AITA
ARTIFICIAL INTELLIGENCE THEORY AND APPLICATIONS

■ RESEARCH ARTICLE

Sentiment Analysis in Turkish Tweets Using Different Machine Learning Algorithms

Hunaida Avvad [a†], **Ecem Ereren** [a]

[a] Department of Management Information Systems, İzmir Bakırçay University, İzmir, Türkiye
[†] hunaida.awwad@bakircay.edu.tr, corresponding author

RECEIVED AUGUST 12, 2024
ACCEPTED SEPTEMBER 19, 2024

CITATION Avvad, H., & Ereren, E. (2024). Sentiment analysis in Turkish tweets using different machine learning algorithms. *Artificial Intelligence Theory and Applications, 4(*2), 107-120.

Abstract

Understanding emotions in any written text is considered a hot topic for many researchers in the field of text mining, especially with the large contribution of users over the web 2.0 and with the growth of the different social media platforms. In this study, we analysed emotions in Turkish text and studied the sentiment within each document using sentiment analysis techniques. Sentiment analysis is the process of identifying and evaluating the emotional states contained in texts. This study aimed to investigate the effect and accuracy rate of sentiment analysis in Turkish texts. Sentiment analysis is an important field of research that helps to obtain important data in many areas, such as marketing, social media analysis, and customer feedback. A comprehensive data set consisting of Turkish tweets from Kaggle was used, and the emotional states of the texts were labelled. This data set consists of a variety of tweets with different topics and emotional tones. Using natural language processing techniques and machine learning algorithms, the data set was processed, and the model was trained. Within the scope of the study, different root extraction methods and a vector space model were used. In addition, machine learning algorithms such as Naive Bayes, Random Forest, Decision Tree, Gradient Boosting, Bernoulli Naive Bayes, Logistic Regression, K-Neighbours-Classifier, and Support Vector Classifier were applied to evaluate accuracy. This study aims to emphasize the importance of sentiment analysis in Turkish texts, examine the impact of the methods used, and form a basis for future studies.

Keywords: sentiment analysis, Turkish text, machine learning, Turkish tweet

1. Introduction

Social media platforms and Web 2.0 allowed people to share their experience and to express their feedback about many products/services that they received, the huge size of the written text on the Web 2.0 is considered a hot research topic for many researchers who focus on text mining in order to analyse emotions in any written text is considered a hot topic for many researchers in the field of text mining. Social media tools such as Twitter and Facebook have an important role to play as big data sources in the process of extracting information from any text. The most important reason for this is that the text data produced by these applications is increasing significantly day by day. Sentiment Analysis (SA) has emerged as a field in which natural language processing, machine learning, and linguistic methods are used to understand the emotional tone of texts and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than AITA must be honoured. Abstracting with credit is permitted, and providing the material is not used for any commercial purposes and is shared in its entire and unmodified form. Request permissions from info@aitajournal.com

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN: 978-605-69730-2-4 © 2024 İzmir Bakırçay University

