

A Review of Turkish Sentiment Analysis and Opinion Mining

H. Karayığit, Ç. Acı and A. Akdağlı

Abstract—Social Media is one of the most frequently used platforms today. Users can easily share their views, ideas, and thoughts on this platform. The data shared on social media platforms is actually a great deal that can be transformed into meaningful information. The obtained big data can be analyzed and evaluated by various data analysis methods. Whether or not the data contain a feeling, if it is included; the type of the feeling (i.e. positive, negative or neutral) can be determined by emotion analysis methods. Sentiment Analysis studies in later times began to turn to analysis indicating different sentiments. Thus the foundations of Opinion Mining were laid. When ideas conveyed by social media information are presented semantically, they are expressed by Opinion Mining. The purpose of this paper is to explain the relationship between the concepts of Sentiment Analysis and Opinion Mining. The terms used in Sentiment Analysis and Opinion Mining are explained and examples of Turkish Sentiment Analysis are given. It has been tried to suggest solutions for the problems encountered in Turkish studies.

Index Terms— Turkish, Sentiment Analysis, social media, Opinion Mining.

I. INTRODUCTION

SENTIMENT analysis and Opinion Mining are the fields of study that analyzes people's views, evaluations, attitudes, and emotions from a written platform. Sentiment Analysis is used to find out what an existing text expresses emotionally. In some studies, the concept of Sentiment Analysis is also referred to by names such as thought mining, thought extraction. In industrial applications, Sentiment Analysis is more commonly used. Sentiment Analysis and Opinion Mining nomenclature are used in academic studies [1].

Opinion Mining is trying to show the meaning of the present text scientifically. Opinion Mining is the concept of Sentiment Analysis as a title [2]. The first study on Sentiment Analysis was conducted by Pang, Lee and Vaithyanatham in 2002 and movie comments existing in Internet Movie Database were taken and classified by various machine learning algorithms [3].

H. KARAYIĞIT, Department of Electrical and Electronics Engineering, Faculty of Engineering, Mersin University, Mersin, Turkey, (e-mail: d2014242@mersin.edu.tr).

Ç. ACI, Department of Computer Engineering, Faculty of Engineering, Mersin University, Mersin, Turkey, (e-mail: caci@mersin.edu.tr).

A. AKDAĞLI, Department of Electrical and Electronics Engineering, Faculty of Engineering, Mersin University, Mersin, Turkey, (e-mail: akdagli@mersin.edu.tr).

Manuscript received August 19, 2017; accepted Nov 16, 2017.

DOI: [10.17694/bajece.419547](https://doi.org/10.17694/bajece.419547)

The first study in the concept of Sentiment Analysis and Opinion Mining are used together, semantic relations were established between emotional expressions and the subject rather than being classified as positive or negative [4].

In the first work under the name of Opinion Mining, a list of both product qualities (quality, characteristics) was created using the Opinion Mining tools, and opinions (weak, mixed, good) were collected about each one [5].

Although they look like different concepts, Sentiment Analysis, and Opinion Mining cannot be considered separately. Very similar methods are used while the significance of the dataset is more important for the Opinion Mining, the aim is to make emotions meaningful in Sentiment Analysis.

The remainder of the paper is organized as follows: In the first part of the second section, the concepts of Sentiment Analysis and Opinion Mining and their relation are mentioned. In the second phase of the second section, the Turkish Sentiment Analysis studies are briefly explained. In the conclusion section, evaluation was made for Sentiment Analysis and Opinion Mining relationship and it has been tried to suggest solutions for the problems encountered in studies conducted in Turkish. In the last section, references are included.

II. TURKISH SENTIMENT ANALYSIS

II. I. TURKISH SENTIMENT ANALYSIS AND OPINION MINING

Grammar rules vary for each language. Verb conjugation in a sentence differs between languages. The Turkish language is a structurally agglutinative language. Structural processing is more difficult than English. In Turkish, verb conjugation and lexical items are different from other languages as shown in Figure 1.

