data collection, classification, summarization and reporting modules. To be able to manage these modules, a web application is also developed using Java programming language and through these web interfaces, data collection, tweet labeling and graphical reporting tools gives extra ability to understand opinions.System design is shown at Figure 1.

2.1 Data Collection and Preprocessing

Twitter provides an API that we can listen to, stream live, or search for historical data using filters. Since the language in the developed system is Java, Twitter4J that uses the Twitter API has been preferred. The collected tweets are stored on MongoDB .

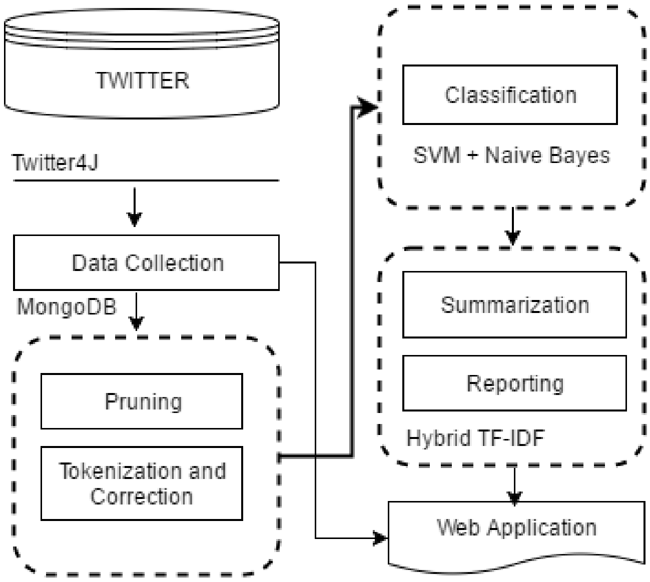

Figure1.System design.

Tweets are generally comprised of very noisy data because of its informal text format, 140 characters constraint and special terms. Therefore, to be evaluated successfully by classifiers, this data must be pre-processed for correction and pruning.

At pruning process which is the first step of pre-processing, usernames, punctuations and words which start with numbers are removed. After that, "#" characters are removed from hashtags and at last all characters are converted to lower case.

At tokenization process which is the second step of pre-processing, firstly the presence of chat abbreviations in the tweet is checked and if it is available, it is changed to its full state.For example,"tşk" is changed to "teşekkür ederim" which means "thank you" in English.After that, the tweet is tokenized. For tokenization and further processes, Zemberek API which is NLP tool for Turkish Language is used.Each word is checked if it is a noun or an abbreviation used in Turkish (eg TSK).If the word is one of them, no correction will be done butif not, the system tries to find the root of the word. If the system can't find the roof of the word, some correction steps are triggered because of the possibility of misspelling. Checking the correction of the misspelled word is controlled by using Turkish dictionary contains approximately one million words.The first step of correction is making deasciifier process. If the system still can't find the word in the dictionary after deascifiying, repeated characters are removed if exists. The reason for not doing this at the beginning is prevent to lose meanings of some words like "saat" which means "clock" in English. If the word isstill not in the dictionary, similar words with 1 letter error rate are found in the Turkish dictionary and the closest one is selected. The "Edit Distance" method is used to calculate proximity.

At the end of these operations, if there are any negative verbs, the expression "_neg" is added to the word as postfix.

2.2 Classification

Separating tweets as positive and negative is inadequate in some cases. For example; tweets about an institution or company may include opinions such as questions, requests, suggestions, information as well as positive and negative opinions. In order to determine these opinions expressed by tweets, multiple labeling method was preferred. For this reason, "Positive", "negative", "question", "requests", "information / news" and "unknown" labels were selected.

To build feature vector, each word in the lexicon is selected as a feature, and if this word is found in a tweet, the value of "1" is assigned to it; respectively, if it is not found, the value of "0" is assigned. Thus, as a result, the tweet's feature vector size will be the same as the size of the lexicon, and the values will be "1" or "0".

SVM and Naive Bayes models are created by using pairs of classes. Since there are 6 classes in total, 15 models are created for each classification method. SVM models are called as SVM Layer which is the first one at classification process and Naive Bayes models are called as NB Layer which comes after SVM Layer. SVM Layer can also be called as parliament because of its role in classification. Tweet is classified by each model in SVM Layer and two of the highest ranked class are selected.After that, to make last decision for the tweet label, the NB model is selected which was trained by using tweets belongs these classes.

