Sentiment Analysis of Turkish Twitter Data

Hatice Kübra Erol
Galatasaray University, Turkey
Email: haticekubrae@gmail.com

***Abstract*—In this paper, we examined sentiment analysis, sentiment analysis in Turkish texts, sentiment analysis with social media data. Then we developed an application where we used machine learning methods for sentiment analysis. In this application, we used a data set of tweets (user comments about gsm-operators) which are labeled negative or positive. After pre-processing the data, we reserved eighty percent of the data for training our model and the rest for testing. It is used 3 different statistical models to train and test the data. These are SVM, Naive Bayes and Logistic Regression models. After applying the models, we calculated the f1Score values and compared our results. According to F1 scores, we got the best results when we used the SVM model.**

Keywords: Sentiment analysis, Twitter, Emotion analysis, Turkish Sentiment analysis.

I. Introduction

Sentiment analysis is the process of detecting positive, negative or neutral sentiment in text. It has become a very popular research area since automatic extraction of the sentiment can be very useful in analyzing what people think about specific issues or items, by analizing large collections of textual data sources such as personal blogs, product review sites, and social media [7].

Twitter, one of the social media channels, contains a lot of easily accessible and processable data. Users of the platform, write about their daily life, their opinions about economics, policy, companies and their products. Especially, businesses interests to detect sentiment in social data to understand their customer's opinion, general trends etc. [6]. As a result, the analysis of Twitter data is a hot topic.

In this study, sentiment analysis will be researched, an application on sentiment analysis will be developed by using Turkish Twitter data and the results will be discussed.

II. A Brief Overview on Sentiment Analysis

Sentiment analysis is a natural language processing (NLP) technique that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources. Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It's also referred as subjectivity analysis, opinion mining, and appraisal extraction. [11]

A. Sentiment Analysis Levels

Sentiment analysis can occur at different levels; Sentence level Sentiment Analysis, Document level Sentiment Analysis, Feature based Sentiment Analysis.

a) Sentence level: It deals with tagging individual sentences with their sentiment. General approach is finding the sentiment polarities of individual sentences or words and combine them together to find the polarity of the document.

b) Document level: It deals with tagging individual documents with their sentiment. General approach is finding the sentiment polarities of individual sentences or words and combine them together to find the polarity of the document. [11]

c) Feature level: It deals with labeling each word with their sentiment and also identifying the entity towards which the sentiment is directed.

B. Sentiment Analysis Approaches

There exist three fundamental approaches for sentiment analysis: lexicon-based approach, machine learning based approach and Hybrid Approach. The first approach has the advantage of being simple, while the second approach is typically more successful since it learns from samples of documents in with known sentiment, in the given domain [7], [1].

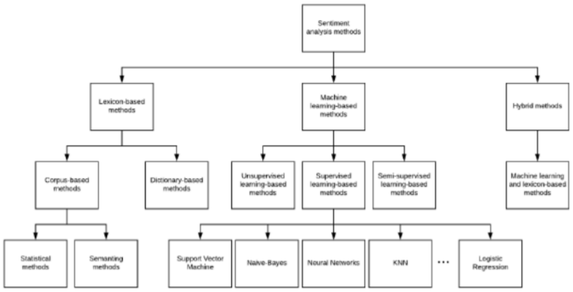

Sentiment analysis methods
Lexicon-based methods
Machine learning-based methods
Hybrid methods
Corpus-based methods
Dictionary-based methods
Unsupervised learning-based methods
Supervised learning-based methods
Semi-supervised learning-based methods
Machine learning and lexicon-based methods
Statistical methods
Semanting methods
Support Vector Machine
Naive-Bayes
Neural Networks
KNN
...
Logistic Regression

Fig. 1: Classification of SA methods (Panday et al., 2017)

In the paper named 'Sentiment Analysis in Turkish Media' [5], the author discussed the comparison of machine learning based experiments and lexicon based experiments, mentioning that machine learning based experiments give higher accuracy rate.

