Global Journal of Computer Sciences: Theory and Research

Volume 8, Issue 3, (2018) 143-148

www.gjcs.eu

Application for sentiment and demographic analysis processes on social media

Harun Ozkisi*, Kesan Yusuf Capraz Applied School, Trakya University, 22800, Edirne, Turkey
Murat Topaloglu, Kesan Yusuf Capraz Applied School, Trakya University, 22800, Edirne, Turkey

Suggested Citation:
Ozkisi, H. & Topaloglu, M. (2018). Application for sentiment and demographic analysis processes on social media. *Global Journal of Computer Sciences: Theory and Research*. *8*(3), 143–148.

Received from April 17, 2018; revised from August 18, 2018; accepted from November 21, 2018.
Selection and peer review under responsibility of Prof. Dr. Dogan Ibrahim, Near East University, Cyprus.

Abstract

Consumers used to make their complaints via phone and mail before the concept of social media was developed. Now, consumers have begun to state their wishes and complaints concerning companies using social media, and the firms' adaptation to social media has been increased. Understanding the negative and positive attitudes of customers towards ads and questions has gained utmost importance for the companies' decisions to be made for the future. Today, companies are able to continue their existence in the modern world's competitive environment by adjusting their advertising and marketing strategies and calculating their budgets to social media analyses they get. In addition, companies depend on those analyses in order to determine their positions in the market and create their action plans. In this study, instant messages sent on social media and demographic information were used in the data analysis in order to determine whether those messages included positive or negative attitudes.

Keywords: Social media, emotion analysis, data mining, natural language processing.

* ADDRESS FOR CORRESPONDENCE: **Harun Ozkisi**, Kesan Yusuf Capraz Applied School, Trakya University, 22800, Edirne, Turkey.
E-mail address: harunozkisi@trakya.edu.tr / Tel.: +90-546-235-0576

1. Introduction

Learning, which has a crucial meaning for human beings, is also a fundamental component of the information technologies in the modern world. For that reason, many studies have been done in order to understand whether computers can learn, comprehend and reason just like human beings. As a result of these studies, various systems were developed in different fields and they are used frequently today. The aim of these systems is to form predictive data based on previous data. Therefore, the computer systems can renew itself based on two situations while solving similar problems faster and more effectively compared to the situations encountered before (Dalyan, 2006). Trying to solve the problems like that has led to the emergence and development of machine learning as a field. Machine learning is defined as the programming of computers in such a way that they make definitive or predictive deductions with the inductive method by making use of sample data or past experiences in the literature (Alpaydin, 2004).

Today, machine perception, artificial vision, natural language processing, syntactic pattern recognition, search engines and medical diagnosis are only some of the areas where machine learning is utilised. One of the most important technological developments in recent years is data mining which emerged as a solution to generate the information that the decision support systems need. Especially, a decision tree and a rule induction are used in machine learning and data mining algorithms (Erdogan, 2004).

Today, companies have to keep up with the others in an extremely competitive environment. Apart from the competition, companies do want to see whether customers have positive and negative attitudes towards the goods and services they offer. For that reason, social media become more important as people tend to express their positive and negative attitudes in this information age. The term 'social network' was first coined in 1954 by Barnes who was an anthropologist (Cetin, 2009) and implies a society formed on the Internet where people interchange their ideas and interact with each other for a common goal (Karal & Kokoc, 2010). Social networks, which have become a significant topic that must be analysed, represent the connection between people and the power of this connection (Onat & Asman, 2008). Recently, many Internet users have been commenting on the goods sold online and almost everything on their personal blogs and social media sites like Facebook, Twitter and Blogger pages (Kaya, 2013). Today, companies do analyses on those comments to make critical decisions for the investments they make. Therefore, how and when to use social media measuring and tracking methods become very important for companies (Sari, 2013). There is a great deal of research done on this subject. Binbir evaluated the use of social media optimisation on communication strategies with the data obtained from online monitoring and online social media services and the resulting effects (Binbir, 2012). Albayrak also examined the relationship between the use of Turkish and the psychological state (Albayrak, 2011), whereas Soysal focused on the definition of a data induction system, which extracts available data and turns it into a data model by processing Turkish radiology reports that were unconstrained text (Soysal, 2010). As a result, the analysis of the data on social media has reached utmost importance for many fields.

2. Methods

In this study, the activities on various social media websites were analysed with artificial intelligence methods. N-Gram model and Naive Bayes algorithm were used for the analysis in order to determine the demographic information for the comments. In addition, a process was followed in order to decide whether a tweet on the social media has a positive or negative comment.

