Sigma J Eng Nat Sci, Vol. 42, No. 4, pp. 1292–1314, August, 2024

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2024.00033

Review Article

Turkish sentiment analysis: A comprehensive review

Ayşe Berna ALTINEL GİRGİN[1,*], Gizem GÜMÜŞÇEKİÇÇİ[2], Nuri Can BİRDEMİR[1]

[1]Department of Computer Engineering, Faculty of Technology, Marmara University, İstanbul, 34722, Türkiye
[2]Department of Computer Engineering, Işık University, İstanbul, 34980, Türkiye

ARTICLE INFO

Article history
Received: 17 July 2023
Revised: 18 November 2023
Accepted: 29 December 2023

Keywords:
Deep Learning; Machine Learning; Natural Language Processing; Polarity Detection; Sentiment Analysis; Sentiment Classification

ABSTRACT

Sentiment analysis (SA) is a very popular research topic in the text mining field. SA is the process of textual mining in which the meaning of a text is detected and extracted. One of the key aspects of SA is to analyze the body of a text to determine its polarity to understand the opinions it expresses. Substantial amounts of data are produced by online resources such as social media sites, blogs, news sites, etc. Due to this reason, it is impossible to process all of this data without automated systems, which has contributed to the rise in popularity of SA in recent years. SA is considered to be extremely essential, mostly due to its ability to analyze mass opinions. SA, and Natural Language Processing (NLP) in particular, has become an overwhelmingly popular topic as social media usage has increased. The data collected from social media has sourced numerous different SA studies due to being versatile and accessible to the masses. This survey presents a comprehensive study categorizing past and present studies by their employed methodologies and levels of sentiment. In this survey, Turkish SA studies were categorized under three sections. These are Dictionary-based, Machine Learning-based, and Hybrid-based. Researchers can discover, compare, and analyze properties of different Turkish SA studies reviewed in this survey, as well as obtain information on the public dataset and the dictionaries used in the studies. The main purpose of this study is to combine Turkish SA approaches and methods while briefly explaining its concepts. This survey uniquely categorizes a large number of related articles and visualizes their properties. To the best of our knowledge, there is no such comprehensive and up-to-date survey that strictly covers Turkish SA which mainly concerns analysis of sentiment levels. Furthermore, this survey contributes to the literature due to its unique property of being the first of its kind.

Cite this article as: Altınel Girgin AB, Gümüşçekiççi G, Birdemir NC. Turkish sentiment analysis: A comprehensive review. Sigma J Eng Nat Sci 2024;42(4):1292–1314.

INTRODUCTION

Sentiment Analysis (SA), also referred to as Opinion Mining (OM), encompasses the process of contextually mining the text which includes detecting, identifying, and extracting properties [1]. To realize and automate sentiment analysis, NLP is used. Sentiment analysis has several applications. For instance, it can be used to determine the polarity of a text, meaning that it can be categorized as positive,

***Corresponding author.**
*E-mail address: berna.altinel@marmara.edu.tr

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

Published by Yıldız Technical University Press, İstanbul, Turkey

neutral, or negative, and to identify individuals' opinions, attitudes, and emotions towards an entity or an event [2]. Since sentiment analysis is such a versatile tool, it can be applied in a variety of fields, including marketing, consumer information, politics, and social networks. An increasing amount of people have started to share their opinions and ideas about significant concepts, events, situations, etc. on social networks, which has led to the vast popularity of SA as a research topic [1]. Anecdotally, some of the most widely used social networks are Twitter, Facebook, Instagram, etc. Collecting data is a significant issue for every type of study. Studies that use more data provide more realistic and accurate results but finding proper data is generally challenging. But the increase in the usage of social networks provides significant sources of versatile data for sentiment analysis and this can be considered a reason behind its popularity. The general process of implementing a sentiment analysis model is given in Figure 1.

In this survey, studies about Turkish sentiment analysis are collected, analyzed, and summarized. All the collected studies are analyzed and summarized in a similar structure which consists of their approach, methodology, and performance. This survey categorizes studies considering the approaches employed and their level of sentiment analysis. There are three primary categories of sentiment analysis. These are Dictionary-based sentiment analysis, Machine Learning-based, and Hybrid-based sentiment analysis. This study also compares different types of SA studies and forms a table accordingly within each category based on the level of sentiment analysis, amount of data used and accuracy. The comparison tables are presented for each category. The advantage of these comparison tables are their efficiencies for presenting the main properties of studies in each category from different aspects. This survey also presents a list of public datasets that can be accessed and used for future sentiment analysis research. This list includes, the type of dataset, the year it was created, its language, size and source link can be found. Additionally, the list of the most used lexicons including their language, the size and information on sentiment polarity is also presented in this survey. This research can be useful for future researchers who are interested in Turkish Sentiment Analysis as it covers many different applications of Turkish Sentiment Analysis in one publication. This survey uniquely categorizes different approaches used in Turkish Sentiment Analysis. The contribution of this survey is significant for various reasons. Firstly, this survey reviews and summarizes a large number of previous and recent articles according to used

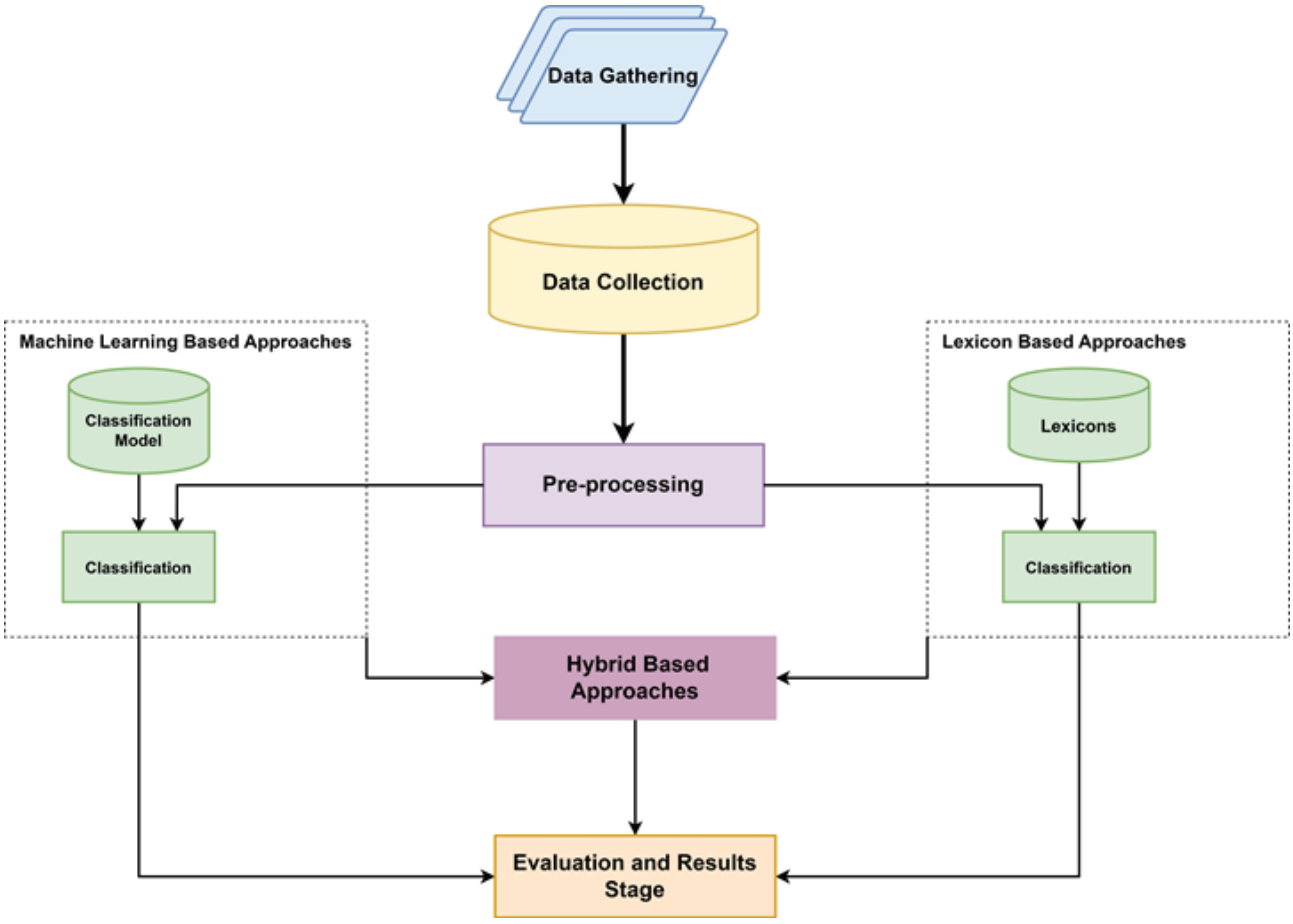

Figure 1. General Process of Turkish Sentiment Analysis.

approaches. This can help interested researchers to easily access the properties of various Turkish Sentiment Analysis studies from past to present and maybe choose the best approaches and techniques for specific research or applications. Second, comparisons of studies from each category are illustrated using tables. This helps visualize the different aspects of studies and allows the easy detection of studies' main features. Third, public datasets and the most frequently used lexicons are collected and listed, which can be accessed by interested researchers to source new research.

This survey is organized as follows: In subsection 1.2, the challenges of Turkish sentiment analysis are presented. In subsection 1.3, an overview of Turkish sentiment analysis is provided. In section 2, the studies using dictionary-based approaches in sentiment analysis are explained. In section 3, the studies using machine learning-based approaches in sentiment analysis are explained. In section 4, the studies using hybrid-based approaches in sentiment analysis are explained. In section 5, different methodologies for constructing a sentiment analysis system are explained. In section 6, the most popular dictionaries and datasets used in Turkish sentiment analysis are presented. In section 7, a general discussion is made about sentiment analysis, including the key points to consider when implementing a sentiment analysis system and current challenges in the field. In section 8, the conclusion of this research is presented.

The Challenges in Sentiment Analysis and Comparison of Turkish Sentiment Analysis with Different Languages

Since sentiment analysis is a language-dependent process, the degree of difficulty associated with performing sentiment analysis in different languages varies. Many linguistic or other types of issues can occur while performing sentiment analysis [3]. In sentiment analysis tasks, preprocessing techniques are generally applied to the data. In the preprocessing stage, the text can be normalized, stemmed, lemmatized, etc. The agglutinative structure of Turkish makes the preprocessing stages generally difficult which complicates Turkish sentiment analysis. In Turkish, the form of a word changes by attaching suffixes to the base (root) word which can change the semantic orientation of the word. This also creates additional challenges in Turkish sentiment analysis [4,5]. In contrast, the English language has a relatively low degree of complexity and inflection, which makes it easier to perform sentiment analysis. Due to the morphology of the Turkish language, creating or finding a proper sentiment lexicon that includes all variants of words can be impossible compared to other languages such as English. Turkish has a large number of unique words and idiomatic expressions that may not be found in other languages. This can make the sentiment analysis process more difficult since the polarities or meanings of these expressions cannot be extracted automatically [3]. Another challenge is that some Turkish characters do not exist in the English alphabet. The transformation of those characters adds difficulties to the sentiment analysis process [4,5,6]. Another important challenge in Turkish sentiment analysis is the limited resources in sentiment lexicons. There may be fewer annotated Turkish texts available for training and evaluating sentiment analysis algorithms compared to English texts [5,6,7]. In addition to the limited resources, the size of the already-limited Turkish lexicons is narrow. Since sentiment analysis is language-dependent, lexicons used in sentiment analysis differ by the language of the sentiment analysis. There are a wide variety of sentiment lexicons available for use in other languages, especially English [6,7]. Lastly, the way that people express sentiment and the words and phrases they use to do so can vary across cultures. Therefore, variations in the way sentiments are expressed can impact the methods of conducting sentiment analysis in different languages. Due to all of these reasons, Sentiment Analysis is most popularly employed in English due to its convenience and there is very limited research published on SA in other complex languages such as Turkish, French etc. [6]. To summarize, the most important challenges of performing Turkish Sentiment Analysis are;

- Language dependency of the sentiment analysis process.
- Language complexity of Turkish.
- Linguistic issues due to the structure of the Turkish language.
- Vocabulary and idiomatic expressions present in Turkish.
- Word Level Sentiment Analysis: Sentiment analysis is conducted on individual words.
- Limited resources in Turkish sentiment lexicons (Annotation availability).
- Narrow capacities of available sentiment lexicons in Turkish.
- Cultural differences between different languages.

The Overview of Turkish Sentiment Analysis

Sentiment Analysis can be performed on many different levels using many different approaches. The levels that the sentiment analysis is performed are aspect level, document level, sentence level and word level.

Aspect Level Sentiment Analysis: In aspect level sentiment analysis, keywords are chosen as entities and sentiment analysis is performed accordingly. Aspect-level sentiment analysis is also referred to as targeted sentiment analysis. Aspect-level sentiment analysis operates under the premise that sentiment is dependent on entities.

Document Level Sentiment Analysis: Document-level sentiment analysis involves analyzing the overall sentiment of an entire document, which may be composed of multiple sentences. The goal of this type of analysis is to understand the sentiment of the entire document.