C. Sentiment Analysis In Turkish

1) Difficulties due to Turkish language structure: There is a lot of research on the normalization of social media data for the English language. However, it is known from previous research results that the methods prepared for English are not compatible with Turkish because of the different language structure of Turkish [7]. Therefore, there is a lack of resources for Turkish.

It is known that Turkish is a difficult language for Sentiment analysis. List of the compelling features of Turkish as follows [7]:

a) Agglutinative Morphology: : Turkish is an agglutinative language: it is possible to generate new and arbitrarily long words by adding suffixes to a root word. These derivational and inflectional suffixes may change the Part-of-Speech (POS) tagging and semantic orientation of the word (e.g. “beğenme” (to like or don’t like), “beğendim” (I liked it), “beğenmedim” (I didn’t like it)). The practical limitation of the agglutinative morphology in speech, handwriting, or sentiment analysis problems is that it makes it infeasible to build a (polarity) lexicon that would need to contain all variants of Turkish words. Hence, sentiment analysis systems for agglutinative languages Sentiment Analysis in Turkish 7 like Turkish face some extra challenges compared to those for which a reasonable size lexicon (e.g. 30,000 as for English) is sufficient for many applications.[7]

b) Negation: : In Turkish, there are many ways a word may be negated in a way that its sentiment polarity will change: with the affixes me/ma or siz/sız (as in “olmadı” (didn’t happen), “başarısız” (unsuccessful)); or using a separate word such as “değil” or “yok” (as in “güzel değil” (not beautiful) or “konusu yok” (didn’t have a topic)). [7]

c) Turkish characters: : Turkish Alphabet: Turkish has several characters that do not exist in the English alphabet: “¸c”, “˘g”, “ı”, “¨o”, “¸s”, “¨u”. In informal writing, people tend to substitute these Turkish letters for the closest ASCII characters (e.g. “¸c” is written as “c”), which complicates the mapping of a string to the words. Thus, one needs a preprocessing step before sentiment analysis known as de-ASCIIfication (i.e., converting the ASCII English characters to their Turkish equivalents, to find the words and obtain their polarities from the lexicon). [7]

D. Sentiment Analysis On Social Media

Social media channels are virtual spaces where to express and share individual opinions, influencing any aspect of life, with implications. Many companies use social media data for marketing and brand success tracking. It also contains data that can be considered valuable for fields of study such as psychology and economics.

One of the most frequently used social media channels in Sentiment analysis is Twitter. While working with Twitter data, it is necessary to consider the Tweet’s data structure. Some of the pecularities of tweets can be listed as follows:

- A Tweet can contain maximum of 140 characters. This situation causes users not to pay attention to grammar and to use abbreviations too much.
- Users are marked with the @ sign.
- In general, topics that is discussed in the tweets are marked with the hashtag sign. Thanks to this, there is no need to employ very complex linguistic tools to determine it.
- Tweets are user generated texts. So, the language employed in subjective tweets includes a specific slang (also called “urban expressions” 2) and emoticons (graphical expressions of emotions through the use of punctuation signs).
- Twitter is available in more than 30 languages.
- In major events, the rate of tweets per minute commenting or retweeting information surpasses the rate of thousands per minute.

III. Related Works

[11] presents a survey on sentiment analysis of Twitter Data. It was seen as a valuable resource for understanding Sentiment analysis and drawing a roadmap for the study.

In [1], a sentiment analysis (SA) was conducted on Turkish tweets collected about an ODE system to monitor students’ views and feelings about the system. This publication is a study similar our work. It is used as one of the main resources.

[7] provides an overview of the sentiment analysis problem and combines supervised learning and dictionary-based approaches by making use of the Turkish polarity dictionary called SentiTurkNet. It reviewed as a useful Turkish study example.

[2], [3], [4],[5],[6] [9], [10] are also found as useful resources for Turkish sentiment analysis and Twitter data analysis.