2.1. N-gram model

N-Gram model, which is used in sentiment analysis, represents *n* character slice(s) of a string. The frequency of the *n*-gram characters used in a document is taken into account when the n-gram based classification method is used. 2-gram, 3-gram and 4-gram formats were used in this study.

This method is based on the definition of n-gram formations in a document. *N*-gram frequency method works independently of the language. As the frequency of the characters used forms, the basis of classification, the keywords chosen related to the subject are important for the classification.

Natural languages include some words that do not change and used more frequently compared to others. This assumption applies to all languages. One of the most familiar ways to state this assumption is the Zipf's (1949) law. According to this law, some words of a language are more frequently used than the others. It is an assumption that seems right for the words related to a specific topic. A threshold value is used in order to separate the weighty n-grams from the ones that are used less frequently. *N*-gram value gives us positive or negative results by using computer learning based on the dataset that we choose. Therefore, whether the content has a positive or a negative meaning can be detected.

2.2. Naive Bayes algorithm

Bayes classification, which is one of the scientific decision-making methods, is a statistical topic that is still being developed. With this method, in order to get an optimal result, we try to combine two sources. The first resource is the objective data value and the other is the degree that a person's acceptance of a theory, an idea that is accepted worldwide, *apriori* knowledge or a subjective thought or phenomenon with many possibilities (Efron, 1986).

3. Data analysis

Data analysis with Hadoop, live data retrieval and sentiment analysis are discussed under this title. *N*-gram model and Naive Bayes Algorithm were used for the analysis.

3.1. Data analysis with Hadoop

The data obtained from Twitter were transferred to Hadoop system first. The content of the data includes hashtags, retweets and co-hashtags. According to the critical words identified on the transferred data by using the K-Means algorithm, the desired results were obtained. K-means algorithm is based on the clustering of the data that has similar properties. The aim of this algorithm is to have clusters that are as homogenous as possible within them while being as distinct as possible from the others.

Besides, by using Ma-Reduce which is a subunit of Hadoop and used for assembling and reducing data, it is possible to access the data. The analysis was done on about 350,000 tweets. The data obtained from the tweets that were sent by Twitter users around the world went through the clustering process in hashtags and popular words. K-Means algorithm also revealed the relationship between the tweets which provided meaningful data for our research. In addition, the tweets sent were subjected to live analysis. Then, they were clustered and categorised with Map-Reduce. The aim of categorisation is to make the data meaningful in itself based on their individual meaningful. After the categorisation, the distance between the data was found by using the K-Means algorithm. While clustering the data by looking at the distance to each other, machine learning is carried out at the same time.

Thus, it is possible to do the demographic analysis of the obtained data using Map-Reduce and K-Means algorithm. In this way, we are able to see both the graphical interface and the digital table of

the data. As shown in Figure 1, a more meaningful and clustered data was formed by a graphical interface.

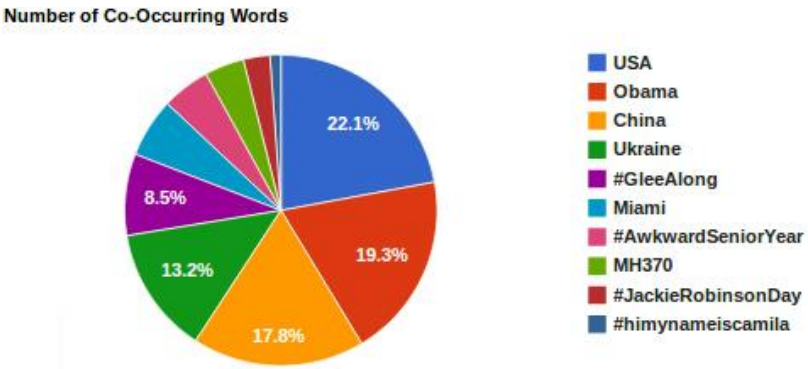

Figure 1. Demographic data analysis with Hadoop

The ratios of the input data and analyses are shown in Figure 1. After the graphic interface categorisation, the hashtags and tweets were calculated.

3.2. Real-time sentiment analysis

The tweets and hashtags are retrieved from Twitter, and the data are saved to MongoDB. The data saved on the system need to be transformed into a format that is appropriate for machine learning and sentiment analysis. This is where the n-gram model is used. The incoming content is identified first as a high priority and low priority based on their niceness values. NLTK, Python's natural language toolkit, includes the language library for sentiment analysis in Turkish. The library has sensitive and insensitive categories in terms of content for the machine learning as the next step. Real-time analysis was done for tweets in Turkish in this study. The tweets obtained in this way were gathered in a pool and then their n-gram values were calculated.