Sentence and Document Level Sentiment Analysis: Sentence-level sentiment analysis involves analyzing the sentiment of a single sentence, which is made up of multiple words. The goal of this type of sentiment analysis is to understand the sentiment of the entire sentence.

Word Level Sentiment Analysis: Sentiment analysis is conducted on individual words. In Sentiment Analysis, different approaches can be used. These approaches are categorized as dictionary-based, machine learning (ML)-based, and hybrid-based.

Dictionary-Based Sentiment Analysis Approaches: An external lexicon is used to perform sentiment analysis.

ML-Based Sentiment Analysis Approaches: Supervised machine learning, unsupervised machine learning, or deep learning methods are used to perform sentiment analysis.

Hybrid-Based Sentiment Analysis Approaches: In hybrid approaches, instead of using one approach, combinations of different approaches are used.

The overview of sentiment analysis is presented in Figure 2.

TURKISH SENTIMENT ANALYSIS WITH DICTIONARY BASED APPROACHES

In this section, Turkish sentiment analysis studies that use dictionary-based approaches are collected, analyzed, and reviewed. Sentiment analysis can be realized at a variety of levels. These are word, sentence, aspect, and document levels. Regardless of the level of sentiment analysis, the dictionary-based approach relies on a sentiment lexicon and a collection of known and precompiled sentiment terms [8]. This means an external lexicon is utilized to perform the sentiment analysis. Generally, if the lexicon used contains polarity scores of words, the word polarities are used to calculate the sentiment value. The lexicons that contain information about word polarities are called polarity lexicons. Additionally, there are important usages of lexicons in sentiment analysis other than encompassing word polarities.

Aspect Level Sentiment Analysis

This section provides an evaluation of a study that applies dictionary-based approaches to perform an aspect-level Turkish SA, focusing on its approaches, methodology, dataset and performance. In aspect level SA, which is also referred to as targeted SA, special keywords are selected as entities from the given text and the SA is performed towards the targeted entities.

Dehkharghani et al. [3] studied SA in and Turkish sentiment analysis was performed at different granularity levels. In this study, a comprehensive sentiment analysis system was built for Turkish. In the study, a large dataset was used, which was collected from various Turkish movie review websites. The dataset that is used contains 60,000 documents. They only used a subset from the Turkish movie dataset. First, they started by manually labeling 2,700 sentences and 1,000 randomly selected documents. This label annotation process was performed manually by three people. They labeled the data as “positive”, “negative”, or ”neutral”. Following the process of labeling, the distribution of labels for sentences was 50% positive, 30% neutral, and 20% negative, and for documents, it was 52% positive, 29% neutral, and 19% negative. Since they made sentiment analysis

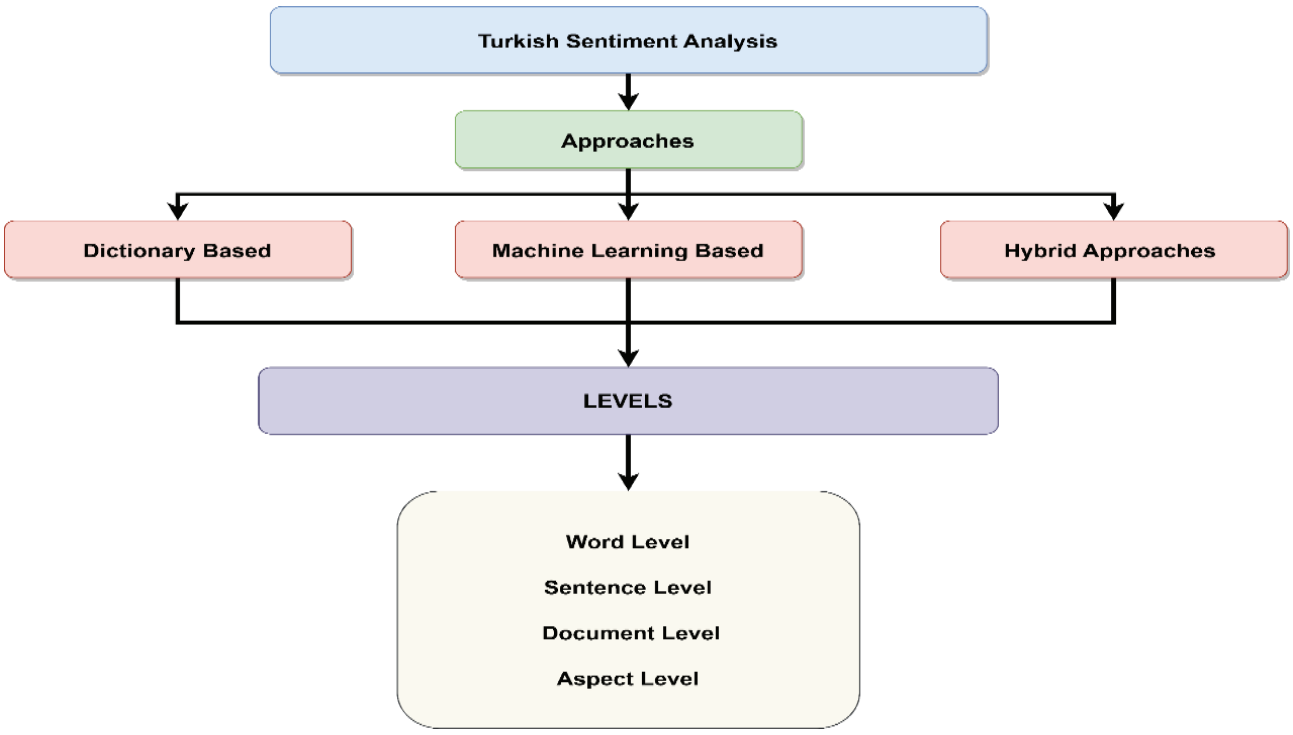

Figure 2. Overview of Turkish Sentiment Analysis.

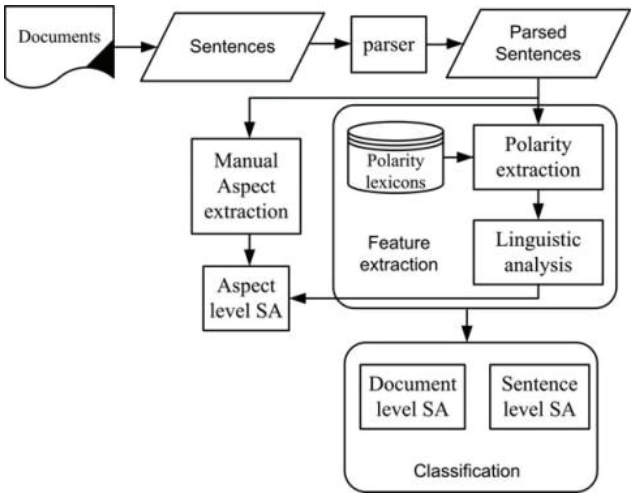

Figure 3. Overview of the system.

at different granularity levels, the proposed system has several different components. The different granularity levels are word level, aspect level, and lastly sentence and document levels. First, the document was segmented into sentences, then they used tokenization to parse the sentences. Following this process, they used a morphological analyzer tool for each word. At the end of this process, they assigned polarity scores to the words' n-grams which are unigram, bigram, and trigram. For the processes, they used different data and NLP tools. They used the ITU Turkish Parser for tokenization and morphological analysis. The SentiTurkNet lexicon was used for finding the polarity scores of words. For the classification, the Logistic Regression classification method was trained and used for different granularity levels. For sentence-level classification, 16 different features were used to train the classifier, for document classification, 20 features were extracted and used and for aspect- level classification, average polarity and number of tokens were used as features. Specifically, word polarities, polar words, sentence types, sentence polarities, emoticons, and linguistic issues such as the number of adjective verbs, number of initial capital words and number of domain-specific indicative terms were used as features. When the proposed system is applied to the chosen data, by utilizing all of the features, they obtained 73.42% and 79.06% accuracy in sentence and document classification. In future works, it will be possible to resolve every issue and sub-problem associated with Turkish sentiment analysis. Additionally, it can be attempted to extend the complexity of the proposed system in the future by exploring the phrase-level sentiment analysis thoroughly. Furthermore, the sentiment analysis process of the study [3] is given in Figure 3.

Sentence and Document Level Sentiment Analysis

In this section, studies on Turkish SA on the sentence-level are reviewed based on methodology, dataset, and performance. All the reviewed studies below use dictionary-based approaches. In sentence-level SA, the sentiment of sentences is extracted based on the information obtained from the whole sentence.

Another recent study had been proposed by Suat and Çınar in [9], wherein the relationships between news about companies and company values were analyzed in the year 2014 using text mining and sentiment analysis. Company publications, news in the media, and social media were used as data sources. In the study, it was stated that in the digitization of textual data, an approach based on sentiment dictionaries would be used to detect the sentiment contents of textual data. For the analysis stage of this study, a polarity lexicon was chosen and used to perform sentiment analysis. The analysis was performed at document-level. According to sentiment analysis, it has been detected that company values are also affected by perceptions and prejudices, unlike priorities such as market investment. This suggests that the quality with which work is accomplished in a company does not hold much significance on its own, without proper marketing and advertising, the real value and effects of the work cannot be reached within the company. No matter how well companies perform, if the performed work cannot be promoted effectively and supported by advertisements, it won't make a big impact and may be forgotten in a short amount of time.

In a new approach proposed by Dehkharghani et al. in [3], sentiment analysis was performed on different granularity levels. The levels that the sentiment analysis performed were aspect level, word level, sentence level, and document level. The polarity of the sentences was extracted from documents and classified with a lexicon-based approach. The details of this sentence-level polarity classification can be found under section 2.1.

The authors in [10] attempted to develop a sentiment lexicon using existing approaches. In order to create the sentiment lexicon, Twitter data was used to conduct sentiment analysis. Sentiment analysis was realized for two distinct topics; the first topic was the effects of weather change on people's feelings. The second topic was to analyze the feelings of people regarding a specific tv-show. To perform sentiment analysis on these two different topics, lexicon-based approaches were used. The details of this study can be found under section 5.

The authors presented a new framework for sentiment analysis classification in [11], using a relatively large Turkish movie site dataset that consists of 60k movie reviews. The dataset is collected from a website called Beyazperde. To create this framework, the authors have customized the SentiStrength sentiment analysis library and used it accordingly. First, the SentiStrength sentiment lexicon is translated into Turkish to be used in sentiment analysis. Then, this newly formed Turkish sentiment lexicon was used to classify the polarities of Turkish movie reviews. The system was evaluated depending on the accuracy scores. The accuracy was calculated by the ratio of the number of reviews

that have the correctly predicted polarity scores and the number of reviews that have incorrectly predicted scores.

In [12], Çoban et al. proposed a sentiment analysis system to classify the polarity scores of tweets. First, the data was gathered and prepared. For the intended purpose, the words in the dataset were stemmed and the stop words removed. Lastly, term frequency methods were used to determine which words in the dataset are more dominant. In this study, three different term frequency methods, TF (Term Frequency), Boolean, and TF-IDF (Term Frequency-Inverse Document Frequency). The authors used both machine learning-based and lexicon-based approaches were separately used to perform sentiment analysis. In the lexicon-based approach, the sentiment analysis was performed on the sentence level using the polarity information of each word in the sentence. The proposed system using the lexicon-based approach achieved a 77.1% accuracy rate while the proposed system that uses a machine learning-based approach to perform sentiment analysis achieved 89% accuracy. A detailed review of this study can be found under section 3.2.

The primary purpose of the study conducted by Albayrak et al. [13] was to analyze and interpret ideas on Twitter, where people share their opinions on certain issues. The tweets that were posted with the hashtag "bedelliaskerlikgeliyor" were collected from Twitter using the "TwitterAPI" with the "Tweepy" library. From this process, a dataset containing 12739 pieces of data was created from the tweets posted on this subject between October 10-12, 2017, and the data were preprocessed using the" NLTK" library. For preprocessing, the punctuation marks, etc. were removed from the dataset. The remaining words in the dataset were analyzed and interpreted by using the" SentiTurkNet" sentiment analysis dictionary, and how people felt about the hashtag topic were analyzed. According to the results, the benefits obtained by combining data analysis and sentiment analysis were determined. One of the advantages of the study is that people's thoughts can be identified about a certain subject with a solution model that is more suitable to the sociological sensitivity of the public.

Karaöz and Gürsoy proposed a new approach in [14] with the aim of indicating that people who work in the social science field with no sufficient software knowledge can also perform sentiment analysis on relatively large data. For this purpose, sentiment analysis was performed by using two dictionaries. Dictionaries were used to determine the polarity values of the words in the dataset. The data used in the study consists of different tweets which are collected over a period longer than eight-months about a TV channel. During this period, a total of 1,200,000 tweets were collected. In this study, only R language and Excel-vba were used. First, the dataset was divided to test and train the system. The division of the used dataset was as follows: 80% for testing of the system and 20% for the experiment. According to the results of this study, the average accuracy rate is calculated as 68.12%. From these results, we can claim that this study underperformed in terms of accuracy. Thus, we can interpret that the studies that include other methodologies such as Machine Learning, etc. to perform sentiment analysis generally possess higher accuracy rates. As a future process, it is aimed to perform better text preprocessing and better usage of sentiment dictionaries to increase the accuracy rate of the system. The advantage of the study is to show that anyone can perform sentiment analysis without having advanced programming knowledge.