In addition, similar studies in resources such as github.com and kaggle.com were also examined during the study. Reviewed source are [12], [13], [14], [15]

IV. Methodology

In this study, we present a sentiment analysis on a data set which contains labeled tweets. We used machine learning methods for the analysis.

A. Tools

The work was carried out using python on Jupyter Notebook. Some libraries used are pandas, sklearn, nltk, numpy, BeatifulSoup, TurkishStemmer.

B. Flow

The steps to be applied in the study are as follows. Also, the flowchart of the proposed methodology is given in Fig. 2.

1) Twitter data will be acquired. If it is not labeled, it will need to be labeled positive or negative.
2) The data will be pre-processed with data cleaning, normalization, tokenization, stemming processes.
3) The text to be analyzed will be digitized.
4) The training and test data will be separated.
5) Classification models will be applied to the separated data.
6) The results obtained will be presented by plotting.

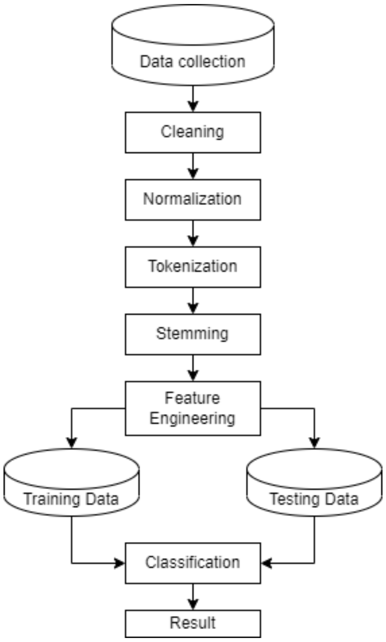


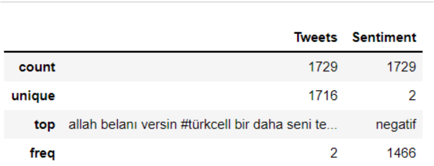
Fig. 2: Flowchart of the proposed methodology

V. Application

A. Data Collection

Twitter data can be collected through the API provided by Twitter. To be able to connect the API, it is necessary to have access keys. To obtain these keys, it is necessary to apply for a Twitter Developer Account.

In this study, since we could not get Twitter authentication keys, we obtained the data from a different source. It is https://www.kaggle.com/ . It contains datasets created by some researchers. The dataset we selected from kaggle consists of tweets about gsm operators. It has two columns; one of them contains tweets and the other contains emotion labels. Emotion labels are ‘negatif’ and ‘pozitif’.



	Tweets	Sentiment
count	1729	1729
unique	1716	2
top	allah belanı versin #türkcell bir daha seni te...	negatif
freq	2	1466

Fig. 3: Information about data

B. Data Preprocessing

Before proceeding to the modeling phase, the dataset needs to be improved. There are 4 processes that we will perform for this: Cleaning, normalization, tokenization, stop word removal.

1) Cleaning: At this stage, we deleted the data with null values.

2) Normalization: Normalization is the process of converting all text to the same case, eliminating punctuations, converting numbers to words, and so on [1]. In this step,

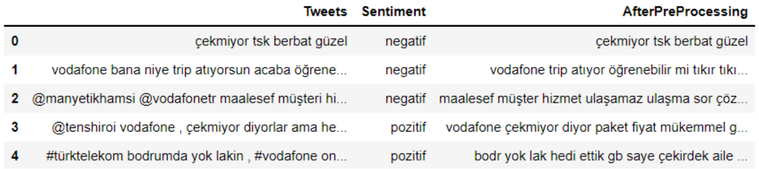
- We removed urls from tweets.
- We converted the tweets to lower case considering Turkish characters.
- We removed usernames and topics from tweets. We could find them searching with @ and hashtag signs.
- We removed punctuations from tweets.
- We removed numeric characters from tweets.

