Turkish Journal of Science & Technology
17(2), 161-166, 2022.

Research Paper
https://doi.org/10.55525/tjst.1113832

Multilingual Text Mining Based Open Source Emotional Intelligence

Aytug BOYACI[1*], Shahin AHMADOV[2]

[1] Department of Computer Engineering, , Air Force Academy, National Defence University, Istanbul, Turkey

[2] Department of Computer Technologies, Hezarfen Aeronautics and Space Technologies Institute, National Defence University, Istanbul, Turkey

[1*] aboyaci@hho.msu.edu.tr, [2] shahinya@code.edu.az

(Geliş/Received: 08/05/2022; Kabul/Accepted: 31/05/2022)

Abstract: The purpose of this study is to learn how people who speak different languages interpret the same issues, and to compare the results obtained and show the difference between their perspectives. To learn this point of view, we must first turn to open source intelligence. In this execution, a sentiment analysis application was designed using the Python programming language and the Natural Language Processing algorithms in the texts, which were taken as a data set of comments in Azerbaijani, Turkish, Russian and English languages from social media. As the data set, the comments made on 4 subjects: the declaration of Hagia Sophia as a mosque, the objection events that started with the natural gas hike in Kazakhstan, the natural disasters in Turkey, the Ukraine crisis. After loading the texts in four languages from the network environment, after preprocessing, the text was divided into 8 different categories (neutral, fear, joy, anger, sadness, surprise, disgust, shame) by means of the application written in Python programming language based on Data Mining and Machine Learning topics. In the study, precision, sensitivity, accuracy and F1 score were obtained by using Random Decision Forests, K - Near Neighbor Algorithm, Decision Trees, Support Vector Machine, Naive Bayes Algorithm, Logistic Regression, which are machine learning methods. By comparing the results, it was determined that the Logistic Regression method obtained the highest result. A sentiment analysis model was created using the Logistic Regression method, and sentiment analysis was performed for each subject at separation and the results were compared.

Key words: Text mining, Emotion detection, Machine learning, Natural language processing.

Açık Kaynak Duygusal Zeka Tabanlı Çok Dilli Metin Madenciliği

Öz: Bu çalışmanın amacı farklı dilleride konuşan insanların aynı konuları nasıl yorumladıklarını öğrenmek ve elde edilen sonuçları kıyaslayarak bakış açıları arasındaki farkı ortaya koymaktadır. Bu bakış açısını öğrenmek için ilk önce açık kaynak istihbaratına başvurmamız gerekmektedir. Bu çalışmada açık kaynak istihbaratı olan sosyal medya platformu Intagram üzerinden Azerbaycan, Türk, Rus ve İngiliz dillerinde Ayasofya'nın cami olarak ilan edilmesi, Kazakistan'da doğal gaz zamıyla başlayan itiraz olayları, Türkiye'de oluşan doğal afetler, Ukrayna krizi olmak üzere 4 konuda yapılan yorumlar veri seti olarak kullanılmıştır. Veri seti üzerinde Python programlama dili ve içerisinde bulunan doğal dil işleme algoritmalarını kullanarak duygu analizi uygulaması tasarlanmıştır. Dört dilde metinler ağ ortamından yüklenerek, ön işlemlerden geçtikten sonra veri madenciliği ve makine öğrenmesi konuları baz alınarak Python programlama dilinde yazılmış uygulama vasıtasıyla metin 8 farklı kategoriye (nötr, korku, sevinç veya eğlence, öfke, üzüntü, hayret, iğrenme, utanç) ayrılmıştır. Çalışmada makine öğrenmesi yöntemlerinden olan Rastgele Karar Ormanları, K - Yakın Komşu Algoritması, Karar Ağaçları, Destek Vektör Makinesi, Naive Bayes Algoritması, Lojistik Regresyon kullanılarak kesinlik,duyarlılık, doğruluk ve F1 skoru sonuç olarak elde edilmiştir. Sonuçlar karşılaştırılmış ve Lojistik Regresyon yönteminin en yüksek netice elde ettiği tespit edilmiştir. Daha sonra Lojistik Regresyon yöntemi kullanılarak duygu analizi modeli oluşturulmuş ve her bir konu için ayrılıkta duygu analizi yapılmış ve sonuçlar karşılaştırılmıştır.