Gender discrimination in languages such as Arabic, English, German does not exist in Turkish. A sentence spoken by multiple words in English can be explained in Turkish by a word. For example; 'Gidemedik' is expressed in English 'We were not going'. There are 8 vowels (a, e, i, o, ö, u, ü) and 21 consonants (b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z) in Turkish. The seven letters are unique to the Turkish language alphabet (ç, ı, ş, ö, ü, ğ, İ).

When doing Sentiment Analysis in Turkish, you cannot use language models like Sentiment Analysis in English but all of

the methods used in the Sentiment Analysis literature can be used for the Turkish Sentiment Analysis.

| TURKISH | ENGLISH |
|---------------------|-------------------------|
| Okula gidiyorum. | I'm going to school. |

Fig.1. Sentence structures in Turkish and English

In Sentiment Analysis studies positive, negative or neutral expressions are searched and analyzed in the dataset. The results of the analysis determine the attitudes of the persons or groups in the study. However, if the dataset is large, the classification of opinions cannot be done individually. The most commonly used method for Sentiment Analysis and Opinion Mining analysis is machine learning with text mining.

Preparing the data set for text classification is the most important step in the pre-processing step. In text mining, it is always necessary to deal with words containing noisy and insignificant information. Pre-processing step is made to solve the problems on the dataset, to be able to make more meaningful and quality analyzes by learning the natural structure of the dataset and to generate more meaningful information from the dataset [6].

Data Pre-processing Stage

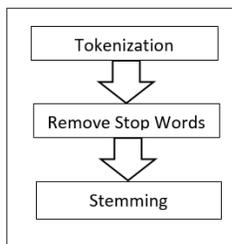


Fig.2. Data pre-processing stage

Pre-processing step is usually carried out in 3 steps as shown in Figure 2.

- Tokenization
- Stop-Word
- Stemming

In the next stage of the qualitative inference and selection process; add-subtract feature, creating a vector space model and attribute selection were made as shown Figure 3.

Feature Extraction

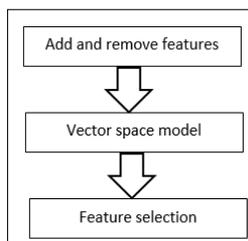


Fig.3. Feature Extraction and Feature Selection Stage

Classification methods are applied for data passing through the text mining stage. The concept of classification is simply to distribute data among the various classes defined on a data set. Classification algorithms learn this distribution form from given training set and then try to classify correctly when the test data arrives that the class is not specified.

The values that specify these classes on the dataset are given a label name and are used to determine the data class during the test, if necessary. Machine learning algorithms are used for classification [7].

Machine learning methods extract features from texts. Some of the methods used in Sentiment Analysis; Naive Bayes (NB), k-NN, Center Based Classifier, Artificial Neural Networks (ANN), Support Vector Machines (SVM).

There are two different approaches for Sentiment Analysis except text mining with machine learning approaches. These are natural language processing (NLP) methods and hybrid approaches.

Using the NLP methods, Sentiment Analysis is performed according to the results obtained after the analysis developed specifically for the language. The hybrid method is a hybrid approach and involves both methods.

When Sentiment Analysis studies go beyond the text-to-label linkage, the studies have turned into open-ended results. This is how the idea of Opinion Mining was formed.

As a result, both Sentiment Analysis and Opinion Mining are research topics for the following common materials;

- Classify personal opinions
- Sentiment classification of texts
- Summarizing thoughts in texts
- Inference of ideas from texts
- Classification of negative, sarcasm or irony-containing statements

II. II. THE STUDIES ON TURKISH SENTIMENT ANALYSIS

Parlar [8] investigated various data preprocessing combinations in Sentiment Analysis of Turkish interpretations and investigated which qualification selection methods were effective results. Sentiment Analysis has proposed a new method of feature selection in selecting the most valuable features. It was tried to improve the accuracy and efficiency characteristics of the Sentiment Analysis process by using qualitative selection methods such as Chi-square, Information Gain, Document Frequency Difference and Optimal Orthogonal Centroid. The proposed method of selection of qualities is compared with these methods. In analyzing Turkish interpretations, it has been concluded that holding certain punctuation marks and ineffective words qualitatively contributes positively and contributes to better results with the quality selection methods used.