Another SA methodology is implemented by Yüksel and Tan in [15]. This methodology proposes to analyze and classify restaurant reviews as positive, negative or neutral. They used a self-collected dataset which was gathered from the Foursquare application and comprises 7086 Turkish reviews from 128 different restaurants. They used the ITU Turkish NLP Web Service, Zemberek and Google Translate API. Their presented approach, Social Information Discovery Algorithm (SIDA), makes decisions in classification based on the presence of some special words which alerts the polarity of the sentiment of a review. According to the experimental results they reported, their algorithm achieved an 81,97% accuracy rate while NB algorithm achieved a 73% accuracy rate in Turkish reviews.

Authors in [16] conducted a sentiment analysis on Twitter data. In this study, the sentiment values of tweets were determined and labeled as positive, negative, or neutral in addition to extracting keywords from tweets and thus creating a sentiment dictionary. In this study a dictionary-based approach with an n-gram model is used for classification. As a dataset, it was created by collecting approximately 7k Turkish tweets from Twitter over a 4-month period. First, all the tweets in the dataset were preprocessed by removing redundant characters and processing special Turkish characters to prepare the data for the classification process. The number of repetitions of words along with word frequencies was found. For the N-gram model, 2,3,4 grams were used. In the dictionary-based approach, the words were classified and grouped as positive, negative, or neutral according to the information in the dictionary. The system was evaluated. As a result, the proposed system achieved the highest accuracy rate of 72%.

Word Level Sentiment Analysis

In this section, Turkish SA studies which employ a word-level SA are reviewed. These studies are reviewed based on methodology, dataset, and performance. The reviewed studies use dictionary-based approaches to perform word-level SA. In word-level SA, the polarities of single words are detected. In the study conducted by Dehkharghani et al. [3], sentiment analysis was performed on different granularity levels. The levels that the sentiment analysis performed were aspect level, word level, sentence level and document level. Word level polarity classification was performed. The word n-grams were assigned polarity scores using lexicon

based approaches. All of the details about this word polarity assignment can be found in section 2.1.

The study by Aydın et al. [17] proposed an approach to generate word and document embeddings for sentiment analysis. The sentiments of words were unstable. Sentiments can differ from one corpus to another. The reason for this instability is the usage of various methods and approaches in SA. This instability causes unbalance in system evaluations. In this study, the authors combined contextual and supervised features with the general semantic representation of words that occur in the dictionary. This research tried to create word vectors while using semantic and sentimental features of words in the vector generation process. The proposed model has many different components in its methodology. In the dictionary approach, they used the TDK lexicon which contains 616.767 different words. From this lexicon, the polarities of words have been obtained. But since the TDK lexicon was not a sentiment lexicon, TDK was combined with the domain-specific scores gathered from the corpus to generate word vectors. In the final component, supervised contextual 4 scores, four supervised scores are assigned to each word obtained from the corpus. Considering this component, a more precise polarity score can be achieved since four scores are analyzed rather than only one score which is the self-score. After different components are created distinctly, they combined the generated output together to receive the best possible result. Lastly, the document vectors were generated accordingly. The flowchart of the proposed system can be seen in Figure 4.

Erşahin et al. proposed a hybrid approach in [18] to apply in Turkish sentiment analysis. A detailed review of this study can be found in section 4.2. The detailed review study by Türkmenoğlu and Tantuğ [19] can be found under 3.2. The authors in [19] used lexicon- and machine-based learning approaches separately.

Performance Comparisons of Dictionary Based Approaches

Table 1 presents Dictionary-Based Turkish Sentiment Analysis studies, categorizing them by year of publication, the utilized approach, analytical level, dataset size, and highest accuracy reported. As presented in Table 1, [20] achieves 91% accuracy, leveraging a substantial dataset of 43,000 samples. Notably, studies [11] and [19] both employ the SentiStrength dictionary; [11] achieves a 73.7% accuracy rate with a dataset comprising 60,000 samples, while [19] demonstrates a slightly higher accuracy of 77.1% using a dataset of 20,000 samples. However, it's worth noting that for a more comprehensive performance comparison between these studies, additional details concerning their respective datasets and approaches are required.

Various approaches can be used to find the sentiment of a given text. These are dictionary-based, ML-based and hybrid based. In this chapter, we analyzed sentiment studies using dictionary-based approaches. While dictionary-based sentiment analysis has its benefits such as offering simplicity and transparency, it is considered outdated when compared to hybrid ML approaches. Dictionary-based methods have limitations as they don't have the ability to recognize context well, making them less effective in capturing subtleties in language [8] which leads to less effective results in performance. They struggle to perform

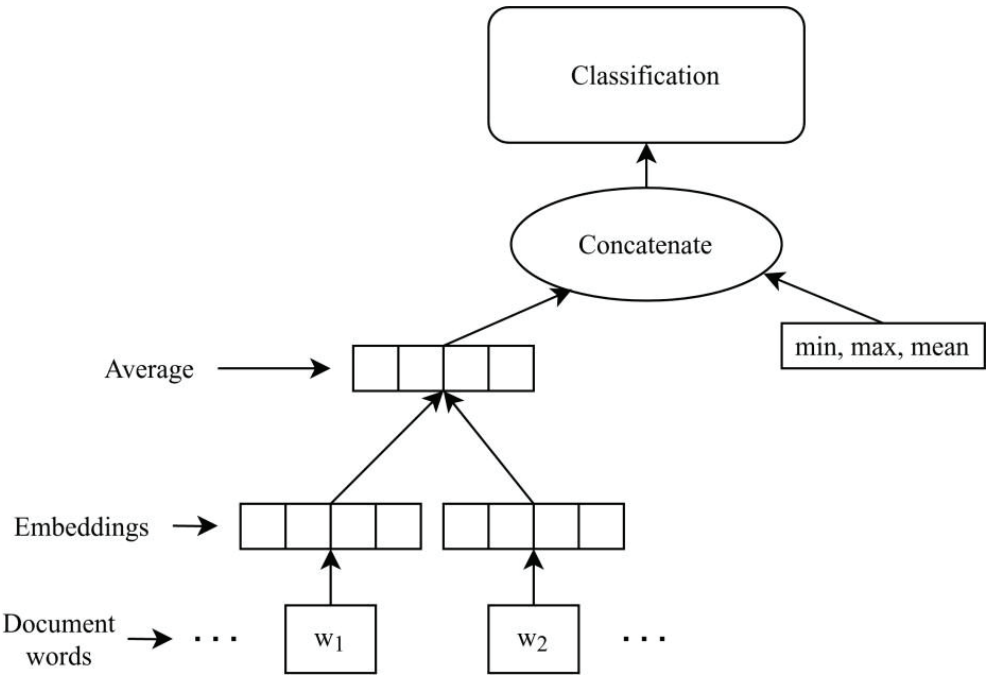

Figure 4. The flowchart of the system.

Table 1. Comparisons of Dictionary-Based Turkish Sentiment Analysis Studies

Study	Year	Level	Data size	Dictionary	Performance metric	Result
Vural et al. [11]	2013	Sentence	60k	SentiStrength	Accuracy	73.7%
Türkmenoglu and Tantug [19]	2014	Sentence	20k	Sentistrength	Accuracy	77.1%
Akgül et al. [16]	2016	Sentence	7k	TDK	Accuracy	72%
Karaöz and Gürsoy [14]	2018	Sentence	1.2M	Custom	Accuracy	68%
Yüksel and Tan [15]	2018	Sentence	7k	Custom	Accuracy	81%
Erşahin et al. [18]	2019	Sentence	220k	SentiTurkNet	Accuracy	74.90%
Toçoğlu and Alpkocak [20]	2019	Sentence	43k	Custom	Accuracy	91%
Aydın et al. [17]	2020	Word	10k	TDK	Accuracy	78.3%

sufficiently with complex languages like Turkish. Besides, hybrid approaches combine the simplicity of dictionaries with the contextual understanding of ML, offering a balanced solution. The evolution of ML-based techniques has significantly improved accuracy and adaptability, making dictionary-based methods less popular in today's dynamic world of sentiment analysis. In particular, the usage of deep-learning techniques has evolved over the years and can provide outstanding results. In chapters 3 and 4, ML and hybrid-based approaches are further explained.

TURKISH SENTIMENT ANALYSIS WITH ML BASED APPROACHES

In this section, Turkish sentiment analysis studies that use ML-based approaches are presented. ML is considered an important concept and a versatile tool that has gained popularity in recent years and is now applied in many different fields. Due to its versatility, ML can be classified as the most popular approach when compared with other methodologies [8]. There are various ML techniques that can be used in Turkish sentiment analysis. ML techniques can be classified as supervised, unsupervised learning and deep learning techniques. All categories of ML techniques are useful in solving a variety of NLP tasks. In sentiment analysis (SA), using ML approaches has many advantages and provides successful results when applied to a well-structured large dataset. Machine learning and deep learning techniques involve training models on large datasets to identify patterns and context in Turkish text. These models then assign sentiment labels to the text based on their learned knowledge. In ML-based Turkish SA, several machine techniques are generally utilized in a single study. Different levels of ML-based Turkish SA can be performed. These levels are aspect level, document level, sentence level, and word level.

Aspect Level Sentiment Analysis

In this section, studies that perform aspect level SA are reviewed based on methodology, dataset, and performance. The studies reviewed below use machine learning-based approaches to perform aspect-level SA.

Another recent study has been conducted by Bayraktar et al. in [21] in which a holistic method has been researched for Turkish aspect-based SA. As a dataset, many restaurant reviews were collected from various resources. Prior to the SA process, preprocessing steps were applied to the dataset. For preprocessing, spelling errors were removed and corrected using Zemberek and Yandex. Then, all the words in the dataset were converted to lowercase and inflectional suffixes were also removed from the words in the dataset using XML tools. This way the preprocessing is done and the root words in the dataset are gathered without losing their meaning. In aspect-based SA, the Latent Dirichlet Allocation (LDA), Pointwise Mutual Information (PMI), C-value, and WSBFE (Web Search Based Feature Extraction) are used during the aspect extraction. In [21], PMI was used to measure the association between two terms P (word1, word2) and indicated the probability of word1 and word2 coexisting, and the formula P(word1) P(word2) represents the probability that the two terms coexist when they are statistically independent. Overall, this research received 56,28% accuracy in the aspect extraction process, and received 52,05% accuracy in sentiment classification, which is relatively low. The main reason that the accuracy is low is the unpredicted aspect-sentiment pairs which are misclassified by the system. As future work, the authors in [21] are considering applying a double propagation method to increase the success of the aspect sentiment matching.

$$PMI(word_1, word_2) = \frac{P(word_1, word_2)}{P(word_1)P(word_2)} \quad (1)$$

$$P(word) \equiv hits(word) \quad (2)$$

$$P(word1, word2) \equiv hits(near(word1, word2)) \quad (3)$$

Ekinci and Omurca proposed a new approach in [22] in which it was aimed to classify the comments written about a product via a subject modeling method according to product features. In this study, GDA was used as the

topic modeling algorithm. The dataset used consisted of 1,000 user comments regarding hotels. These reviews were collected from the website "www.otelpuan.com". The data consist of a total of 5364 sentences. First, the dataset was preprocessed using the Zemberek library. Then, the classification was performed and the system was evaluated. According to the results, the success rate of the system was exactly 99%. The advantage of this study is that it achieves high accuracy by using the LDA algorithm. The disadvantage of this study is that it used a very small dataset to perform classification. So, the rating of the system's accuracy, which is a 99% accuracy rate, may not be reliable.

Mutlu and Özgür proposed a recent study in [23] that focuses on performing a targeted SA on Turkish text. The difference of targeted SA over normal SA is that targeted SA tries to predict the sentiment of a text according to a specific text rather than identifying the overall sentiment. To perform targeted SA, Bert-based models were implemented using different architectures. Bert is a neural network-based model which is widely used in Natural Language processing. The dataset was collected from Twitter. 19% of the dataset includes sentimentally positive tweets, 58% includes negative tweets and 23% includes neutral tweets. The model is trained and tested in the collected dataset. Lastly, the results of the proposed Bert model and different baseline Bert models were evaluated. According to the results, the proposed model outperforms other baseline models by achieving a 67% F1-score when tested on the same dataset.

Sentence and Document Level Sentiment Analysis

In this section, studies performing sentence-level SA are reviewed based on methodology, dataset, and performance. The studies reviewed below use machine learning-based approaches to perform sentence level SA.