Anahtar kelimeler: Metin madenciliği, Duygu analizi, Makine öğrenmesi, Doğal dil işleme.

*Corresponding author: aboyaci@hho.msu.edu.tr. ORCID Number of authors: [1]0000-0003-1016-3439, [2]0000-0002-3068-9377

1. Introduction

Throughout life, people have a desire to learn about various topics and to learn ideas. Lately, we usually get this information from the Internet. Later, we would like to discuss about this information or learn the opinion of others about the said topic. So, if in ancient times people were asking their friends or books for information or ideas, now with the rapid development of information and computing technologies and especially the global internet, a valuable alternative has emerged to find the necessary information and help choose something. The leading method for obtaining this information is open source intelligence. OSINT-open source intelligence is based on online publications, chats, blogs, social networks and streaming platforms. Apart from these, open source intelligence also takes into account the print media such as magazines, newspapers, professional research articles, business data, telephone directories, and other reports that may be useful in extracting data from individuals or organizations representing the interest[1]. As an example, we can say that 80-90% of the intelligence needs of countries in our age are obtained from open source intelligence [2]. Since the aim of the study is to understand people's feelings about current events by making sentiment analysis on texts, open source intelligence helps us to achieve an efficient result. We have decided to use social media platforms as we aim to extract text data from open source intelligence.

Various texts created on social media, such as blog articles, product reviews, social media posts, news comments, etc., contain a lot of valuable information. With the help of this valuable information, we can read people's minds or direct them if needed. One of these methods is sentiment analysis [3]. In social media, sentiment is an emotional evaluation that expresses the direction of thoughts in comments and other texts. The importance of the task of identifying emotions, i.e. sentiment analysis, is that it is possible to evaluate the attitude of society towards a message, product or idea of the text based on text information. This sentiment analysis can be used, for example, to evaluate the success of an advertising campaign, political and economic reforms, or to determine the attitude of the press and media towards a particular person, organization or event. In another area, it provides information on how consumers relate to certain products, services, and organizations. Therefore, this type of information is very important for marketers, sociologists, economists, politicians and all experts whose activities depend on people's opinions [4].

The sentiment analysis conducted in this study aims to determine what people who express themselves in four languages on different topics think and how their emotions change based on the comment texts collected through social media.

This study consists of four parts. The first chapter deals with the introduction and the aim of the study. The second chapter deals with the formation of the machine learning model and the methods of collecting the data used and preprocessing the texts. The third chapter explains the results of the categorization of the text according to the sentiment analysis performed in the study and compares the results according to the applications to the data in different languages. In the last part, the results were evaluated and suggestions for future studies were presented.

2. Material and Methods

In this part of the study, information is given about the procedures before the analysis. First, it will be shown that the model we will use in the study for sentiment analysis consists of the following elements.

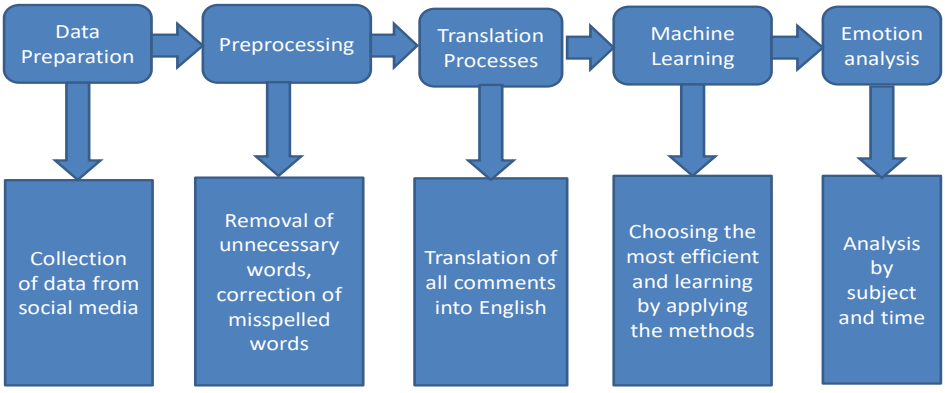

Figure 1. Creating the Model

2.1 Dataset and Preprocessing

Machine Learning algorithms produce predictive models that perform well only if they are trained with sufficient number and quality of data[5]. In this regard, the machine learning process starts with the creation of the dataset.

In the data collection phase, the social media platform Instagram, which is widely used in many countries, was used [6].