3) Tokenization: For tokenization, we used trtokenizer library. In the tokenization step, tweets were split into words as a token by whitespaces.

4) Stop Word Removal: Stop word removal is one of the most commonly used preprocessing steps to reduce the vector space and enhance the classifier performance. Stop word file is imported from kaggle ([16]). it contains 231 Turkish words.

5) Stemming: We used TurkishStemmer library for stemming operation. Details about library can be found in [17]

After the completion of the processes, the change in the data was observed. A small piece of data is listed in Fig. 3.



	Tweets	Sentiment	AfterPreProcessing
0	çekmiyor tsk berbat güzel	negatif	çekmiyor tsk berbat güzel
1	vodafone bana niye trip atıyorsun acaba öğrene...	negatif	vodafone trip atıyor öğrenebilir mi tıkır tıkı...
2	@manyetikhamsi @vodafonetr maalesef müşteri hi...	negatif	maalesef müşter hizmet ulaşamaz ulaşma sor çöz...
3	@tenshiroi vodafone , çekmiyor diyorlar ama he...	pozitif	vodafone çekmiyor diyor paket fiyat mükemmel g...
4	#türktelekom bodrumda yok lakin , #vodafone on...	pozitif	bodr yok lak hedi ettik gb saye çekirdek aile ...

Fig. 4: A small piece of data after preprocessing

C. Feature Engineering

Text needs to be represented as numerical feature vectors before applying machine learning algorithms. Bag of words (BoW), term frequency-inverse document frequency (TF-IDF) are used as vector space models in the study.

BoW model generates a feature vector that contains the counts of each unique word in the text without using semantics and order of the words.[1]

Used countvectorizer parameters: **mindf = 0, maxdf=1, ngramrange=(1,3), binary=False**

TF-IDF is a well-known method to evaluate a word’s importance in a text and multiplication of term frequency (TF) and the inverse document frequency (IDF). (TF) of a particular term (t) is calculated as the number of times a term occurs in a document to the total number of words in the text (Ahuja et al., 2019). IDF is the log of the inverse probability of a term being in the text. [1]

Used TfidfVectorizer parameters: **mindf = 0, maxdf=1, ngramrange=(1,3), useidf =True**

D. Training and Testing

Supervised learning is an important technique for solving classification problems. Training the classifier makes it easier for future predictions for unknown data. [11] We split 20 percentage of the data as test data and trained the rest.

We used 3 different statistical models to training and testing: Logistic Regression, SVM, Naïve Bayes. After training and testing, then we calculated F-score values for each model. We then compared the results.

The models used are detailed below.

1) Logistic Regression: Logistic Regression (LR) is a statistical classification model for the prediction of a binary categorical variable. The aim of Logistic Regression is to find a relationship between features and the probability of the outcome [1]

2) Support Vector Machines: SVM is a used optimization-based supervised machine learning algorithm for classification and regression. The aim of the SVM is to find a hyperplane in N-dimensional space (N-the number of features) that correctly classifies the data set by maximizing the margin between the two classes. [1]

3) Naive Bayes: Naive Bayes is a probabilistic classifier, meaning that for a document d, out of all classes c C the classifier returns the class c which has the maximum posterior probability given the document. [18]

E. Evaluation

The performances of the sentiment classifications were evaluated by calculating F1-score. Formulations to calculate the score are given below. TP, FN, FP and TN refer respectively to the number of true positive instances, the number of false negative instances, the number of false positive instances and the number of true negative instances,

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

F1 = (2×Precision×Recall)/(Precision+Recall)

To calculate the F1-score, confusion matrices are constructed.

F. Computational Results

F-score of classifiers according to TF-IDF and BoW model is in below table.