Starting with 2,000 tweets, we formed two groups which are sensitive and insensitive data criterion after the machine learning process. When we search for a word or sentence in the system, the system makes a comparison by using the database available in the system and the data it has learned. This comparison process is done by means of the n-grams that helped the system learn. At this stage, the information of the tweets whether they are positive or not was obtained through the keywords written on the program. The program interface by which the real-time analysis was done is presented in Figure 2.

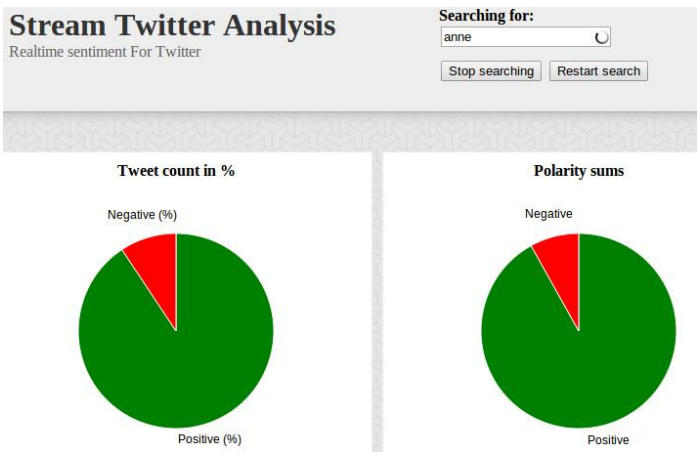

Figure 2. Real-time analysis interface

As an example, the word 'anne' (which is the word for *mother* in Turkish) was searched on the program. The results change instantaneously in graphics. The positive and negative tweets are shown in green and red, respectively. The program running background was sending tweets while saving data by means of machine learning at the same time. After the tweet frequencies were obtained, an analysis was done considering the n-gram values for the data in addition to the sentiment analysis for the tweets appearing instantaneously. Two examples of positive and negative tweets that underwent analysis are shown in Figures 3 and 4.