identify emotional trends or moods in the text. SA can be applied in many areas, such as social media analytics, customer feedback, marketing strategies, product reviews, and survey responses. For example, a company may try to understand customer satisfaction and perception by analysing tweets about their brand or products on social media platforms. Similarly, a movie studio can gauge the overall emotional response of movies by analysing the comments that viewers share about the movies on social media. It is of great importance to analyse the big data, which has emerged with the increase in the use of the internet and social media, which has become widespread today, and to transform it into meaningful information. SA is the process of systematically examining the data containing opinions in a text and determining the emotion category and emotion polarity of the text. SA approaches are frequently used not only in linguistics, but also in many different fields such as financial markets, marketing, and social media analysis, Tokcaer [1]. In SA studies, it is questioned and analysed whether the texts have positive, negative, or neutral content. According to the results of this analysis, the attitude of individuals or a certain group about the subject related to the study is determined. In this respect, SA can guide businesses on issues such as preliminary market research for a new product to be launched, how a decision to be taken for a community will receive a positive or negative reaction, and whether people who will watch the movie decide to watch the movie according to previous comments. However, the large amount of data from which a positive or negative opinion can be obtained makes it difficult to make this analysis by examining it one by one. Therefore, sentiment analysis has become one of the most important and studied topics in the fields of text mining and machine learning, Kaynar et al. [2]. Danisman and Alpkocak [3] used different machine learning algorithms such as Vector Space Model, Naïve Bayes, and SVM classifiers to compare between the performances using the ISEAR data set that contains 5 classes of emotions: anger, disgust, fear, joy, and sadness. The training set is enriched with WordNet Affect and WPARD (Wisconsin Perceptual Attribute Rating Database) data sources, Medler et al. [4]. Stop word removal and root removal operations were applied, and the term frequency – inverse document frequency (TF-IDF) method was chosen as feature weighting. According to the results obtained, an overall classification accuracy of 70.2% was achieved. There are many studies on sentiment analysis on text data sources, especially in English. Sentiment analysis generally uses two basic approaches: the rule-based approach and the machine learning-based approach, Alpkoçak et al. [5]. Rule-based approaches use predefined rules that include certain emotional words or phrases that indicate a particular emotion. Machine learning-based approaches, on the other hand, use large data sets to automatically learn how emotional expressions relate to specific emotions and create classification models. Identifying and classifying emotional expressions in Turkish tweets is an important issue among text classification problems. Social media platforms, which are easily accessible platforms, provide remarkable resources to get feedback from target audiences, but it is impossible to analyse these feedbacks with human labour. Therefore, automated sentiment analysis tools are essential for companies' customer service to be able to capture complaints and/or positive feedback at the right time. Processing by computer allows this data to be used in the market, Türkmenoğlu [6].

Studies such as Kozareva et al. [7], Mohammad [8], and Chaffar and Inkpen [9] have carried out important contributions on sentiment analysis. In the Turkish language, there are limited number of studies on the subject of sentiment analysis, Boynukalın [10]. In the Turkish language, there are many to be done because of the lack of studies dealing with SA different aspects and subdomains. The Turkish language is known as a sticky language. With the use of derived suffixes, the root of a word can be transformed into a completely different type of word, for example, from a noun to a verb. These derivatives can be applied consecutively more than once. Since each derived suffix has the potential to change the meaning of a word, each derived suffix must be examined separately to obtain the true meaning of a word. Previous studies have often focused on official data sources, such as newspaper headlines and surveys. Recently, however, research on informal data sources such as instant messaging, blog posts, and Twitter has become popular. Twitter is a social micro-blogging service that allows users to publish and read messages in real time, and these messages are called "tweets". People share their thoughts, daily life events, and feelings on Twitter. Although there are many micro-blogging platforms, Twitter is the most popular. The large volume of user-generated content makes Twitter a suitable space for sentiment analysis. The similarity of the tweets also makes them effectively actionable for sentiment detection tasks. Boynukalın [10] used a translation of the ISEAR data set and a manually marked data set to classify the Turkish texts. Apart from emotion classes, the determination of emotion levels was also attempted. Different combinations of n-gram features were used. A weighted log probability algorithm was used to score the features and identify the most important ones. Kaya et al. [11] applied supervised classification algorithms for the sentiment classes positive and negative in Turkish news columns. With the exception of SVM, Maximum Entropy, and Naïve Bayes classifiers, the character-based n-gram language model was used. This language model uses characters instead of words as a unit. Their idea is that statistical methods may not yield promising results due to the fact that Turkish is a morphologically rich language.

In this study, we analysed emotions in Turkish text and studied the sentiment within each document using Sentiment Analysis (SA) techniques. This article aims to show the different methods and approaches that can be used to evaluate sentiment analysis and classify emotional expressions in Turkish texts. This study was carried out in order to evaluate the effectiveness of existing methods and algorithms for sentiment analysis in Turkish and to propose new approaches to obtain better results for sentiment analysis in Turkish texts. In this study, a ready-made Turkish tweet data set downloaded from the Kaggle platform was used. The data set consists of Turkish tweets shared by different topics and users. The distribution of classes in the used data set was not balanced. For imbalanced data sets, some classification algorithms may perform better; for example, Decision Trees, Random Forests, or Gradient Boosting models have been used in this data set because they can work well in unbalanced data sets. We also split our data set into training and testing data, making it available for training and testing our machine learning models. After performing the data preprocessing steps, we used the processed data for building the ML models, testing them, and evaluating them one by one.