1) Logistic Regression :

	precision	recall	f1 Score
negative (BoW)	0.86	1.00	0.92
positivw (BoW)	1.00	0.04	0.08
negative (TF-IDF)	0.85	1.00	0.92
positive (TF-IDF)	0.00	0.00	0.00

2) SVM:

	precision	recall	f1 Score
negative (BoW)	0.87	0.99	0.93
positivw (BoW)	0.67	0.16	0.25
negative (TF-IDF)	0.86	0.99	0.92
positive (TF-IDF)	0.71	0.10	0.17

3) Naive Bayes:

	precision	recall	f1 Score
negative (BoW)	0.85	1.00	0.92
positivw (BoW)	1.00	0.04	0.08
negative (TF-IDF)	0.85	1.00	0.92
positive (TF-IDF)	0.00	0.00	0.00

When we compare the results of the 3 models, it is observed that the best result is obtained with the SVM model. In the comparison of Bow and TFIdf, it is observed that the BoW calculation gives better results.

G. Future works

We can say that we did not get good results with these models. The high f1-scores of negative values are due to the fact that the estimation assigns too many 0s.

It is thought that this result is due to the small size of our data set and the difficulty of the Turkish language. Better results can be obtained by using a large set and applying polarization and using a hybrid method.

VI. Conclusion

Sentiment analysis,sentiment analysis in Turkish texts, sentiment analysis with social media data, sentiment analysis on Twitter data were investigated in this study. Related studies reviewed. Then, a sentiment analysis application was studied with a machine learning approach using a data set consisting of tagged Turkish tweets obtained from a website. Although a successful result cannot be obtained in practice, it can be considered as a useful study for an initial work. In the next study, it is planned to apply a hybrid method, to focus on the use of the appropriate model and parameterization, taking into account the data set.

References

[1] Z.Aydın and Z. Ozturk and Z. Cıcek, *TURKISH SENTIMENT ANALYSIS FOR OPEN AND DISTANCE EDUCATION SYSTEMS*, Turkish Online Journal of Distance Education-TOJDE, July 2021.

[2] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede *Lexicon-Based Methods for Sentiment Analysis* Association for Computational Linguistics

[3] G. Yurtalan, M. Koyuncu and Ç. Turhan A polarity calculation approach for lexicon-based Turkish sentiment analysis Turkish Journal of Electrical Engineering Computer Sciences, 2019.

[4] G. Eryiğit, T.Temel, İ. Çiçekli, M. Yanık, F.S. Çetin *TURKSENT: A Sentiment Annotation Tool for Social Media* Association for Computational Linguistics, 2013

[5] C. Turkmenoglu, A.C. Tantuğ *Sentiment Analysis in Turkish Media* Istanbul Technicla University

[6] G. Eryiğit, E. Yıldırım, T. Temel, F.S. Çetin *The Impact of NLP on Turkish Sentiment Analysis*, İstanbul Technical University .

[7] G. Gezici, B. Yanikoglu *Sentiment Analysis in Turkish*, Sabancı University

[8] G. Eryigit, D. Torunoglu *A Cascaded Approach for Social Media Text Normalization of Turkish*, Istanbul Technical University

[9] O. Kolchyna, T.P. Souza, P. C. Treleaven and T. Aste *Methodology for Twitter Sentiment Analysis*, Department of Computer Science, UCL, Gower Street, London, UK

[10] R. DEHKHARGHANI *SENTIMENT ANALYSIS IN TURKISH: RESOURCES AND TECHNIQUES* , SABANCI UNIVERSITY, 2015

[11] Kharde, Vishal Sonawane, Sheetal. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications. 139. 5-15. 10.5120/ijca2016908625.

[12] https://www.kaggle.com/rolmez/twitter-sentiment-analysis1.-Introduction

[13] https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/

[14] https://towardsdatascience.com/step-by-step-twitter-sentiment-analysis-in-python-d6f650ade58d

[15] https://medium.com/@bernazeyrekk/makine-

[16] https://www.kaggle.com/rolmez/turkce-stop-words

[17] https://github.com/skroutz/turkishstemme

[18] https://web.stanford.edu/ jurafsky/slp3/4.pdf