Comments on news posts on the Turkish 'Ist1hbarat', 'Haberglobal', 'BBCTurkce' pages on the Instagram platform, comments on news posts on the Russian 'Ria_novosti', 'NTVru', 'BBCRussian' pages, comments on the news shared on the English 'BBCnews' page, In Azerbaijani, the comments made on the news shared on 'BBCAzeri', 'Azerbaycan.24', 'Maraqlı.tv' pages were compiled and used manually.

At this stage, the first step is to check the correct collection of data. The collected data is reread according to the topic and time and wrong data formation is prevented. Then, incorrectly added or classified comments are manually filtered and cleaned. In the next step, the collected data is converted to a suitable format (CSV) so that it can be compiled by the Python programming language. After the format conversion, the row and column layout of the dataset is recompiled. Information loss is minimized by completing missing data. Some data sets are combined into a single data set.

The new datasets created after these preprocessing steps are classified and grouped. The preprocessed and grouped data are translated from Turkish, Russian and Azerbaijani languages into English using 'Microsoft Translator for Business' and 'Google Translate'. Each translated comment is manually compiled and the accuracy of the translation is checked. The translations that occur at this stage, resulting in incorrect or ambiguous translations, are converted into the appropriate text by manual changes. The translated texts are prepared again by applying preprocessing steps. After these steps, the dataset can now be considered ready for use.

2.2. Machine Learning

After the preprocessing steps are done, the evaluation of the operation results made by the Machine Learning methods [7] on the prepared dataset is made according to the obtained F1 score, accuracy, sensitivity and precision values. Six different Machine Learning methods were applied to the pre-prepared dataset and value results were obtained for each of them.

For the Random Decision Forests method, the Random Forest Classifier module of the Scikit-learn library in the Python programming language was used. The "Gini" index was used as the common objective function. An attempt was made to obtain the best result by changing the number of trees.

The K-neighbors Classifier module of the Python library Scikit-learn is used for the K - Near Neighbor Algorithm classification method. In this module, the number of neighbors was set to 3 and 5, and the Minkowski distance and Euclidean distance were chosen as distance measures.

The Decision Tree Classifier module of the Scikit-learn library in the Python programming language was used for the decision tree method. The "Gini" index was used as the common objective function. An attempt was made to obtain the best result by changing the maximum depth settings.

For the Support Vector Machine classification method, the SVC (Support Vector Classification) module of the Scikit-learn library was used. The kernel function is used as a Radial Based Function. The variables are set to 'auto'.

For the classification method with the Naive Bayes algorithm, the module MultinomialNB of the Scikit-learn library was used. Many variants of the Naive Bayes algorithm are used in the scikit-learn library. The main reason for using the Multinomial Naive Bayes algorithm is its application to texts.

For the Logistic Regression classification method, the Logistic Regression module of the Scikit-learn library was used. The experiments were performed by determining the random variables.

2.3. Emotion Detection

Sentiment analysis analyzes human emotions and moods on the dataset, that is, it is a method that analyzes data obtained from customer reviews, financial news, social media comments or other sources according to emotion criteria [8].

With the advent of social media, people are becoming more receptive about their experiences with online products and services through blogs, vlogs, social media stories, reviews, recommendations, reviews, hashtags, comments, direct messages, news articles, and various other platforms. Such sentimental perceptions leave a digital

footprint of how a person expresses their experience in an online space [9]. Most sentiment analysis methods can evaluate these ideas as positive, negative, or simply neutral [10]. If we want to get a more detailed result, we can use methods that include many emotions. With these methods, we can divide the data into categories such as fear, joy, anger, sadness, surprise, disgust, and so on. This helps us a lot in getting the necessary information or forming an opinion about a topic. In the mood analysis phase, the machine learning model was trained using the Logistic Regression module of the Scikit-learn library in the Python programming language [11]. Using the trained model, sentiment analysis was performed on the data categorized according to the mentioned topics.

Using the sentiment analysis, the texts were categorized into 8 sentiment categories such as neutral, fear, joy or fun, anger, sadness, surprise, disgust, and shame.

3. Findings

The results of the machine learning experiments are shown in Table 1. In this table, precision, sensitivity, f1-score and accuracy values of 6 machine learning methods were compared and it was determined that the Logistic Regression result was the highest.