After carrying out the needed preprocessing steps, a classification model was developed to conduct sentiment analysis using various machine learning algorithms. These algorithms include popular methods such as Support Vector Machines (SVM), Decision Trees (DT), and Random forests (RF), in addition to, Logistic Regression (LR), K-nearest Neighbours (KNN), and Gradient Boosting (GB). The developed models were used to classify emotional expressions in Turkish tweets as positive, negative, or neutral. During the training process, the data set was split, and the performance of the models was evaluated using the five-fold cross-validation method. Performance metrics such as accuracy rates and confusion matrices were calculated, and the results were analysed. In addition, the in-class performance values of the model were calculated.

2. Literature review

In recent years, many studies have been carried out on sentiment analysis in Turkish texts. Especially considering that most of the social media data language used in Turkey is expressed using Turkish language, that leads to the increase of the importance of sentiment analysis in Turkish texts. In many studies, various approaches have been adopted to detect and classify emotional expressions in Turkish texts by using different methods such as machine learning algorithms, natural language processing techniques, and deep learning models. The researchers evaluated the results obtained using performance metrics such as accuracy, precision, recall and F1 score, and proposed new methods and improvements to increase the success of sentiment analysis in Turkish texts.

There are many studies on SA, especially in English, on text data sources. Researchers such as Kozareva et al. [7], Mohammad [8] and Chaffar and Inkpen [9] have done important studies on this subject. In Turkish, there are fewer studies on the subject of SA, Boynukalın [10]. In the Turkish language, more work has been done on Natural Language Processing (NLP) rather than sentiment analysis because NLP field is more developed. Danisman and Alpkocak [3] compared the performances of Vector Space Model, Naïve Bayes and SVM classifiers using ISEAR dataset for 5 emotion classes: anger, disgust, fear, joy, and sadness. The training set is enriched with WordNet Affect and WPARD (Wisconsin Perceptual Attribute Rating Database) data sources, Medler et al. [4]. Stop word removal and root removal operations were applied, and the TF-IDF method was chosen as feature weighting. According to the results obtained, an overall classification accuracy of 70.2% was achieved.

The Turkish language is known as a sticky language. Using derived suffixes, the root of a word can be transformed into a completely different type of word, for example, from a noun to a verb. These derivations can be applied sequentially more than once, Oflazer [12]. Since each derived suffix has the potential to change the meaning of the word, each derived suffix must be examined separately to get the true meaning of a word. Previous research has often focused on official data sources such as newspaper headlines and surveys. Recently, however, research on informal data sources such as instant messaging, Neviarouskaya [13], blog posts, Wang [14] and Twitter, Mohammad [8] has become popular. Twitter is a social micro-blogging service that allows users to post and read messages in real time, and these messages are called "tweets". People share their thoughts, daily life events and feelings on Twitter. Although there are many micro-blogging platforms, Twitter is the most popular. The sheer volume of user-generated content makes Twitter a viable space for sentiment analysis. The similarity of tweets also makes them effectively workable for emotion detection tasks. Boynukalın [10] used a translation of the ISEAR dataset and a manually marked dataset to classify Turkish texts. Apart from emotion classes, determination of emotion levels was also attempted. Different combinations of n-gram features were used.

Kaya et al. [11] applied supervised classification algorithms for positive and negative emotion classes in Turkish news columns. Except for SVM, Maximum Entropy\, and Naïve Bayes classifiers, the character-based n-gram Language Model was used. This language model uses characters instead of words as units. Their thoughts are that statistical methods may not yield promising results because Turkish is a morphologically rich language. Erogul [15] created a dataset from a Turkish movie review site. Reviews were labelled by their authors with positive, negative, or neutral icons. In the generated dataset, an emotionally labelled data item was created by using the text and symbol of the review together. A polarity dataset was created from another movie review site, which includes the ratings given to the movies by the users. Combinations of n-grams and POS information for the classification task were used for morphological analysis using the Zemberek tool. For the polarity labelled dataset, scores were estimated using regression and single comparison techniques.