Evirgen [9] has proposed a general framework in terms of the fact that Turkish sentences, which express emotions in a

short and concise manner, are not properly formatted so that words can be directly processed and processed by R programming language, and that this is the starting point. It has been stated that there is not yet a robust and usable web or client application on Turkish Sentiment Analysis and that it is a starting point for one step forward in the study.

Kama [10] has implemented Turkish application for blogs and forums, comments and reviews collected by the collection of data created by previously developed for English.

Demirci [11] focused on emotional analysis on microblogs. Automatically detecting emotions in micro-blogs is a new research area which gains importance with the rapid growth of the micro-blogs in the last few years. Mining emotions in micro-blogs have some practical uses which can improve human-computer interaction. As opposed to regular text used in text mining studies, micro-blog entries are short and not well-formed enough to process directly. He has proposed a general framework that special uses, symbols, and facilities that can be used in micro-blogs can greatly affect the emotion of the text. He has created new Turkish Twitter dataset for Sentiment Analysis.

Çelik [12] developed A custom data preparation algorithm for the Turkish language in order to maximize the accuracy. He applied Naïve Bayes, Support Vector Machines, Logistic Regression and Decision Tree machine learning algorithms to the data sets. Naïve Bayes has worked 20 or 30 times faster than his nearest competitor, the Support Vector Machines algorithm. Besides the accuracy, the execution time of Naïve Bayes algorithm has been much faster than the others. Positive comments about everything have been 20-25 times more than negative comments. Accuracy has to decrease while n-gram size increases. Thresholding has generally little positive effect on accuracy but has improved execution time gradually on algorithms other than Naïve Bayes.

Boynukalın [13] has focused on emotion analysis of Turkish texts and it has been shown that using ML methods for Turkish texts on analysis of emotions is feasible and gives promising results. Several methods have been applied to get the best results with Turkish language and experimental results have been evaluated. She has defined a new set of data with using two different sources and this set was used in the study. Different automatic learning techniques have been tried and results are compared. Additions to the methods used due to the features that are separated between Turkish and English were made and the success of the analysis in Turkish texts has been tried to be increased.

Yelmen [14] focused on the selection of attributes from the Turkish texts written in daily conversation language in his work and used support vector machines, artificial neural networks, and centroid based classification algorithms on the detailed pre-processing data. SVM, ANN, and Centroid Based Classification Algorithms have been used on the data that is

processed through detailed pre-processing. Gini Index, Data Gain and Genetic Algorithm have been used as hybrids together with 3 different classification algorithms on Twitter messages belonging to followers of three different GSM operators.

Çoban [15] has studied two categorized sentimental analysis on Turkish messages from Twitter. Sentiment Analysis was considered as a text classification problem; Sentiment Analysis techniques, as well as classical text classification techniques, have been used. Machine learning methods are used to automatically detect the dominant sensation in Twitter messages. In this study, in which both text classification and emotion analysis experiments were performed, the main goal was to increase the success of emotional analysis. For this purpose, the effects of different preprocessing, labeling, classification and similarity methods on the Twitter emotional analysis in Turkish have been examined. In addition, a labeling method based on topical information was proposed and the highest success rate of 92.50% was achieved. Thus, the emotion analysis success can be achieved higher than previous studies.

Kaplan [16] has analyzed Twitter messages of Twitter users who are social media networks. The emotions expressed by shared Twitter messages are classified into four different classes. These classes are; 'Happiness', 'Anger', 'Sadness' and 'Surprise' groups. All typo mistakes of tweets were proofread with the help of 'Zemberek' library for classifying these accurately. Proofread tweets were labeled on these four categories of volunteers. In the study, messages collected from Twitter were analyzed with a decision tree and fuzzy learning techniques. Despite the fact that the fuzzy rules and decision tree methods have very close success rates, it has been found that both methods of emotion categorization give results that cannot give confidence in different categories. Nevertheless, it is suggested that the fuzzy decision method is a more valid method when closely examined in the categories.