For example, if a tweet is classified as "Negative" by 5 models and "Positive" by 4 models in SVM Layer, for the last classification to decide the tweet label, the NB model is selected which was trained with "Negative" and "Positive" labeled tweets. And so the tweet will be classified by the selected NB model.

2.3 Summarization and Reporting

After the classification, tweets deemed as those of high priority are clustered to be summarized. Sum Basic, TF-ISF (Term Frequency-Inverse Sentence Frequency) and Hybrid TF-IDF methods are used for summarization. SumBasic [16] uses simple word probabilities with an update function to compute the best k posts. TF-ISF takes into account the word frequency in sentences and the word frequency in clusters to compute the best k posts [17]. At last, Hybrid TF-IDF method is a customized version of TF-IDF for tweet classification, which is shown at Equation 1.

$$\mathrm{TF\_IDF} = \mathrm{tf_{ij}} * \log_2 \frac{\mathrm{N}}{\mathrm{dfj}} \qquad (1)$$

In Equation (1), $\mathrm{tf_{ij}}$ (Term Frequency) is a frequency of the word $\mathrm{T_j}$ in the Di document; N is a document count and dfj is a count of document which contains the $\mathrm{T_j}$ word. Thus, IDF (Inverse Document Frequency) represents the information value of a word in that document, and is calculated by taking the logarithm of the document count containing the word divided by the total document count. However, tweet is not a standard document and so if all tweets represent one document, IDF will lose its value due to the singularity of the document; otherwise, if each tweet represents a document to keep the value of IDF intact, TF's value will be minimal and nearly the same in all cases. To solve this problem at study [8], TF-IDF's definition is customized according to the hybrid documentation. In the hybrid documentation, all tweets represent one document while the TF value is being calculated, and each tweet represents a document while the IDF value is being calculated. In addition, the weight of a post is normalized by dividing it by a normalization factor, as in Equation (2), since this algorithm will always have a bias towards longer posts.

$$P(t) = \frac{Hybrid\ TF\_IDF}{\max[threshold, word\ count]} \qquad (2)$$

In Equation (2), P(t) is summarization score while t represents the tweet and is calculated by dividing the Hybrid TF-IDF score by the maximum value between the word count of the tweet and the threshold.

After completion of summarizing process, the system creates reports automaticly according to results. These reports includes graphical charts which shows classification rates for each class, summarization result for each class and tag cloud that consists most used words for additional functionality.

3 Experiments

The official accounts of the 10 municipal offices in Istanbul were used as keywords for the tweets to be collected in order to evaluate the success of the developed system. Today, Twitter users frequently express their opinions about municipalities, questions, requests by labeling official accounts of municipalities. For this reason, 1135 tweets about municipalities were labeled by two people. Of these, 169 are positive, 279 are negative, 156 are uncertain, 170 are questions, 157 are requests and 154 are information / news.

While evaluating the system's classification success, the tweets of each municipality were taken once as test data and the rest as training data. For example, one step is the test report of the Üsküdar municipality, while the other one is the test report of Kadikoy municipality. The results are shown in Table 1.

Table 1. Test results of the system (MA : Municipal Office account, TC: Total Tweet Count, SC: Successfully classified tweet count, A%: System Actuary)

MA	TC	SC	A%
@FatihBelediye	111	75	68
@GOPBelediye	107	76	71
@SarıyerBelediye	119	77	65
@Sislibelediyesi	122	87	72
@Uskudarbld	120	79	66
@besiktasbel	112	81	73
@beylikduzubeltr	103	72	70
@eyupbelediyesi	102	72	71
@kadikoybelediye	134	85	64
@pendik_belediye	105	73	69
	1135	**777**	**68**

As seen in Table 1, the success rate of the system in classifying tweetswas measured as 68%. After the classification evaluation, the system was trained by using all data. Then, by selecting the keyword @Uskudarbld from the municipal accounts, opinions were found, summarized and reported on the newly shared tweets which were not available in the train data.316 tweets were collected using @Uskudarbld keyword. Of these, 27% were identified as information/news, 22% as questions, 14% as positive, 13% as negative, 13% as uncertain and 11% as requests.
The positively identified tweet example: "Mosque toilets are both free and extraordinarily clean. I breathe, I kiss your forehead @uskudarbld". For this tweet, 5 of the SVM models classified as positive and 4 of them classified as negative. The Naive Bayes model, which was trained with positive and negative tweets, made a final decision and labeled it as positive. Another tweet example that shows the importance of Naive Bayes model is: "We participated in the Sahur program which our mayor Mr. @hilmiturkmen34 has organized @uskudarbld". In SVM layer, 4 models labeled this tweet as information/news and 4 models labeled as negative. The Naive Bayes model, trained with information/news tweets and negative tweets, labeled it as information/news. After detecting the opinions, next process is summarizing the tweets related to each opinion. Using Hybrid TF-IDF, tweets were selected as summary tweet with 3 of the highest scoring. To summarize the results, only one summary tweet which has highest score is given in Table 2.