Kilimci proposed a new SA system in [24] in which the direction of the Borsa Istanbul index is predicted using SA. For datasets, two distinct datasets were created by utilizing Turkish and English tweets on the tags BIST100 and XU100 in Twitter and used in this research. These datasets were enriched by using word embedding methods such as Word2vec, Glove etc. Although more than one method was emphasized in the study, ensemble learning was unanimously endorsed in the end. Since it is necessary to use different classifiers in ensemble learning, the heterogeneous ensemble system is emphasized. In the study, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) algorithms were used as basic learners to ensure community diversity. The MV and "STCK" methods were used to combine community decisions. Firstly, ant colony optimization and selection of the features in the dataset were performed for the dataset. Then, these obtained features were embedded using the word2vec and glove methods. Following this step, document representations were created using Term Frequency Inverse Document Frequency (TF-IDF) with Avg(Word2vec), Avg(Glove), Avg(Word2vec) +Avg(GloVe), TF-IDF+Avg(Word2vec), TF-IDF+Avg(GloVe). Ensemble classification is obtained by applying CNN, RNN, and LSTM algorithms to the obtained document representations. The results obtained in the community classification have been achieved utilizing the majority voting and heap community strategy methods. As a result, the study achieved a 78.07% classification performance rating for the Turkish dataset [24].

The author in [25] proposed an approach to perform SA on Turkish tweets from Twitter based on the Latent Dirichlet allocation algorithm. In this study, the data from Twitter were tried to be classified as positive, negative, or neutral. Two approaches can be used to perform SA, machine learning-based approaches and dictionary-based approaches. The dataset used in this study was collected from Twitter. A total of 10600 tweets were retrieved from Twitter. 5300 out of 10600 tweets were positive tweets and the remaining 5300 tweets were negative. First, the data were preprocessed, then machine learning methods were utilized in tweet classification. The Naive Bayes algorithm was chosen as the main classification algorithm. The proposed model was evaluated using the "wekada" cross-validation method. According to the system evaluation, the proposed system achieved a 78.34% accuracy rate in tweet classification.

Kaynar et al. proposed a model in [26] for Turkish SA, performed using machine learning and deep learning methods. In this study, Turkish data was collected from social platforms as a dataset. These data were then classified as positive or negative. For the dataset, 2000 movie reviews were collected from an online resource of which 1000 these reviews were negative, and the remaining were positive. For SA, artificial neural networks(ANN), support vector machines, Naive Bayes, and center-based classifier methods were used. This study mostly focuses on ANN. The ANN structure used is given in Figure 5.

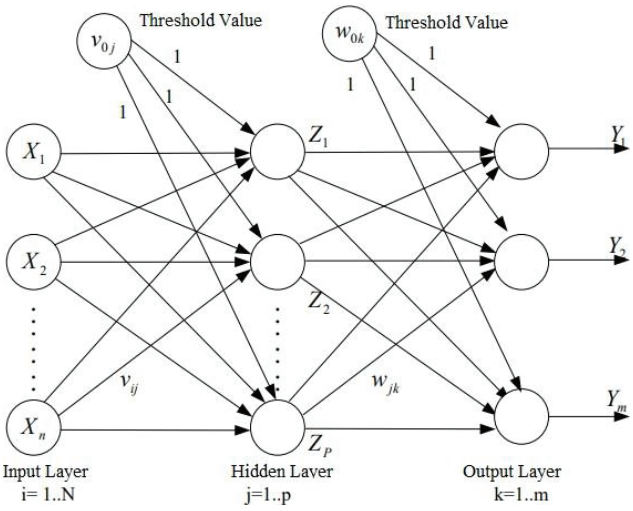

Figure 5. The structure of artificial neural networks.

Aytekin Keskin proposed a SA method in [27] on Turkish texts regarding interest-free finance systems. Specifically, perceptions of potential customers towards interest-free financial systems were aimed to be detected. The dataset used consisted of Turkish reviews of customers about the interest-free finance system in January 2019. The content of the dataset was collected from various internet resources and social media. In this study, a sentence-level SA was realized using a Machine learning approach on Turkish texts. The SA was performed using an online program from the "www.mediatoolkit.com" website. In the classification process, the most common concepts related to interest-free finance systems were identified and the dataset was examined using keywords to classify information about interest-free finance systems. The SA of the texts in the dataset that mentioned concepts related to interest-free finance systems resulted in them being classified as positive, negative, or neutral. These results were analyzed and recommendations were put forward based on the findings. This study aimed to understand customer perceptions of interest-free financial systems. The results showed that the mention of "participation banks" alongside the concept of "interest" in the press creates a negative bias towards these institutions. Additionally, the association of commercial institutions that provide interest-free financing systems with various concepts, namely banks, leads to varied feelings about these institutions within society.

The authors in [12], proposed a SA system to classify the polarity scores of tweets. In this study, the dataset was collected from Twitter. Twitter API was used in the data collection process to retrieve tweets efficiently. This process resulted in the creation of a dataset that consists of 20k tweets. To classify tweets, machine learning approaches were used. Specifically, SVM, Multinomial NB, and K-nearest Neighbors (KNN) were used in the classification process. In this system, it is stated that since each tweets in the dataset can be classified into different categories such as positive, negative, etc., and the basic document classification algorithms were applied. Before the classification process began, preprocessing methods were applied to the content of the tweets. Then, the attribute selection was performed. The system was evaluated for every machine learning approach used in the classification. According to the results, it is seen that different machine learning approaches have different accuracy yields in this system. The accuracies of machine learning approaches are NB:62.04%, MNB:66.06%, SVM: 62.25%, KNN:65.79%. Thus, based on the accuracies achieved in the system, the Multinomial NB outperforms other approaches. In conclusion, this study used widely known classification methods to perform SA on Turkish tweets. For future improvements to the system, an external lexicon could be used besides Machine learning methods. The authors stated that adding an additional lexicon would increase the system's accuracy. The major decrease in the accuracy of the system is the result of using random sentiment score based methods to determine the sentiment value of tweets.

In the study by Türkmenoglu and Tantug [19], opinion extraction was performed and the polarity of these opinions were analyzed. To achieve these, two different approaches were used. These approaches were Machine Learning-based and Lexicon-based SA methods. In this study, two distinct datasets were used. One of these datasets consisted of tweets collected from Twitter. The other dataset consisted of movie reviews which were collected from the website "beyazperde.com". In the machine learning approach, SVM and NB methods were used in the classification of opinions. To use these machine learning methods, a feature set which contains word root words and n-grams was represented with a bundle of words. For the lexicon-based approach, it is attempted to predict the sentimental orientation of an input text using sentiment scores of words and phrases in text using the information in the sentiment lexicon. As a result, the lexicon approach achieved an accuracy of 77.1% and the machine learning approach achieved an accuracy of 89%. From these results, it is detected that the machine learning approach outperforms the lexicon approach.

In the study conducted by Kaya et al. [28], SA in the Turkish language was investigated using different sentiment classification techniques. These techniques were different supervised machine approaches. A total of four different supervised machine learning algorithms were used. These are Naive Bayes, Maximum Entropy, SVM and the character-based N-Gram Language Model. The N-gram language models are used to create the N-gram based character language model. Instead of words, this model uses characters as the basic unit in the algorithm. For strings "s," the model provides p(s). The chain rule is given for a character (c) and a string (s). The context is limited to the previous (n-1) characters due to the N-gram Markovian assumption. As a result, the maximum likelihood estimator for N-grams is given, where C(sc) is the number of occurrences of the sequence in the training data and the denominator is the number of single-character extensions of sc.

$$p(sc) = p(s) \times p(c|s) \tag{4}$$

$$p\left(c_n|s_{c1\cdots c_n-1}\right) = p(c_n|c_1 \cdots c_{n-1}) \tag{5}$$

$$p(cs) = \frac{C(sc)}{\sum_{c} C(sc)} \tag{6}$$

For this study, a dataset was created containing Turkish political news articles collected from different news sites' political sections. Before the sentiment classification was performed, some preprocessing steps were applied to the dataset. These preprocessing steps were removing the HTML tags and stemming the words in the dataset using the Zemberek tool. Then possible roots and suffixes are

found automatically. Additionally, a list of words which people tend to use to express strong sentiments were created to be used in this study. Then, the classification process began. To evaluate the system, the K-fold-cross-validation was used. The experiments of this study were conducted with Bigram and Unigram features. However, results show that bigram information is not as useful as information obtained through unigram for the Sentiment Classification of news. Overall the system received a 76% accuracy rate. For future work, the authors thought of adding Named Entity Recognition to the system which might be used to identify which columnists write about which political party or politician etc.

In this research conducted by Ciftci and Apaydın [29], a modern deep learning method called RNN was developed using LSTM units on the dataset instead of machine learning methods that use the word bundle model such as Logistic Regression and Naive Bayes for SA and comparing the results of RNN based methods with the results obtained from traditional machine learning methods. It was revealed how much improvement modern natural language processing and deep learning methods can bring about in emotion analysis results. An external library was not used for this research. First of all, a 355-thousand-word dataset consisting of 283 thousand positive and 72 thousand negative words was collected from Turkish shopping and movie reviews. After the tags were removed and normalization was applied, punctuation marks and stop words were removed. Then, the datasets were separated as 80% training and 20% testing. After Naive Bayes and Logistic Regression models sentences were vectorized using TF-IDF, a hyper parameter search was conducted for training. TF-IDF was not used for the RNN based algorithm, word vectors directly fed the RNN architecture. The results showed that the RNN-based deep learning method improved the classification accuracy. As a result, the biggest advantage of this research is that despite the unbalanced data problems of modern deep learning methods and the dominance of positive comments, RNN-based algorithms have been shown to improve accuracy in emotion analysis, but the long training times of deep learning methods can be viewed as a disadvantage.

A hybrid machine learning approach is proposed by Shehu and Tokat in [30] using "Random Forest (RF)" and "Support Vector Machine (SVM)" algorithms used for Turkish SA. Turkish letters of 3,000 and 10,500 words were used as a dataset. A sentiment dictionary translated from 27,000 English words was used as the SA dictionary. According to this proposed hybrid model, "Zemberek" was used as a resource to find the roots of the words as a first step. According to the authors, it has been observed that the RF algorithm gives better results than the SVM algorithm in the classification of positive words, while the SVM algorithm gives better results than the RF algorithm in the detection of negative and neutral words. Based on these two results, it was planned to classify the data according to two classes, 'Positive' and 'Others', using the RF algorithm. The data in the 'Others' class is intended to classify the data according to three classes, positive, negative, and neutral, using the SVM algorithm. While SVM achieved 76.4% and RF achieved 75.9% accuracy in the dataset with 3,000 items of content, the mixed method achieved an accuracy rate of 86.4%. In the large data set consisting of 10,500 items of content, SVM achieved an accuracy of 67.6% and RF 71.2%, while the mixed method achieved 82.8% accuracy. The accuracy of this approach can be improved and tested with the use of other data sets. The biggest advantages of the study are that the accuracy rates in classification can be increased by using larger emotion dictionaries, it is open to development due to the improvement of classification algorithms, and it is clear that the mixed model has achieved higher accuracy compared to other methods.

A new SA study was proposed by Acikalin et al. in [31], using various data written in Turkish. In this study, two models based on the BERT model were proposed. These were developed by Google using the transformer architecture. In the first model, a multilingual BERT was adapted to the proposed system. In the second model, Turkish texts were translated to English after which the main model of BERT, which was developed for English, was adapted to this Turkish dataset. The mTranslate library was used in the translation process of the system. Two distinct datasets were used in this proposed system, these datasets included movie and hotel reviews collected from online resources. The size of the movie reviews dataset includes 53,400 words, of which 26700 are positive and 26700 are negative. The size of the hotel review dataset includes 11,600 words, 5800 of which are positive and 5800 are negative. The models that were developed in this study were compared with 4 different models which are fastText-Tr, and fastText-En. Experiments show that when enough data is obtained, the BERT model can learn enough features to successfully classify the given data. In the future, adapting other general models for the purpose of text representation on different datasets and combining the results with BERT are considered. Additionally, developing a Turkish version of the BERT model is being given considerable thought.

The study by Balli et al. [33] aimed to find the sentimental analysis of Turkish tweets. The main purpose of this study was to detect the emotional states of Twitter users by classifying the tweets they wrote as positive, negative, or neutral. For this purpose, 2 different Twitter datasets were used. The first dataset was a public dataset and the second one was a custom-made dataset named SentimentSet which included 11k tweets and is publicly available. After the dataset preparations, preprocessing was applied to the datasets to clean and prepare the data for classification. For classification machine learning models were used. Before the classification, the data was vectorized using the TF-IDF vectorizer to convert the text into numerical form to be used by the machine learning models. Logistic Regression, Random Forest, Naive Bayes, SVM were implemented for machine learning models, and for deep learning, LSTM

models were implemented. Then, the models were trained and tested with the prepared data. According to the results, random forest and LSTM models performed the best by achieving an 85% accuracy rate.

The authors in [34] tried to classify tweets about music, art, etc. The analysis was conducted within categories. In the study, the tweets were taken from Twitter using the Twitter API. In this study, the libchart library was used to graphically display the output of SA and categorical classification. The Bayesian algorithm was used for the classification of tweets. In this study, the tweets of a Twitter user are categorically classified and SA is performed on those tweets. In future studies, it is aimed to increase the number of tweets used in the study and increase the accuracy of the system accordingly. The advantage of the study is that SA and classification of any given Twitter user can be performed through this created interface. The main disadvantage of the study is that only 100 tweets from a selected user can be processed at a time.