Yurt [17] has tested the achievements of emotion analysis using pre-designed NLP algorithms in Turkish texts. Turkish texts were taken from the network environment, pre-processed, analyzed with tools that help in data mining and machine learning and the results are discussed. The study done using other machine learning algorithms has resulted in better results than Naïve Bayes.

Türkmenoğlu [18] has performed Two separate methods of sentimental analysis for Turkish, from machine learning methods and dictionary-based Sentiment Analysis, previously studied for English and Turkish. These methods were applied to two different sets of Turkish data, short and long texts, and their performance was measured. They have carried out preliminary operations considering the structural characteristics of the Turkish language.

Akba [19] has examined emotional analysis methods by

comparing success rates. According to some experimental results, he has tried to establish a system which can respond in a short time and needs less human power. Using data (Turkish movie comments) are interpreted and scored by the users on the website. Furthermore, in the evaluation of the experiment, it was tried to determine the number of the most appropriate words to be used in determining the feelings in the sentences written in Turkish. It was observed that the most successful results were obtained with 375 terms.

Nangir [20] has tried and succeeded in a multi-classifier approach which is an unexamined approach for the Turkish language. He has shown that the performance of the classifiers can be improved by parameter optimization operation and the accuracy of the classification operation can be increased. This success has shown that the multiple classifier approaches is a successful machine learning approach and can be used in many studies. The multiple classifier approaches have been used with three singular classifiers and a majority voting method.

III. CONCLUSIONS

Turkish Sentiment Analysis and Opinion Mining are almost identical concepts. Sentiment Analysis, when looking for emotional words/expressions in a medium, the Opinion Mining outlines and analyzes the opinions of people about an entity.

While Sentiment Analysis is mostly used in industry, Opinion Mining is more frequently used in the academic field. Both concepts work together on a common research theme.

The following solutions are proposed for the problems observed in the Turkish emotional analysis thesis studies;

The lack of any open data sets in Turkish that can be used in the field of emotion analysis can be overcome by the data obtained from social media by various methods.

The difficulties related to the inadequacy of the studies done in the field of emotion analysis in Turkish language, Researchers can be overcome by providing project support from the economic side.

Although it is a Turkish WordNet database, the number of terms in this database is low and the result is that the opinion dictionary to be formed is composed of few terms. This problem can be solved if each researcher who works on emotion analysis increases the number of terms in the database.

In document-based classifications, there is a lack of work that has been reduced to sentence and sub-sentence structures and even subject-based classification. Studies on these issues can be increased.

Correct adjustment of the quality and categorical distribution of the training set has a direct impact on the analysis. The fact that the training data to be used for emotion analysis is sector-based will increase the success of the results of emotional analysis and quality extraction to be achieved.