Table 2. Summarization Results

Class	Summary Tweet
Positive	Tam anlamiyla sokak iftari... Tesekkurler @uskudarbld su güzel organizasyon icin tebrikler @hilmiturkmen34! *(it was exactly street iftar... Thanks @uskudarbld for this good organization, congratulations @hilmiturkmen34!)*
Negative	Sizin ecdat sevginiz işte bu kadar! RANT,RANT,RANT @uskudarbld @hilmiturkmen34 *(That's all your ancestry love! RANT,RANT,RANT @uskudarbld @hilmiturkmen34)*
Question	@uskudarbld Kız kulesine giden yolda sahildeki korkuluklar ne zaman yapilacak? Bu kadar ihmarkarlik olur mu????? *(@uskudarbld When will the fence on the beach which is on the road to the Maiden's Tower be put? Would it be so negligent?)*
Request	@hilmiturkmen34 @uskudarbld maganda şöförler çalıştırmak yakışmamaktadır. Gereğini yapmanızı rica ederim. İyi çalışmalar. *(@hilmiturkmen34 @uskudarbld employment of roughneck drivers isn't befit you. I would ask you to do the necessary. Good works.)*
Information /News	Acibadem TiBAŞ Parkı önünün @uskudarbld since 30 yıllığına Eğitim ve Sosyal amaçlı kullanımı için oylama bugün! *(The voting for Acibadem TIBAS Park's use for Education and Social use for 30 years by @uskudarbld today!)*

Because the summarization results are relative, the summaries of the tweets were manually evaluated in order to best represent the class to which they belong. According to the results obtained, summary tweets allowed to get general idea about the tweets of the detected opinion.

4 Result

It was seen that the developed system in this study enabled to get opinions about the distribution of the general view and the idea of what is going on by classifying tweets. This system, in which the results of the reporting can be examined through the web interfaces and the results that are thought to be wrongly determined can be corrected, offers an alternative scientific solution for manual evaluation in the environment where there are many tweets for the people, institutions and organizations who value the user opinions.

The first goal in future studies is to expand the data set to make this system more successful than 68% success rate. Furthermore, since the scope of the study is Turkish tweets, it is considered to further improve the pre-processing steps and to evaluate the expression of the emojis.

Another goal is to improve the summarization performance by ensuring that the results are similar in the summarization process and if they are similar, they are not selected as a summary tweet.

5 Acknowledgment

This study was presented in "Akıllı Sistemlerde Yenilikler ve Uygulamaları"-ASYU 2016

6 References

[1] Java, A.,"Why we Twitter: Understanding microblogging usage and communities", *Proceeding of the 9th Web KDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ACM, 2007.

[2] Niharika, S.,Latha, V.S. and Lavany, D.R., "A Survey On Text Categorization", International Journal of Computer Trends and Technology- Volume 3, Issue1, 2012.

[3] Neethu, M.S. and Rajasree, R., "Sentiment Analysis in Twitter using Machine Learning Techniques", *4th ICCNT* 2013 July 4-6, Tiruchengode, India, 2013.

[4] Bahrainian, S.A. and Dengel, A., "Sentiment Analysis Using Sentiment Feautures",*IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)*, 2013.

[5] Lima, C.E.S. and De Castro, L.N., "Automatic Sentiment Analysis of Twitter Messages*", Fourth International Conference on Computational Aspects Of Social Networks*, 2012.

[6] Bahrainian, S.A. and Dengel, A., "Sentiment Analysis and Summarization of Twitter Data",*Proceedings of the 2013 IEEE 16th International Conference on Computational Science and Engineering*, p. 227-234, 2013.

[7] Sharifi, B.,Hutton, M.A. and Kalita, J.K., "Experiments in Microblog Summarization", *In Proc. of IEEE Second International Conference on Social Computing* ,2010.

[8] Inouye, D. and Jugal, K.,"Comparing Twitter Summarization Algorithms for Multiple Post Summaries", *IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, 2011.

[9] Kebabcı, K. and Karslıgil, M.E., "High priority tweet detection and summarization in natural disasters", *23nd Signal Processing and Communications Applications Conference (SIU)*, p. 1280 – 1283, 2015.