Table 1. Results of Machine Learning Methods

Machine Learning Methods	Precision	Sensitivity	Accuracy	F1 score
Random Decision Forests	0.74	0.60	0.58	0.60
K Near Neighbor Algorithm	0.47	0.24	0.22	0.24
Decision Trees	0.52	0.52	0.52	0.52
Support Vector Machine	0.71	0.61	0.59	0.61
Naive Bayes Algorithm	0.61	0.57	0.54	0.57
Logistic Regression	**0.65**	**0.63**	**0.60**	**0.63**

Based on these results, the Logistic Regression method will be used for the Sentiment Analysis to be done later [12].

3.1. Comparative Analysis of the Announcement of Hagia Sophia as a Mosque

In the sentiment analysis phase, all the topics for each language were analyzed first, and in the next phase, the results of the sentiment analysis of the comments of the people speaking in different languages were compared according to the topic.

After the Sentiment Analysis of the comments made in four different languages on the subject mentioned here, the three emotions with the most comments are shown.

Table 2. Comparative Analysis of the Announcement of Hagia Sophia as a Mosque

Language	Emotions		
Turkish	Joy	Neutral	Fear
Russian	Neutral	Joy	Fear
English	Joy	Neutral	Fear
Azerbaijani	Joy	Neutral	Sadness

As shown in Table 2, we see that the emotion of joy is dominant in the comments written in all languages about the declaration of Hagia Sophia as a mosque. It is also shown that, unlike the others, the comments with a neutral emotion in the comments made in Russian are more than the emotion of joy.

For example, in Russian texts 'Church', 'Mosque', 'Tourist', 'Christian' and 'Prayer' words, in Turkish texts 'Hagia Sophia', 'Mosque', 'People', 'Allah', 'Day' and 'Satisfied' words, in English texts 'Turkey', 'Mosque', 'Muslim', 'Beautiful', 'Love', 'Erdogan', in Azerbaijani texts 'Allah', 'Acceptance', 'Prayer', 'Beautiful', 'Possible' words are used more.

3.2. Comparative Analysis of Objection Events in Kazakhstan

After the Sentiment Analysis of the comments made in four different languages on the subject mentioned in Table 3, the three emotions with the most comments are shown.

Table 3. Comparative Analysis of Objection Events in Kazakhstan

Language	**Emotions**		
Turkish	Fear	Neutral	Joy
Russian	Joy	Fear	Neutral
English	Fear	Neutral	Joy
Azerbaijani	Neutral	Joy	Sadness

As shown in Table 3, we see that the emotion of fear is dominant in the comments written in English and Turkish regarding the protest events that started with the natural gas hike in Kazakhstan. And we see that the emotion of joy in Russian interpretations, neutral in the Azerbaijani language.

For example, in Russian texts 'People', 'Well', 'Russia', 'Natural Gas', 'Price', 'Time', in Turkish texts 'People', 'Kazakhstan', 'Russia', 'Will', 'Turkish' and 'Street', in English texts 'People', 'Kazakhstan', 'Russia', 'Demand', 'Terrorist', 'State', in Azerbaijani texts 'Nation', 'Power', 'Russia', 'Day', 'Demand' words are used more often.

3.3. Comparative Analysis of Natural Disasters in Turkey

In Table 4, after the Sentiment Analysis of the comments made in four different languages about the natural disasters that have occurred in Turkey in recent years, the three emotions that contain the most comments are shown.

Table 4. Comparative Analysis of Natural Disasters in Turkey

Language	**Emotions**		
Turkish	Joy	Neutral	Fear
Russian	Joy	Neutral	Fear
English	Joy	Sadness	Fear
Azerbaijani	Neutral	Joy	Fear

As shown in Table 4, we see that the emotion of joy is dominant in the comments written in English, Russian and Turkish about natural disasters in Turkey. In the Azerbaijani, we see that the emotions of neutrality and later joy is superior. It is used in large numbers in comments that contain a emotion of fear, usually in all interpretations.

For example, in Russian Texts 'God', 'Turkey', 'Always', 'Good', 'Pain', in Turkish texts 'Hotel', 'Aim', 'Fire', 'Burn', 'Immediate', In English texts, 'People', 'Care', 'Smooth', 'God', and 'Walk', in Azerbaijani texts the words 'Allah', 'Help', 'Prayer', 'Protect' and 'Remember' are used more often.