In this study by Tuncer and Çetintaş [35], the authors tried to sentimentally classify the tweets on Twitter as positive, neutral and negative and also to detect malicious tweets. For classification, the Decision Tree algorithm and Naive Bayes algorithm were used. For the dataset, 20,000 tweets were collected and used in the system. Knime program and excel macro were used as utility programs. According to the evaluation of the system, the system achieved an average of 75.2% accuracy rate using the Decision Tree algorithm and 56% with using the Naive Bayes algorithm. The advantage of the study was that the classification stages were performed in pairs, which resulted in a more observable result. The disadvantage of the study is that some Turkish translation processes were incomplete.

In [36], Güran et al. studied the sentiment polarity detection problem with social media data. They applied a grid search method in order to discover the most suitable kernel function in SVM. They used 3 datasets: 1) VS1(3 classes, 3000 data), 2) VS2(4 classes, 157 data), 3) VS3(3 classes, 105 data). According to the experimental results they reported, their proposed model received an average of 75.2% accuracy.

A SA model was proposed in [37] for Turkish text. For this purpose, three different types of data were used, which were reviews of products, movies, and books. A machine learning model was used to process and perform SA on all of these three types of data. The data was attempted to be classified as positive, negative, or neutral. First, the most frequently used 20,000 words in the dataset were determined. Scores were assigned between 0 and 1 by normalization. By removing the missing and neutral data, 105,220 data were obtained from each dataset. A mixed set of data was created by taking equal amounts of data from all three datasets. TensorFlow was used for the models. With the Sklearn library, the data was split as 90% training and 10% testing. While each model provided a very high (85-95%) accuracy in its own channel, the success rate was greatly reduced in other channels (50%). While the mixed model had success rates of 79.8% and 85.8% in mixed data, it also achieved a high success rate of 77.8% and 85.8% in other channels.

In the study by Aytuğ [38], SA was performed by analyzing tweets collected from Twitter. Three different machine learning models were used in the classification process of tweets. A special dataset was created for this research by collecting tweets from Twitter using the Twitter API with the Python programming language. This created dataset contained a total of 10600 Tweets, 5300 of these tweets were positive and 5300 are negative. Before the data classification process, preprocessing was applied to the dataset. For preprocessing, the stop words and characters such as "@", "#" were cleaned and all the words in the dataset were rooted using the Zemberek Library and duplicate words were removed from the dataset as well. The N-gram model was used to determine the appropriate features of the data. For classification, NB, SVM, and (Logistic Regression) LR models were used. The Weka software was used to perform a classification procedure on these three classification models. The proposed SA system was evaluated using the 10-fold cross validation method. According to the results, different classification models received different rates of accuracy. The LR model received 77.23% accuracy, the SVM model received 73.68% accuracy and the model NB received 77.78% accuracy. Even though the accuracy rates of the three models are close to each other, it is seen that the NB model achieved the best accuracy rate compared to other models.

In [39], Sarıman and Mutaf have attempted to analyze people's feelings about Coronavirus through social media from the date of the spread of the virus until the present day. People's feelings have been attempted to be analyzed with the help of textual material shared on social media, specifically, tweets shared on Twitter. As of March 11, 2020, 2 million tweets were collected using the Twitter API to form a dataset. To process the data, Python's "pandas", "numpy" and "sklearn" libraries have been used. Preprocessing was applied to the created dataset. Preprocessing is applied by removing unnecessary Turkish characters in the dataset and also by deleting the recurring tweets in the dataset. The tweets underwent SA using machine learning methods. The meaning of the tweets in the dataset was obtained through word sequences, sentence analysis, and emotion analysis. Logistic regression was used in the classification of tweets. The system was evaluated and the AUC metric was used as the basis metric for success. As a result of the classification process, five main classes were created as follows: Eba, Mask, State Support, Curfew, and Short Working Allowance. After predicting the sentiment of the tweets, the AUC value was obtained. The AUC classification results are as follows: Mask (0.97), Eba (0.94), Curfew (0.98), State Support (0.86) and Short Working Allowance (0.91) results were obtained.

The authors in [40] developed an application to analyze tweets using the Tweepy, Odoo, and NLTK modules in the Python programming language. Analyses were made by accessing the tweets posted over the hashtags on the Twitter platform. The Odoo module was used to display data and results in an organized structure. The NLTK module was used to apply natural language processing techniques to the data. The dataset used was created by collecting tweets from Twitter. As a result of the system, it was observed that Twitter tags can be analyzed in a versatile way within a single interface.

In the study by Kaynar et al. [41], a comparison was made regarding the Feature Reduction Methods with Deep Autoencoder Machine Learning in SA. The autoencoder architecture is given in Figure 6. In the autoencoder architecture, the first layer is the input layer, the second is the hidden layer and the final one is the output layer. The space between the input layer and the hidden layer is called the encoder. The space between the hidden layer and the output layer is called the decoder. The results of using linear and nonlinear dimension reduction techniques in combination with machine learning methods were compared to the results of using machine learning methods by themselves. It has been observed that the model that uses linear and nonlinear methods with machine learning methods performs better than the model that only uses machine learning models.

In another recent study proposed by Tuzcu in [43], Turkish texts were classified through SA. In this study it was aimed to receive high accuracy for Turkish SA. A Sentence and document-level SA was performed using machine learning-based approaches. For the dataset, a total of 91309 book reviews were retrieved from an online website. Prior to the sentiment classification, all of the data in the dataset was preprocessed using the methods of the NLTK library. Various machine learning algorithms were used in the classification process of SA. These classifiers are MLP, NB, SVM and LR. The sentence- and document-level SA was performed on the same dataset using all these classifiers and were compared according to the classification success. The system was evaluated using accuracy as a metric. According to the evaluation, it can be observed that different machine learning algorithms received different accuracy rates when tested on the same dataset. The MLP algorithm received an 89% accuracy rate, the LR algorithm received 84% accuracy rate, the SVM algorithm received 80% accuracy rate and the NB algorithm received a 77% accuracy rate. According to these results, the MLP algorithm performs the best, achieving an 89% accuracy rate, when compared with other classification algorithms.

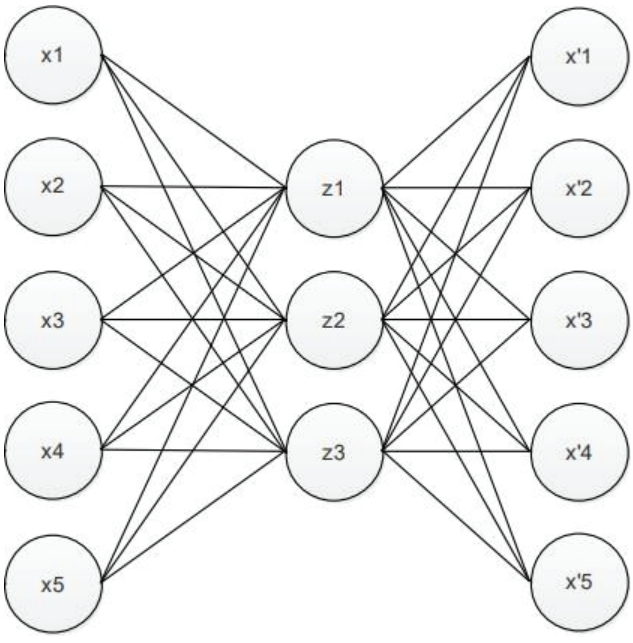

Figure 6. The Architecture of AutoEncoder.

In [44], the authors conducted a SA of Turkish language text using a dataset collected from various social media platforms. The study employed three algorithms: Random Forest, Logistic Regression, and LSTM. The dataset consisted of 28,189 data points collected from five social media platforms and manually labeled as positive, negative, or neutral. Of these, 5,712 were labeled as positive, 11,567 as negative, and 11,247 as neutral. The experimental results showed that deep learning models outperformed machine learning approaches in terms of performance. In the study, the LSTM model achieved the highest accuracy with a rate of 84.46%.

Performance Comparisons of ML Based Approaches

Table 2 presents the Machine Learning-Based SA studies including their approaches and performance results. In this table, ML-based Turkish SA studies are listed based on their year of publication, approach, sentiment level, the size of the data used, and highest results achieved in the study.

According to the results, several parameters are required to draw a meaningful comparison; approach, data, data size, performance metric etc. In this case, to make this comparison between the given studies, we should choose the studies that used the same performance metric since we don't have the complete information about their dataset. According to the given information, the results on Table 2 shows a wide range of performance in ML-based Turkish SA. Studies like Acikalin et al. [31] and Kemaloğlu et al. [44] achieve notably high accuracy rates, while others, such as Çoban et al. [12] and Tuncer et al. [35], report comparatively lower results. The choice of the machine learning algorithm, dataset size, and specific problem domain can significantly affect the outcomes. ML techniques, particularly deep learning approaches, have gained popularity in recent years to provide solutions for many specific NLP tasks, especially in SA [45]. Additionally, the recent utilization of deep learning techniques, as seen in studies like Acikalin et al. [31], has demonstrated impressive accuracy, underlining

Table 2. Comparisons of ML-Based Turkish sentiment analysis studies

Study	Year	Level	Data Size	Performance Metric	Result
Kaya et al. [28]	2012	Sentence Level	400	Accuracy	NB:72.05%, ME:76.78%, N gram:76.78%, SVM:76.31%
Güran et al. [36]	2014	Sentence Level	105	Accuracy	SVM: 75.2%
Türkmenoglu and Tantuğ [19]	2014	Sentence Level	20k	Accuracy	SVM:89% , NB:89%
Çoban et al. [12]	2015	Sentence Level	20k	Accuracy	NB:62.04%, MNB:66.06%, SVM: 62.25%, KNN:65.79%
Kaynar et al. [26]	2016	Sentence Level	2k	F-measure Sensitivity	83%
Onan [25]	2017	Sentence Level	10k	Accuracy	NB:78%, SVM:77%, KNN:72%, RF:72%
Alpkoçak et al. [42]	2019	Sentence Level	27k	Accuracy	SVM:86.2%, ANN:86.6%, RF:82.5%, KNN:70,9%
Erşahin et al. [18]	2019	Sentence Level	220k	Accuracy	NB:82.07%, SVM:85.4%, J48:76.7%
Tuncer et al. [35]	2019	Sentence Level	20k	Accuracy	DT:75.2% , NB:56%
Acikalin et al. [31]	2020	Sentence Level	11k	Accuracy	Bert: 93.32%
Aksu et al. [32]	2020	Sentence Level	49k	F-Score	SVM:91%, NB:86%
Sarıman and Mutaf [39]	2020	Sentence Level	2M	AUC	LR:93%
Tuzcu [43]	2020	Sentence Level	91k	Accuracy	MLP:89%, LR:84%,SVM:80%, NB:77%
Kilimci [24]	2020	Sentence Level	30k	Accuracy	CNN:74%, RNN:72%, LTSM:76%,MV:77%, STCK:78%
Kemaloğlu et al. [44]	2021	Sentence	25k	Accuracy	84.46%
Aktaş et al. [48]	2022	Sentence Level	1M	Accuracy	NN:86.37%, NB:83.55%, KNN:81.88%

the evolution of these approaches in the field of Turkish SA. We can conclude that deep learning tools such as transformers can provide immense results. Bert (Bidirectional Encoder Representations from Transformers) is a type of transformer model which is used extremely frequently for NLP tasks [45], including in language understanding, translation, and SA due to its ability for understanding the context and relationships between words in a sentence, which is essential for many NLP tasks [45].

TURKISH SENTIMENT ANALYSIS WITH HYBRID APPROACHES

In this category, Turkish SA studies that use hybrid-based approaches are presented. Hybrid-based approaches are highly preferred for Turkish SA over single approaches like dictionary-based approaches. The use of a hybrid approach in SA combines elements of both dictionary-based and machine learning-based methods to improve the performance and effectiveness of sentiment classification. This approach is used to improve the strength and efficiency of each method. As a result, in hybrid approaches, machine learning techniques are used as a baseline approach. In addition to the ML- based approach, dictionary-based approaches can be used. In the hybrid approach that uses dictionary-based methods, an external lexicon is used with ML techniques. Hybrid-based Turkish SA can be executed on different levels. These levels are aspect, document, word and sentence.

Aspect Level Sentiment Analysis

In this section, a study performed on an aspect level SA is reviewed based on methodology, dataset, and performance. The study reviewed below uses hybrid-based approaches to perform aspect-level SA.

In the study by Çetin and Eryiğit [46], SA was performed to analyze people's feelings to classify opinions as positive, negative, or neutral. As a data set, a set of Turkish restaurant reviews was used. The dataset consists of a total of 1415 sentences of restaurant reviews. Prior to the classification process, the data is prepared by using preprocessing techniques. A Word segmenter, Turkish character corrector, morphological analyzer, morphological disambiguation and dependency parser provided by ITU NLP tool were used to prepare the data. In this study, the CRF algorithm is used as the primary method in tagging the data. Logistic regression was used to separate the generated data. In order to determine the relationship between the words in a sentence, two different methods were used both separately and in combination. One of these is the relationship of neighboring words at a certain distance and the other is the detection of words associated with the loyalty tree. For the criteria aspect, the F1 criterion was used in the evaluation

of the system. According to results, it was observed that the system received a 76% F1-score. In the future, it is aimed to improve the methods that can overcome the difficulties created by the Turkish dataset and to use deep learning methods to improve the accuracy of the systems.