REFERENCES

- [1] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.
- [2] S.E. Şeker, "Duygu Analizi (Sentimental Analysis)", YBS Ansiklopedi, 2016, pp.21-36.
- [3] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques", Proceedings of EMNLP 2002, pp. 79-86, 2002.
- [4] T. Nasukawa and J. Yi, "Sentiment Analysis: Capturing favorability using natural language processing", Proceedings of the 2nd international conference on Knowledge capture, Sanibel Island, FL, USA, 2003.
- [5] K. Dave, S. Lawrence, and M. P. David, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews", Proceedings of the 12th international conference on World Wide Web, ACM, 2003.
- [6] T. Kaşıkçı, H. Gökçen, "Metin Madenciliği İle E-Ticaret Sitelerinin Belirlenmesi, Bilişim Teknolojileri Dergisi", 7(1), 2014, pp. 25-32.
- [7] S.E. Şeker, *Sınıflandırma*, 08 December 2017. [Online]. Available: <http://bilgisayarkavramlari.sadievrenseker.com>
- [8] T. Parlar, *Feature Selection for Sentiment Analysis in Turkish Texts*, Published Ph.D. Thesis, Adana, 2016.
- [9] E. Evirgen, *Sentiment Analysis of Turkish Tweets*, Published Master's Thesis, İstanbul, 2016.
- [10] B. Kama, *Feature Based Sentiment Analysis on Informal Turkish Texts*, Published Master's Thesis, Ankara, 2016.
- [11] S. Demirci, *Sentiment Analysis on Turkish tweets*, Published Master's Thesis, Ankara, 2014.
- [12] H. Çelik, *Sentiment Analysis for Turkish Language*, Published Master's Thesis, İstanbul, 2013.
- [13] Z. Boynukalın, *Sentiment Analysis of Turkish Texts by Using Machine Learning Methods*, Published Master's Thesis, Ankara, 2012.
- [14] İ. Yelmen, *Sentiment Analysis on Turkish Social Media Data with Natural Language Processing Methods*, Published Master's Thesis, İstanbul, 2016.
- [15] Ö. Çoban, *Turkish Twitter Feelings Analysis with Text Classification Techniques*, Published Master's Thesis, Erzurum, 2016.
- [16] B. A. Kaplan, *Sentiment Analysis on Turkish Twitter Messages by Using Data Mining*, Published Master's Thesis, İstanbul, 2016.
- [17] E. A. Yurt, *Sentiment Analysis in Turkish Documents*, Published Master's Thesis, İstanbul, 2015.
- [18] C. TÜRKMENÖĞLU, *Sentiment Analysis in Turkish Texts*, Published Master's Thesis, İstanbul, 2015.
- [19] F. Akba, *Assesment of Feature Selection Metrics for Sentiment Analysis: Turkish Movie Reviews*, Published Master's Thesis, Ankara, 2014.
- [20] M. Nağır, *Classification of Emotions with Multiple Classifier Methods for Turkish Language*, Published Master's Thesis, İstanbul, 2013.

BIOGRAPHIES



HABİBE KARAYIĞIT received the B.S. degree from the Electronic-Computer Technical Education department, Fırat University, Elazığ, Turkey, in 2002, and the M.S. degrees from the Electrical and Electronics Engineering department, Mustafa Kemal University, Hatay, Turkey in 2014. She is now a Ph.D.

student at the Mersin University. She is currently technical computer teacher and undergraduate student in computer engineering at the İSTE (İskenderun Technical University, İskenderun, Turkey). Her research interests Social Media Analysis, Text Mining, Machine Learning, Web Programming.



ÇİĞDEM ACI received her B.S. degree in Computer Engineering from Fırat University, Elazığ, in 2007. M.S. and Ph.D. degrees in Computer Engineering from the University of Çukurova, Adana, in 2009 and 2013. From 2008 to 2013, she was a Research Assistant with the Department of Computer Engineering at Çukurova University. Between 2013-2015 years, she worked as an Assistant Professor in the Computer Engineering Department, Çukurova University. Since 2015, she has been an Assistant Professor and Deputy Dean in the Faculty of Engineering, Mersin University. Her research areas are Machine Learning, Data Mining, and High-Performance Computing.



ALİ AKDAĞLI received his B.S. degree in Electrical and Electronics Engineering from Erciyes University, Kayseri, in 1995. M.S. and Ph.D. degrees in Electrical and Electronics Engineering from the University of Erciyes, Kayseri, in 1997 and 2002. From 2006 to 2008, he was Assistant Professor in the Department of Electrical and Electronics Engineering at Mersin University. Between 2008-2013 years, he worked as an Associate professor in the Electrical and Electronics Engineering Department, Mersin University. Since 2013, he has been a Professor and Dean in the Faculty of Engineering, Mersin University. His research areas are Embedded Systems, Wireless Communication Systems, Computer Networks and Security, Intelligent Antennas, Microwave Technology, RFID.