3.4. Comparative Analysis of the Ukraine Crisis

In Table 5, after the Sentiment Analysis of the comments made in four different languages regarding the crisis in Ukraine in 2022, the three emotions with the most comments are shown.

As shown in Table 5, we see comments that contain more sadness in the comments written in English and Azerbaijani on the subject. Neutral and fearful interpretations are used a lot in Turkish, while interpretations containing fear are used a lot in Russian.

For example, in Russian texts 'Ukraine', 'Russia', 'War', 'People' and 'Thinking', in Turkish texts 'World War', 'Russia', 'Putin', 'Countries', in English texts 'War', 'Russia', 'Stop', 'Need', in Azerbaijani texts 'War', 'Ukraine', 'Soldier', 'Nation', 'World' and 'Persistence' words are used more common.

Table 5. Comparative Analysis of the Ukraine Crisis

Language	**Emotions**		
Turkish	Neutral	Fear	Joy
Russian	Fear	Neutral	Joy
English	Sadness	Fear	Anger
Azerbaijani	Sadness	Neutral	Joy

4.Conclusion

We see that open source intelligence is very important to understand the emotions of people speaking different languages, which is our in four languages: Turkish, Russian, English and Azerbaijani The second chapter talks in the study. It is easy to access all kinds of data and information we need on the Internet. However, the open-source nature of this data creates the opportunity to act within the legal framework. For this reason, intelligence data collected on the Internet should be open source. More than six thousand comments were collected on the social media platform Instagram, which has open-source intelligence data in four languages, Turkish, Russian, English, and Azerbaijani, because it is legal and contains information on the topics needed for the study. After the preprocessing steps were done on the collected data set, machine learning methods were used for sentiment analysis and the results were compared. In this phase, machine learning methods such as Random Decision Forests, KNN algorithm, Decision Trees, Support Vector Machine, Naive Bayes algorithm and Logistic Regression were used. Logistic Regression was selected with the highest result, an F1 score of 0.63. A machine learning model was created and preparations for sentiment analysis were seen Sentiment analysis was performed according to the timing and announcement of Hagia Sophia as a mosque, the objection events that started with the natural gas surge in Kazakhstan, the natural disasters that occurred in Turkey, and the Ukraine crisis. The sentiment analysis was performed separately for the comments made in four languages on each topic and the results were displayed. As a result of the sentiment analysis, we see that the indicator of the emotion of joy we obtained is the highest. The reason is that the sentiment analysis perceives the allusions and ironic expressions used in the comments as joy. For example, in the comparative analysis of natural disasters in Turkey, many comments about the construction of hotels in fire areas were rated as feelings of joy. For future studies, it is suggested to develop models to detect such expressions.

References

[1] Eliot Higgins, “We Are Bellingcat: An Intelligence Agency for the People”, Bloomsbury Publishing, 2021, pp. 9-63.
[2] Stevyn Gibson, “Open Source Intelligence, An Intelligence Lifeline”, Royal United Services Institute Journal, 2004, pp 5-6.
[3] Svetlana Tupikova, “Cognitive Foundations of Communicative Tonality”, Lambert Academic Publishing, 2020, pp. 28-44.
[4] A.G. Dodonov, D.V. Lande, V.V. Prishchepa, V.G. Putyatin, “Computer competitive intelligence”, Engineering, 2021, pp. 15-18.
[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., “Scikit–learn: Machine learning in Python“, J. Mach. Learn. Res. 2012, pp. 2825–2830.
[6] We are social and Hootsuite, “Digital 2020” report, 2021.
[7] Engin Sorhun, “Machine Learning with Python”, Abakus, 2021, pp. 9-43.
[8] Hassan Saif, “Semantic Sentiment Analysis in Social Streams”, IOS Press, 2017, pp. 26-36.
[9] Harkamal Preet Pal Singh Ubhi, “The Social Media Guide”, Rakuten, 2019, pp. 7-31.
[10] Bing Liu, “Sentiment Analysis”, Cambridge University Press, 2020, pp. 16-46.
[11] Michael Bowles, “Machine Learning with Spark and Python: Essential Techniques for Predictive Analytics”, Wiley, 2019, pp. 129-166.
[12] Joseph M. Hilbe, “Practical Guide to Logistic Regression”, CRC Press, 2016, pp. 49-70.