Sentence and Document Level Sentiment Analysis

In this section, studies performing sentence-level SA are reviewed based on methodology, dataset and performance. The studies reviewed below use hybrid based approaches to perform sentence-level and document-level SA. In hybrid approaches, different methodologies are combined and used to propose efficient models for the task of SA.

In this study by Onan [47], SA was performed on Turkish text documents to evaluate the predictive performance of 36 word embeddings based on representations obtained by three-word embedding methods (i.e., word-2vec, fastText and DOC2vec), two basic weighting functions (i.e., inverse document frequency and smooth inverse document frequency) and three vector pooling schemes (namely, weighted sum, center based approach and delta rule). For the dataset, a total of 21000 Twitter messages were used, of which 10500 were positive and 10500 were negative. Preprocessing steps such as string splitting, removing stop words, and finding roots were applied on the dataset. The area under the ROC curve (AUC) was used to evaluate the effectiveness of the methods. Experimental analysis shows that word2vec-based representation in conjunction with inverse document frequency-based weighting and center-based pooling yields promising results for SA in Turkish. (0.9131 AUC). The advantage of this study is that there is no study conducted on Turkish text documents that comprehensively analyzes the weighted word vectors, and this study constitutes a basis in this manner.

In the research by Gezici and Yankıoğlu [4], a model was built to estimate the sentiment value of movie reviews in Turkish. This model combines supervised learning and lexicon-based approaches. Firstly, it computes the average polarity of the words in the text and trains a classifier (Naive Bayes or SVM). Then, the effectiveness of more complex processing techniques and features are measured: handling negation; considering the effects of booster words; and using other features derived from the seed words. In this research, the basic approach obtained 67.49% accuracy with the Naive Bayes classifier and 67.61% with the SVM classifier; while the best results were obtained using all of the features, achieving 74.28% accuracy with the Naive Bayes and 75.52% with the SVM classifiers, respectively. The advantages of this work is that it considers the seed word occurrences and, in this way, classification accuracy significantly improves. Secondly, if a larger lexicon can be used, the system can achieve an improved classification performance.

In the study by Aktaş et al. [48], a SA was conducted on reviews of an online food delivery service using machine learning models in Turkish. The goal was to predict whether a given review was positive or negative. To create the dataset, 676,000 reviews were collected from "yemeksepeti.com", with 338,000 labeled as positive and 338,000 labeled as negative. The dataset was preprocessed via lemmatization and normalization. The study compared the performance of several machine learning algorithms and deep learning algorithms, including KNN, NB, and a neural network. The results showed that the neural network model had the highest accuracy rate, at 86.37%, followed by the NB model at 83.55% and the KNN model at 81.88%.

In the study conducted by Erşahin et al. [18], a SA approach is presented using a hybrid approach which consists of both dictionary-based and machine learning approaches. In the dictionary-based approach, a new lexicon was created by expanding the STN lexicon via the ASDICT model called eSTN. The STN lexicon was improved by adding synonyms of words to the dictionary using the "ASDICT" model. For the ML side, the classification problem is handled by using different ML models, which are, NB, SVM and "J48" classifiers. ASDICT, which was developed to identify synonyms and contains 70,000 words which were used to expand an external dictionary. In this study, different datasets containing hotel and movie reviews and tweets were utilized. The selected movie reviews, hotel reviews and tweets included a total of 1,345,726, 738,216 and 19,056 words respectively. The techniques that were used in experiments are: "NB", "NB+active learning", "Logistic regression+QER", "Lexicon" and "SVM". The results show that a hybrid approach to SA provides better results. The highest relative accuracy, at 83%, was achieved by the system. Future consideration is given to improving the overall performance of this system by using aspect-based SA and other subtasks of SA.

In this study conducted by Aydın et al. [51], it has been attempted to develop a hybrid SA method with a higher success rate than the dictionary-based and machine learning-based methods. In the study, tweets related to the Apple, Google and Microsoft companies were used as a dataset. These tweets were divided into 3 sections which were positive, negative and neutral. Specifically, three datasets were used. The first two datasets consisted of 479,988 tweets. These tweets were classified using the Pso and k-eyk algorithms. With the use of the multiple Pso classifier, it has been observed that the Pso population is oriented towards their own class from the 3 separated classes, and as a result, each particle can be classified according to its original class. The system was evaluated by comparing three methods on four parameters. These parameters are precision, sensitivity, f1-score and accuracy. The system achieved a 73% precision score.

In the study by Demir et al. [52], the authors tried to sentimentally analyze large amounts of Turkish text data in a short amount of time. In this study, three varying datasets were used. The first dataset contains about 5000 Turkish words, and the second and third dataset contains a total of 25000 words. In this study, mostly a dictionary-based

approach was used with machine-learning methods to conduct SA. In the dictionary-based approach, four different dictionaries were used. These dictionaries were by Afinn, Bing, NRC and SentiTurkNet. A word-level SA method was performed using a dictionary-based approach. Then, for classification of words in the dataset, the NB method was used. This hybrid system was evaluated. According to results, the system achieved a peak accuracy rate of 82%.

In the proposed paper [6], "The Turkish movie reviews" dataset (composed of 34990 positive and negative movie reviews in Turkish) had been used for the sentiment classification task. The model was fed with pre-processed and vectorized data. The pre-processing included the steps of tokenization, stopwords, special characters removal, fixing misspelled words, stemming, and detecting negation. Vectorization was performed with the VSM (Vector Space Model). This vectorization model uses features like TF, TF-IDF, and Word Embeddings (Word2Vec, GloVe). For the classification part, these extracted features were applied to the three different well-known ensemble algorithms which were AdaBoost Classifier (AdaBoost), Random Forest (RF), and GradientBoostingClassifier (GBC). The evaluation of the classifiers was made according to standard metrics such as precision, recall, and F1 score. The best result achieved among these models belongs to the RF classifier with 86% accuracy.

In [53], Saed Alqaraleh proposed a model based on Deep Convolutional Neural Networks (ConvNet). "Turkish movie reviews" dataset (composed of 34990 positive and negative movie reviews in Turkish) is used for the training and the evaluation of the model. The pre-processing of the model consisted of tokenization, cleaning, correcting any misspelled words, and negation handling steps. The Word2Vec word embedding model was used for the feature extraction stage. In the classification part, CNN based architecture model with an embedding layer, GlobalMaxPool1D layer, Dropout layer, and a Dense layer was built to obtain sentiment classification results. Standard metrics were used for the evaluation part. The average F1 result achieved was 82.36%. It was also found that 64 was the optimal number of filters for the model.

Harisu et. al. in [54] proposed a paper that compared TML algorithms (RSVM, RANF, MAXE, SVMs, and DECT) and deep learning (DL) models which are built using three main DL models called RNN, CNN, and hierarchical attention network (HAN). They used stemmed Turkish Twitter data to perform the SA task on. They also tried to increase the size of training data by applying a few data augmentation techniques called shift, shuffle, and hybrid (a combination of shifting and shuffling). To evaluate their proposed model, different evaluation metrics are used such as ACC, AUC, F1S, and RTM. While the TML algorithms are better in training time and runtime compared to the DL algorithms, in the case of AUC, ACC and F1S metrics, the DL algorithms outperformed the TML algorithms.

The authors in [55] proposed to achieve better results by combining different types of word embeddings (Word2Vec, fastText, character-level embedding) with different deep learning methods (LSTM, GRU, BiLSTM, CNN). The evaluation metrics were accuracy, precision, recall, and F1 score. The proposed model was evaluated on a dataset which was collected from Twitter regarding GSM operators in Turkey. It consists of 17,289 Turkish tweets and has 3 different labels for the sentiment (positive, negative, neutral). Features were extracted and used by word embeddings and DL algorithms both separately and in combination. The pre-trained BERT, ALBERT, ELECTRA, and DistilBERT models were used since they have infrequently been used in Turkish literature. The best model, ELECTRA, achieved 98.38% accuracy (k=10) for the hotel dataset and 92.21% accuracy (k=10) for the movie dataset. It was discovered that using TFM with a transformer would yield better results.

Another up-to-date study has been proposed in [57] including a new pretraining objective known as SSP. The main advantage of the proposed new pre-training task over NSP and SOP is that it makes better use of the dataset and generates more training input from it. As a result, models can be trained with more steps than in NSP and SOP. Models trained with SSP and SOP outperformed models trained with NSP. This demonstrated that using NSP alone in training is insufficient and that the model can achieve better results by performing other tasks. The SSP models performed similarly to the SOP models but outperformed the SOP models in masked word prediction.

Performance Comparisons of Hybrid Based Approaches

Table 3 presents the Hybrid-Based SA studies including their approaches and accuracy results. In this table, Hybrid-Based Turkish SA studies are listed based on their year of publication, applied approach, the level at which it is performed, the size of the data used, and highest performance achieved in the study.

In order to make a more meaningful comparison between these studies, more information must be known about the dataset. Without the information in the dataset, all these results present a range of performance in SA. Onan [47] stands out with an impressive accuracy of 91%, while Aydın et al. [51] and Demir et al. [52] also achieved high accuracy rates. On the other hand, Dehkharghani et al. [3] and Çetin and Eryiğit [46] achieves slightly lower accuracy results. However, it's worth noting that the techniques available during the year of the study's publication can also have a significant impact on the results. Additionally, the choice of the dataset size, analysis level, and specific approach significantly contributes to the variations in outcomes. Notably, Onan [47] demonstrates exceptionally accurate results in this category by utilizing different word-embedding models. Word embedding models are very effective in SA because they include contextual and semantic understanding of words. Their pre-trained state, along with their ability to handle context, contributes to their high performance in SA tasks [2].

Table 3. Comparisons of hybrid based sentiment analysis studies

Study	Year	Level	Data Size	Performance Metric	Result
Gezici and Yanıkoğlu [4]	2018	Sentence	10k	Accuracy	73.7%
Aydın et al. [51]	2018	Sentence	2k	Accuracy	77.1%
Çetin and Eryiğit [46]	2018	Sentence	1k	Accuracy	72%
Dehkharghani et al. [3]	2019	Document	60k	Accuracy	68%
Demir et al. [52]	2019	Word	91k	Accuracy	81%
Erşahin et al. [18]	2019	Sentence	220k	Accuracy	74.90%
Onan [47]	2020	Sentence	21k	Accuracy	91%
Kılıç et al. [50]	2020	Sentence	2k	Accuracy	78.3%
Köksal and Özgür [59]	2021	Sentence	5k	Accuracy	72.9%

TURKISH SENTIMENT ANALYSIS WITH HYBRID APPROACHES

In this section, it is explained how polarity lexicons are constructed and studies that work on constructing polarity lexicons are analyzed and reviewed. Polarity lexicons can be constructed using different methods. These methods are classified as statistical and dictionary based methods. In statistical-based lexicon construction, the statistical structure of the dataset that is used to construct the lexicon is used. In dictionary-based lexicon construction, an external lexicon is used as a baseline in the construction process. In the following two studies, polarity lexicons were constructed using statistical-based approaches.

In the study by Sağlam et al. [58], a Turkish sentiment lexicon was created for SA. This study aimed to develop an existing Turkish sentiment lexicon using data from online news media. In the aforementioned literature, there is not much useful research about developing an existing lexicon. In many studies about SA in Turkish, the online sources are collected and sentimentally analyzed but the data that is used has not been converted into a sentiment lexicon. This research primarily aimed to create an extended Turkish sentiment lexicon that can be used in various different studies. For the base Turkish sentiment lexicon, "SWNetTR" was used. "SWNetTR" has a capacity of 27,000 words. The data was collected using GDELT datasets. From the GDELT dataset, 100,000 online news documents were randomly chosen. After this process, raw text inputs were extracted from the selected documents. To create the dictionary that will be merged with "SWNetTR", first the collected raw text sentences were tokenized. As a result of this process, 14,000 words were received. Then, using the "Zemberek" morphological tool, the stems of all the words were received. Following all these processes, polarity scores of all the words were determined. All the words along with their morphological analysis and polarity scores were saved to a file. This file was named SWNetTR-GDELT. The "SWNetTR-GDELT" and base lexicon, which is "SWNetTR", was compared to prevent having duplicates. From this comparison, it was shown that the newly created dictionary "SWNetTR-GDELT had 10,000 unique words that did not exist in the base dictionary "SWNetTR. These found 10,000 unique words were added to the "SWNetTR" dictionary with their polarity scores. This way, the capacity of" SWNetTR" was increased by 10,000 words. Before "SWNetTR" had 27,000 words and the increased version had a capacity of 37,000 words. This new extended dictionary was now called "SWNetTR-PLUS". After this process, the accuracy of "SWNetTR" and "SWNetTR-PLUS" were calculated. According to the results, "SWNetTR" had a 60.6% accuracy rate of polarity classification performance. The extended dictionary "SWNetTR PLUS" had an accuracy rate of 72.2%.