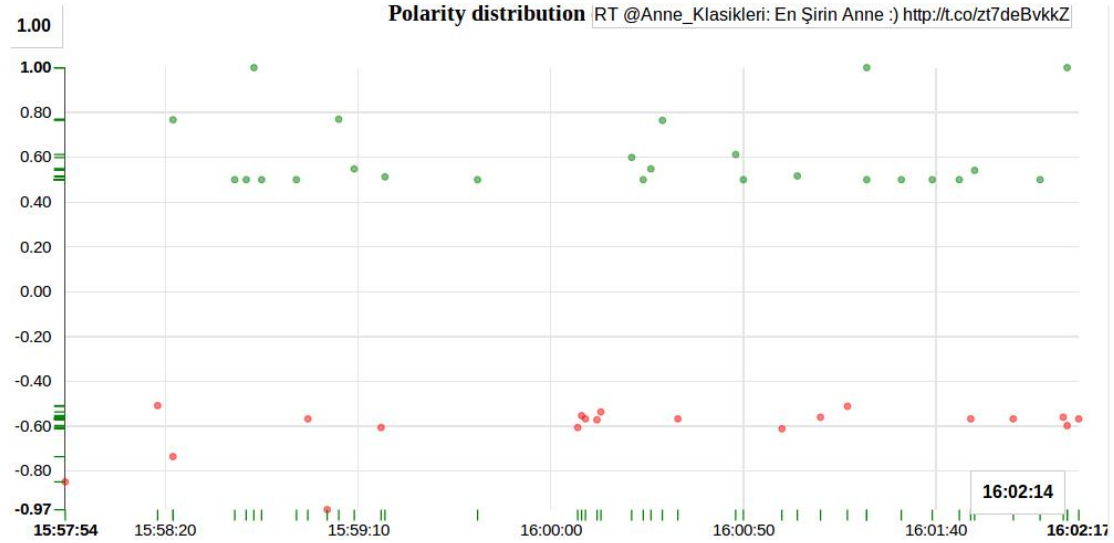

Figure 3. Positive tweet examples on which sentiment analysis were done

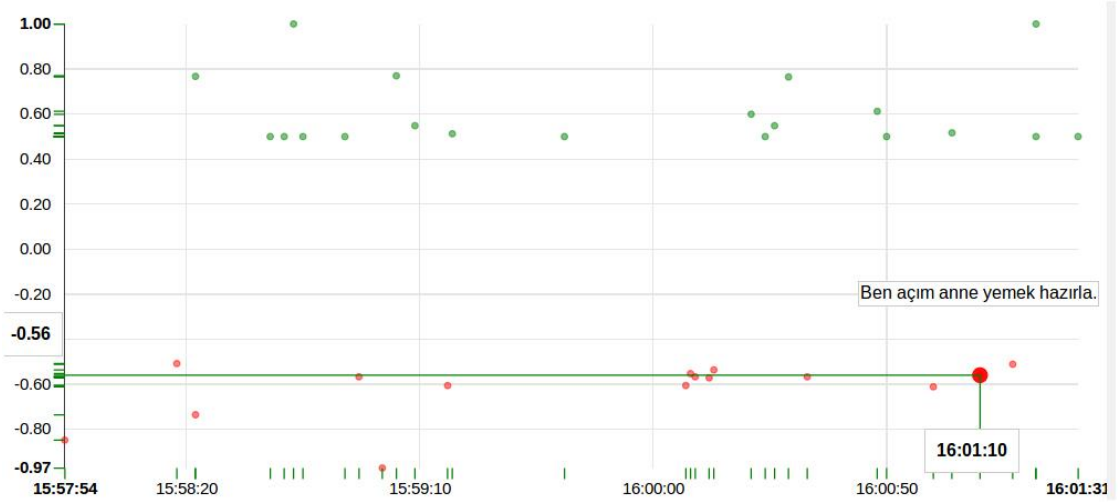

Figure 4. Negative tweet examples on which sentiment analysis were done

The values below and above zero on the graphic are interpreted as negative and positive, respectively. A tweet being close to zero shows that it may have a neutral meaning. As seen in the figure, the tweet *En Sirin Anne :)* (which means *the cutest mom* in Turkish) is included in the positive group while the tweet *Ben acim anne yemek hazirla* (which means *I'm hungry mom get the meal ready*) is placed in the negative group.

4. Results and discussion

With the development of information technologies and social media applications, people have changed the places and media where they communicate with others and sharing what they think. Thus, the analysis of the comments and communications happening there is of importance for the world of science, as well as for the companies trying to carry on in a harsh competition. Understanding the negative and positive attitudes of customers towards ads and questions has gained utmost importance for the companies' decisions to be made for the future.

In this study, some points were made to highlight the data created on social media in terms of being positive or negative. It is thought that these highlights can be useful for companies and their decisions for the future. The results may also be informative for customer relations and brand values.

References

Alpaydin, E. (2004). *Introduction to machine learning*. The MIT Press.

Albayrak, N. B. (2011). *Opinion mining and sentiment analysis using natural language processing techniques* (MSc Thesis). Istanbul, Turkey: Fatih University.

Binbir, S. (2012). *Social media optimisation as corporate communication tool: business social media optimization practices as corporate communication tool* (MSc Thesis). Izmir, Turkey: Yasar University.

Cetin, H. (2009). *Social support and social network for elder migrants from Bulgaria* (MSc Thesis). Izmir, Turkey: Ege University Izmir.

Dalyan, T. (2006). *1R algorithm in machine learning and forming second rule (2R)* (MSc Thesis). Turkey: Kocaeli University.

Efron, B. (1986). Why isn't everyone a Bayesian. *American Statistician, 40*, 1–11.

Erdogan, S. Z. (2004). *Data mining and K-means algorithm in data mining and an application to a student database* (MSc Thesis), Istanbul, Turkey: Istanbul University.

Karal, H. & Kokoc, M. (2010). Universite Ogrencilerinin Sosyal Ag Siteleri Kullanim Amaclarini Belirlemeye Yonelik Bir Olcek Gelistirme Calismasi. *Turkish Journal of Computer and Mathematics Education, 1*(3), 251–263.

Kaya, M. (2013). *Sentiment analysis of Turkish political columns with transfer learning* (MSc Thesis), Ankara, Turkey.

Onat, F. & Asman, A. O. (2008). Sosyal ag sitelerinin reklam ve halkla iliskiler ortamlari olarak degerlendirilmesi. *Journal of Yasar University*, 1111–1143.

Sari, S. N. (2013). *Social media monitoring and measurement methods: a research study*, (MSc Thesis). Istanbul, Turkey: Yeditepe University.

Soysal, E. (2010). *Ontology based information extraction on free text radiological reports using natural language processing approach* (PhD Thesis). Ankara, Turkey: Middle East Technical University.

Zipf, G. (1949). *Human behavior and the principle of last effort*. Cambridge, MA: Addison-Wesle.