In the following study, a polarity lexicon was constructed using dictionary-based approaches. In this research by Ayvaz et al. [10] it has been tried to extract significant information data using SA to create a Turkish sentiment lexicon for the purpose of contributing to the Turkish SA. In this research, existing SA lexicons are analyzed and a new sentiment lexicon is created by extending the content of these existing lexicons using the retrieved data from social platforms. In addition to the existing lexicons, basic emojis and scoring structure was added to this newly created sentiment lexicon. Furthermore, to evaluate the effectiveness of this newly formed sentiment lexicon, SA was performed on data retrieved from Twitter with specific tags. The SA studies were performed on two topics. The SA methodology for this study is presented in Figure 7.

In [60], Altınel, Buzlu and İpek created two sentiment polarity lexicons for Turkish and implemented statistical-based semantic algorithms in Turkish SA tasks. The first sentiment polarity dictionary, possessing a size of 159,876 Turkish words, is built with the use of a translator. The second sentiment polarity dictionary, with a size of 84,744 Turkish words, is built using GDELT (Global Data on Events, Languages, and Tone). They implemented baseline state-of-the-art models in order to compare the performance between the proposed system and state-of-the-art models. According to the experiment results, the algorithms developed in this study are beneficial because they can achieve higher classification performance than the baseline models on the Turkish

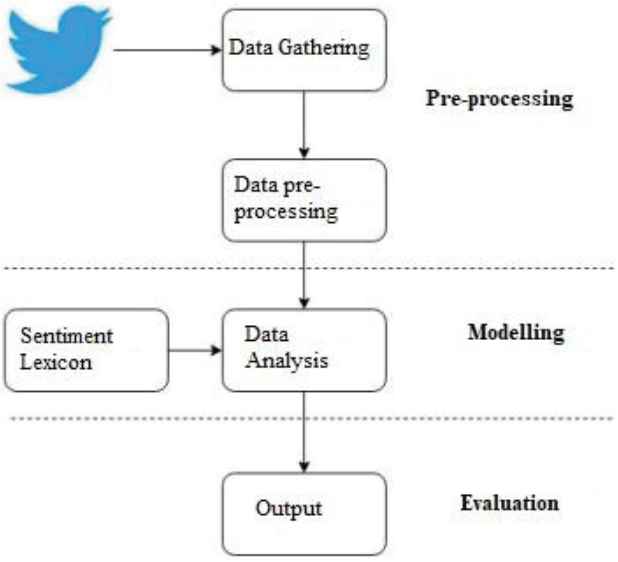

Figure 7. Sentiment analysis of the algorithm.

sentiment polarity detection task. In Figure 8. the representation of the GDELT website is presented.

POPULAR DATASETS AND LEXICONS

Public Datasets

In this section, the publicly available datasets used in the reviewed studies are collected and presented in Table 4. In this table, the information of the dataset type, the year of its creation, the language of the dataset, the size of the dataset, and the source link of the datasets can be found. To retrieve different types of data in multiple languages, the given sources and the table can be used.

Figure 8. The website of GDELT.

Most Used Lexicons in Turkish Sentiment Analysis

In Table 5, the most widely known lexicons in Turkish SA are presented. In this table, the names of the lexicons, their language, size, and information about the polarity state of the lexicons can be found. If a lexicon includes polarities of words, it is called a polarity lexicon. According to this table, the lexicon which has the highest size is the TDK lexicon which includes 616k words. But since the TDK lexicon does not include the information of word polarities, the lexicon has the highest size that also includes the information

Table 4. Public Datasets Used in Sentiment Analysis Studies

Dataset Type	Year	Language	Size	Source
Tweets	2021	Turkish	11k	https://www.kaggle.com/datasets/mrtbeyz/trke-sosyal-medya-paylam-veri-seti
Movie Reviews	2018	Turkish, English	348k	https://doi.org/10.1016/j.tele.2017.10.006
Lexicon	2018	Turkish	5k	GitHub - nazaan/Sentiment-Analysis
Movie Reviews	2017	Turkish	60k	http://myweb.sabanciuniv.edu/rdehkharghani/sentiment-analysis-in-turkish/
Text	2016	English	10k	https://github.com/dennybritz/cnn-text-classification-tf
Lexicon	2019	Turkish	616k	https://sozluk.gov.tr/
Text	2021	Multi-language	2Tb	https://www.gdeltproject.org/
Hotel, Movie Reviews	2016	Turkish	220k	http://humirapps.cs.hacettepe.edu.tr/tsad.aspx
Tweets	2019	Turkish	3k	http://www.kemik.yildiz.edu.tr/veri kumelerimiz.html
Tweets	2014	English	2k	https://drive.google.com/file/d/0BwPSGZHAP
Text	2020	Turkish	157	yoN2pZcVl1Qmp1OEU/view
Movie Reviews	2020	Turkish	105	http://www.kemik.yildiz.edu.tr/veri kumelerimiz.html
Survey(text)	2019	Turkish	27k	http://www.kemik.yildiz.edu.tr/veri kumelerimiz.html
Movie Reviews	2004	English	2k	http://demir.cs.deu.edu.tr/tremo-dataset/
Lexicon	2019	Turkish	13k	https://www.cs.cornell.edu/people/pabo/movie-review-data/
Lexicon	2019	Turkish	43k	https://github.com/sayvaz/turkish-lexicon/
Movie and Product Reviews	2013	Turkish	10k	http://demir.cs.deu.edu.tr/turkish-emotion-lexicon-tel-dataset/

Table 5. Most Used Lexicons in Turkish Sentiment Analysis

Name	Language	Size	Does it İnclude Polarity
SentiTurkNet	Turkish	14k	Yes
SenticNet	English	25k	Yes
Turkish WordNet	Turkish	77k	No
SentiWordNet	English	117k	Yes
TDK	Turkish	616k	No

of word polarities, which is the SentiWordNet lexicon, consisting of 117k words including polarities. Among the lexicons, which include the polarity information of words, the SentiWordNet lexicon has the maximum size which is 117k. Even though SentiWordNet has the largest size, SentiWordNet is an English dictionary. The second lexicon which has the highest size and includes polarity information of words is SentiTurkNet but compared to SentiWordNet, SentiTurkNet is a Turkish lexicon. The lexicon which has the largest size but does not include polarity information of words is TDK.

In Turkish SA, the most widely used lexicon is SentiTurkNet. Even though the TDK is the lexicon that has the largest size among other Turkish lexicons, the TDK library does not include polarities of words, which is why it is not commonly used in Turkish SA studies.

CONTROVERSIES, DISCUSSIONS AND COMPARISON OF THE APPROACHES FOR TURKISH SENTIMENT ANALYSIS

Researchers encounter numerous difficulties while implementing a sentiment text categorization methodology: The presence of base knowledge regarding a certain language: There are only a limited number of lexicon resources accessible for some languages (i.e., English, German, Chinese, etc.), and these languages are widely used languages internationally. In fact, the majority of languages lack their own lexical databases. This fact suggests that these knowledge-based systems can only be constructed for these specific languages. Additionally, knowledge-based systems for a specific language cannot be used effectively for another. Therefore, they are majorly language-dependent systems. The constantly expanding nature of language causes such resources to be typically expensive to be maintained and often unavailable in certain domains. Therefore, researchers should be encouraged to create knowledge bases for the other languages due to its effect on the improvement of the classification performance, or automatic translators should be made available.

The processing complexity of a knowledge base of this size: Knowledge-based automatic systems have substantial processing costs. These systems include the pre-processing part of a big corpus/data. This becomes an extra expense and radically grows according to the proportion of the corpus/data size. Researchers that implement these systems must optimize their algorithms/methodology to decrease the processing time/complexity.

Accessing excessive amounts of unlabeled/labeled data: DL algorithms are especially useful for supervised and unsupervised learning when there is a large amount of unlabeled data, and they typically learn data according to the weights in layers. The amount of data required for taking advantage of DL algorithms is sometimes difficult to access or collect. There are a lot of expenses associated with gathering, storing, and curating these data. Hardware computations: Implementing a machine learning-based system can be computationally expensive. In the training process of a machine learning-based system, the use of costly hardware resources may be a necessity. Prior to starting a project based on machine learning, this fact should be taken into account. Analysis: For a number of classification algorithms based on machine learning (e.g., decision trees, LR, and so on), it is possible to comprehend and explain the learned model, as well as the model's decision. This causes a number of issues in real-life applications. The privacy of the public must be considered by the researchers. To researchers that study SA, we recommend using various data mining methods to apply pre-processing and cleaning the data in their systems.

Furthermore, various ML algorithms can be used to detect and remove irrelevant data from large text corpora. However, in the methodology of the model, deciding to use a rule-based or lexicon-based methodology for sentiment classification is a difficult task that is dependent on certain factors such as the availability and size of the dataset and knowledge bases, as well as the problem itself. Different machine-learning methods have been used to produce a variety of text classification technologies. However, text classification offers more of a challenge since it contains semantic links between words, which are too complex for computers to model. For text categorization, algorithms to achieve high performance, it is essential to extract semantic relations in the correct forms. According to our analysis, selecting the best text sentiment classification method is dependent on the interconnected element size. Each category of approach has advantages over others as well as limitations as described above.

Current Challenges and Research Gaps

SA can be considered a difficult task to perform. SA is language dependent. The difficulty of SA is highly related to the structure of the language that is sentimentally analyzed. For languages similar to Turkish, performing SA can be more challenging compared to other languages due to the structure of the language. SA is a multiprocessing task, in order to perform SA many preprocessing stages must be applied to the data that will be used for SA. For preprocessing, many different techniques can be used to prepare the data for SA. These techniques are normalization, lemmatization, stemming and many more. All of these preprocessing stages are language-dependent as well, meaning that the process for these techniques vary between languages and, due to this reason, many challenges can occur in the process of SA for certain languages. If the preprocessing stages cannot be applied properly, the SA could not be performed effectively or accurately.

Turkish is a complex and rich language; it has an agglutinative structure. For languages that have an agglutinative structure, sentiment classification is a major problem. SA is more challenging to perform in these languages. Aside from the challenges resulting from the morphology and structure of languages, there are other difficulties that can cause major problems to the SA process. Even though SA is a very popular research topic in the Natural Language Analysis field, limited research is conducted on SA for the Turkish language. Due to this reality, there are also limited resources of data or other necessary resources to conduct Turkish SA. Additionally, privacy concerns may arise by the public that must be considered by researchers because the data extracted is often derived from people's posts and ideas from social media.

SA can be performed using various approaches such as lexicon-based, ML-based, and hybrid-based. For different approaches of SA, there are different requirements for resources. For dictionary-based approaches, an external lexicon is needed to perform SA and for ML-based and hybrid-based approaches, a sufficient amount of labeled data is necessary for performing SA. Thus, current resources may not be sufficient to source Turkish SA studies. Finding an adequately-sized Turkish polarity lexicon is difficult. The existing Turkish polarity lexicons are narrow in size. For languages like English, there are a wide variety of resources for SA. This widens the research area because with ready access to proper resources, conducting a SA study can be easier. Even though there are excessive amounts of data in online sources, processing and labeling these data requires sufficient hardware and this process is very time-consuming, contributing to the fact that there are limited resources for data as well. To summarize, the key points of Turkish SA are morphological complexity of Turkish, lack of resources and data privacy and ethics.

CONCLUSION

This paper presents an overview of Turkish SA. Many studies were compiled regarding Turkish SA. These studies were thoroughly reviewed and analyzed. In this paper, Turkish SA studies were categorized under four different categories according to their approaches. These studies are also categorized according to the level of SA performed. The SA studies were categorized into three different categories, these categories are dictionary-based SA, machine learning-based, and hybrid-based SA. Moreover, SA can be performed on four different sentiment levels which are word-level, sentence-level, document-level, and aspect-level. The main purpose of this survey was to collect different Turkish SA studies that use varying techniques and approaches and to analyze and compare the techniques of these studies to detect how different studies approach and propose solutions in the field of Turkish SA. Each Turkish sentiment study is reviewed and analyzed according to the approaches and techniques used to perform SA and the sources of datasets were presented. In each section, the studies are compared according to sentiment level, performance metric, and data size. SA is becoming a very popular research topic. In SA, three different approaches are used; these are lexicon-based, ml-based, and hybrid-based approaches. In summary; Lexicon-based methods offer simplicity and transparency, relying on predefined word lists. While these methods provide a base approach for the SA process, it is a manual process and it may struggle with the Turkish language's rich morphology, complex expressions, and the dynamic nature of language, which restrains their accuracy. Machine learning and deep learning techniques distinguish in capturing context and patterns in Turkish text. They have significantly improved the overall performance of Turkish SA, enabling a finer understanding of sentiments in various contexts, which is crucial for practical SA applications. Using ML techniques would be an improvement over using the lexicon approach alone. Lastly, with hybrid approaches, the aim is to combine the strengths of lexicon-based and ML-based techniques and it can enhance SA by incorporating domain-specific word lists while benefiting from machine learning models.

Using only lexicon-based methods is considered outdated due to all the reasons mentioned above. Instead, machine learning approaches, such as deep learning, are more effective as they have the ability to understand the context and can adapt well to certain situations. Hybrid methods combine ML's power with the benefits of sentiment lexicons and can offer a more balanced solution. Even though the lexicon approaches are outdated, we should not overlook its benefits to the SA process and, when combined with ML techniques, it can even outperform approaches that only employ ML techniques. This shift to the ML approaches is due to ML's increased ability and its future developments to handle the issues we encounter in languages such as idiomatic expressions, language structure etc.

Currently, interest in this field for non-English languages is still growing. In this study, we collected existing studies related strictly to Turkish SA to present a comprehensive study summarizing different approaches, and techniques used in Turkish SA. This survey can be a very helpful resource to gain detailed information about Turkish SA and to obtain general information about Turkish Sentiment analysis.

ACKNOWLEDGEMENT

This work was funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK) with grant number 120E187. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of TÜBİTAK.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

REFERENCES

[1] Girgin A, Gümüşçekiçci G. From past to present: Spam detection and identifying opinion leaders in social networks. Sigma J Eng Nat Sci 2022;40:441-463.

[2] Bilgin M. Classification of Turkish Tweets by Document Vectors and Investigation of the Effects of Parameter Changes On Classification Success. Sigma J Eng Nat Sci 2020;38:1581-1592.

[3] Dehkharghani R, Yanikoglu B, Saygin Y, Oflazer K. Sentiment analysis in Turkish at different granularity levels. Nat Lang Eng 2017;23:535-559. [CrossRef]

[4] Gezici G, Yanıkoğlu B. Sentiment analysis in Turkish. Turkish Nat Lang Process 2018;255-271. [CrossRef]

[5] Yıldırım E, Çetin FS, Eryiğit G, Temel T. The impact of NLP on Turkish sentiment analysis. Türkiye Bilisim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi. 2015;7:43-51.

[6] Alqaraleh S. Turkish sentiment analysis system via ensemble learning. Avrupa Bilim Teknol Derg 2020;122-129. [CrossRef]

[7] Can EF, Ezen-Can A, Can F. Multilingual sentiment analysis: An RNN-based framework for limited data. arXiv Prepr arXiv:1806.04511. 2018.

[8] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng J 2014;5:1093-1113. [CrossRef]

[9] Suat AT, Çınar Y. Borsa İstanbul'da finansal haberler ile piyasa değeri ilişkisinin metin madenciliği ve duygu (sentiment) analizi ile incelenmesi. Ankara Univ SBF Derg 2019;74:1-34. [CrossRef]

[10] Ayvaz S, Yıldırım S, Salman YB. Türkçe duygu kütüphanesi geliştirme: Sosyal medya verileriyle duygu analizi çalışması. Avrupa Bilim Teknol Derg 2019;16:51-60. [CrossRef]

[11] Vural AG, Cambazoglu B, Senkul P, Tokgoz ZO. A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish. In: Computer and Information Sciences III. 2013; Springer; London; 437-445. [CrossRef]

[12] Çoban Ö, Özyer B, Özyer GT. Sentiment analysis for Turkish Twitter feeds. In: 2015 23nd Signal Processing and Communications Applications Conference (SIU). 2015; IEEE; 2388-2391. [CrossRef]

[13] Albayrak M, Topal K, Altıntaş V. Sosyal medya üzerinde veri analizi: Twitter. Suleyman Demirel Univ İktisadi ve İdari Bilimler Fakültesi Derg 2017;22(Kayfor 15 Özel Sayısı):1991-1998.

[14] Karaöz B, Gürsoy UT. Adaptif Öğrenme Sözlüğü Temelli Duygu Analiz Algoritması Önerisi. Bilisim Teknol Derg 2018;11:245-253. [CrossRef]

[15] Yüksel AS, Tan FG. Metin madenciliği teknikleri ile sosyal ağlarda bilgi keşfi. Mühendislik Bilimleri ve Tasarım Derg 2018;6:324-333. [CrossRef]

[16] Akgül ES, Ertano C, Diri B. Twitter verileri ile duygu analizi. Pamukkale Univ J Eng Sci 2016;22(2). [CrossRef]

[17] Aydın CR, Güngör T, Erkan A. Generating word and document embeddings for sentiment analysis. arXiv Prepr arXiv:2001.01269. 2020.

[18] Erşahin B, Aktaş Ö, Kilinc D, Erşahin M. A hybrid sentiment analysis method for Turkish. Turk J Electr Eng Comput Sci 2019;27:1780-1793. [CrossRef]

[19] Türkmenoglu C, Tantug AC. Sentiment analysis in Turkish media. In: International Conference on Machine Learning (ICML). 2014.

[20] Toçoğlu MA, Alpkocak A. Lexicon-based emotion analysis in Turkish. Turk J Electr Eng Comput Sci 2019;27:1213-1227. [CrossRef]

[21] Bayraktar K, Yavanoglu U, Ozbilen A. A Rule-Based Holistic Approach for Turkish Aspect-Based Sentiment Analysis. In: 2019 IEEE International Conference on Big Data (Big Data) 2019; IEEE; 2154-2158. [CrossRef]

[22] Ekinci E, Omurca Sİ. Ürün özelliklerinin konu modelleme yöntemi ile çıkartılması. Türkiye Bilisim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi 2017;9:51-58.
[23] Mutlu M, Özgür A. A Dataset and BERT-based models for targeted sentiment Analysis on Turkish Texts. arXiv Prepr arXiv:2205.04185. 2022. [CrossRef]
[24] Kilimci ZH. Borsa tahmini için Derin Topluluk Modelleri (DTM) ile finansal duygu analizi. Gazi Univ Mühendislik Mimarlık Fakültesi Derg 2020;35:635-650. [CrossRef]
[25] Onan A. Türkçe Twitter Mesajlarında Gizli Dirichlet Tahsisine Dayalı Duygu Analizi. 2017.
[26] Kaynar O, Yıldız M, Görmez Y, Albayrak A. Makine öğrenmesi yöntemleri ile Duygu Analizi. In: International Artificial Intelligence and Data Processing Symposium (IDAP'16). 2016;17-18.
[27] Aytekin YE, Keskin Ö. Türkiye'de faizsiz finans sisteminin duygu analizi bağlamında değerlendirilmesi. Uluslararası Islam Ekonomisi ve Finansı Araştırmaları Dergisi 2019;5:87-112.
[28] Kaya M, Fidan G, Toroslu IH. Sentiment analysis of Turkish political news. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. 2012; IEEE; Vol. 1; 174-180. [CrossRef]
[29] Ciftci B, Apaydin MS. A deep learning approach to sentiment analysis in Turkish. In: 2018 International Conference on Artificial Intelligence and Data Processing (IDAP). 2018; IEEE; 1-5. [CrossRef]
[30] Shehu HA, Tokat S. A hybrid approach for the sentiment analysis of Turkish Twitter data. In: The International Conference on Artificial Intelligence and Applied Mathematics in Engineering. Cham: Springer; 2019. p. 182-190. [CrossRef]
[31] Acikalin U, Bardak B, Kutlu M. Turkish Sentiment Analysis Using BERT. In: 2020 28th Signal Processing and Communications Applications Conference (SIU). 2020; IEEE; 1-4. [CrossRef]
[32] Aksu MÇ, Karaman E. FastText ve kelime çantası kelime temsil yöntemlerinin turistik mekanlar için yapılan Türkçe incelemeler kullanılarak karşılaştırılması. Avrupa Bilim Teknol Derg 2020;20:311-320. [CrossRef]
[33] Ballı C, Guzel MS, Bostanci E, Mishra A. Sentimental analysis of Twitter users from Turkish content with natural language processing. Comput Intell Neurosci 2022:2455160. [CrossRef]
[34] Baykara M, Gürtürk U. Sosyal Medya Paylaşımlarının Duygu Analizi Yöntemiyle Sınıflandırılması. In: 2nd International Conferance on Computer Science and Engineering 2017;911-916.
[35] Tuncer T, Çetintaş D. Bir sosyal ağdan alınan verilerin anlamsal kutuplandırılması. Bilgisayar Bilim 2019;4:1-6.
[36] Güran A, Uysal M, Doğrusöz Ö. Destek vektör makineleri parametre optimizasyonunun duygu analizi üzerindeki etkisi. Dokuz Eylül Univ Müh Fak Fen Müh Derg 2014;16:86-93.
[37] Aytekin Ç, Bayram MA. Türkçe metinler için duygu analizi yaklaşımı ile iletişimde bağlamdan bağımsız modellerin geliştirilmesi üzerine bir araştırma: karma veri modeli önerisi. Yeni Medya Elektron Derg 2021;5:12-25.
[38] Aytuğ O. Twitter mesajları üzerinde makine öğrenmesi yöntemlerine dayalı duygu analizi. Yönetim Bilisim Sistemleri Derg 2017;3(2):1-14.
[39] Sarıman G, Mutaf E. Covid-19 sürecinde Twitter mesajlarinin duygu analizi. Euroasia J Math Eng Nat Med Sci 2020;7:137-148. [CrossRef]
[40] Karabulut YE, Küçüksille EU. Twitter profesyonel izleme ve analiz aracı. Teknik Bilim Derg 2018;8:17-24.
[41] Kaynar O, Aydın Z, Görmez Y. Sentiment analizinde öznitelik düşürme yöntemlerinin oto kodlayıcılı derin öğrenme makinaları ile karşılaştırılması. Bilisim Teknol Derg 2017;10:319-326. [CrossRef]
[42] Alpkoçak A, Tocoglu MA, Çelikten A, Aygün İ. Türkçe Metinlerde Duygu Analizi için Farklı Makine Öğrenmesi Yöntemlerinin Karşılaştırılması. Dokuz Eylül Univ Müh Fak Fen Müh Derg 2019;21:719-725. [CrossRef]
[43] Tuzcu S. Çevrimiçi kullanıcı yorumlarının duygu analizi ile sınıflandırılması. Eskisehir Turk Dünyası Uygulama ve Araştırma Merkezi Bilisim Derg 2020;1:1-5.
[44] Kemaloğlu N, Küçüksille E, Özgünsever ME. Turkish Sentiment Analysis on Social Media. Sakarya Univ J Sci 2021;25:629-638. [CrossRef]
[45] Bharadiya J. A comprehensive survey of deep learning techniques natural language processing. Eur J Technol. 2023;7:58-66. [CrossRef]
[46] Çetin FS, Eryiğit G. Türkçe hedef tabanlı duygu analizi için alt görevlerin incelenmesi-hedef terim, hedef kategori ve duygu sınıfı belirleme. Bilisim Teknol Derg 2018;11:43-56. [CrossRef]
[47] Onan A. Sentiment Analysis in Turkish Based on Weighted Word Embeddings. In: 2020 28th Signal Processing and Communications Applications Conference (SIU). 2020; IEEE; 1-4. [CrossRef]
[48] Aktaş Ö, Coşkuner B, Soner İ. Turkish Sentiment Analysis Using Machine Learning Methods: Application on Online Food Order Site Reviews. arXiv Prepr arXiv:2201.03848. 2022.
[49] Seker SE. Sentimental analysis. YBS Ansiklopedi. 2016;3:21-36.
[50] Kılıç G, Budak I, Kılıç BS. Kara cuma etiketlerinin Tweet istatistikleri ve duygu analizi ile sıralanması. Selçuk Univ Sosyal Bilimler Meslek Yüksekokulu Derg 2020;23:131-140. [CrossRef]

[51] Aydın İ, Salur MU, Başkaya F. Duygu analizi için çoklu popülasyon tabanlı parçacık sürü optimizasyonu. Türkiye Bilisim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi. 2018;11:52-64.

[52] Demir Ö, Chawai AIB, Doğan B. Türkçe metinlerde sözlük tabanli yaklaşimla duygu analizi ve görselleştirme. Int Period Recent Technol Appl Eng 2019;1:58-66. [CrossRef]

[53] Alqaraleh S. Novel Turkish Sentiment Analysis System using ConvNet. 2021. [CrossRef]

[54] Shehu HA, Sharif MH, Sharif MHU, Datta R, Tokat S, Uyaver S, et al. Deep sentiment analysis: a case study on stemmed Turkish Twitter data. IEEE Access 2021;9:56836-56854. [CrossRef]

[55] Salur MU, Aydin I. A novel hybrid deep learning model for sentiment classification. IEEE Access 2020;8:58080-58093. [CrossRef]

[56] Guven ZA. The Comparison of Language Models with a Novel Text Filtering Approach for Turkish Sentiment Analysis. Trans Asian Low-Resource Lang Inf Process 2022;55:1-6. [CrossRef]

[57] Sonmezoz K, Amasyali MF. Same sentence prediction: A new pre-training task for BERT. In: 2021 Innovations in Intelligent Systems and Applications Conference (ASYU). 2021; IEEE; 1-6. [CrossRef]

[58] Sağlam F, Sever H, Genç B. Developing Turkish sentiment lexicon for sentiment analysis using online news media. In: 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA). 2016; IEEE; 1-5. [CrossRef]

[59] Köksal A, Özgür A. Twitter dataset and evaluation of transformers for Turkish sentiment analysis. In: 2021 29th Signal Processing and Communications Applications Conference (SIU). 2021; IEEE; 1-4. [CrossRef]

[60] Altınel AB, Buzlu K, İpek K. Performance analysis of different sentiment polarity dictionaries on Turkish sentiment detection. In: 2022 International Conference on Innovations in Intelligent Systems and Applications (INISTA). 2022;1-6. [CrossRef]