3. Methodology

We used Python with the data set downloaded from Kaggle to detect and evaluate different ML algorithms. In this study, various ML algorithms have been adopted to detect and classify emotional expressions in Turkish texts after applying the data pre-processing steps. We evaluated the results obtained using performance metrics such as accuracy, precision, recall, and F1 score, and proposed new methods and improvements to increase the success of sentiment analysis in Turkish texts.

3.1. Data Collection and Pre-Processing

In this study, we used a Turkish tweet data set from the Kaggle platform to perform sentiment analysis in Turkish texts. The data set consists of a variety of tweets with different topics and emotional tones. We obtained this data set, which included a total of 4201 tweets, and labelled the emotional states.

In the data pre-processing phase, we removed unnecessary characters, special symbols, and punctuation marks from the texts. In addition, we converted the texts to lowercase and performed stemming by using Turkish language processing libraries. Thus, we brought it to a simpler format without affecting the meaning of the texts and purified it from unnecessary information. Stemming is a text processing technique used in Natural Language Processing (NLP) and Computer Language Processing (CLP). Basically, it aims to extract word roots, or the basic form of the word. This is used to identify similarity between different variations or trends of a word and make text analysis or information extraction easier and more effective.

3.2. Text Representation

In order for the texts to be processed with ML algorithms, text must be converted to a vector format. There are three different ways that can help in text representations; bag of words (BOW), n-grams, and Term Frequency-Inverse Document Frequency (TF-IDF). In the bag of words, each word in a document will be added to a bag without any repetition and without keeping the sequence of words existence in the document. A matrix of 1Xn in which n represents the number of words in the bag will be used with a word frequency; words with higher occurrences show that they are more common in the document. One drawback of BOW that words in stop words list are included in the bag without removing them, and they have the highest frequencies, which will affect the vectors negatively. Another drawback of BOW is that it can’t perform well when you have similar documents with small changes. In the n- gram method, the grams (words) will be treated as 2- gram or commonly called bigram, so the model will check the frequency of words in a document as pairs. N-gram can keep the relationship between the consecutive words better than BOW but because of data sparsity n-gram can fail in building a good model specially with there is low frequencies of the n-grams.

For the aforementioned drawbacks of both BOW and n-grams, we applied text representation methods such as TF-IDF to provide vector representation of Turkish texts. In particular, by obtaining TF-IDF word vectors, we aimed to better capture the meaning and emotional content in the texts. TF-IDF is a text mining technique used in NLP applications such as text mining, text classification, and information extraction. TF-IDF is used to determine the importance of a term within a given document and to compare the importance of those terms within a given collection of documents. The main purpose of TF-IDF is to determine the weight and importance of terms between textual documents. TF-IDF increases the applicability of text mining algorithms.

We applied the TF-IDF technique to the processed text after the pre-processing phase and used it as a feature extraction method to convert text documents into numerical data. Term frequency is calculated by finding the number of occurrences of each term (word) in a document, then it will be multiplied by the inverse document frequency, which represents how common a word is in the corpus.

3.3. Machine Learning Algorithms

After completing the data pre-processing and text representation steps, we performed SA by using different machine learning algorithms in the training and evaluation processes. Within the scope of our experiments, we evaluated popular classification algorithms such as Random Forest (RF), Logistic Regression (LG), K-Nearest Neighbours (KNN), Bernoulli Naive Bayes (BNB), Decision Tree (DT), and Support Vector Classifier (SVC).

Random Forest (RF): RF is an algorithm that produces and classifies multiple decision trees by training each one on a different observation sample. The algorithm creates a decision tree for each sample, and the estimated value result of each decision tree is formed. Voting is performed for each value formed as a result of the prediction. Observation is assigned to the class with the most votes.

Logistic Regression (LR): LR is a supervised ML classification algorithm that aims to predict the probability that an instance belongs to a given class or not. Then the data point will be assigned to the class with the highest probability.

K-Nearest Neighbours (KNN) Classifier: KNN classifier classifies using similar samples around labelled data points, KNN is based on deciding the class of the data point depending on the class that is nearest neighbours of the vector. K here represents how many neighbour points we are going to check. The distance will be calculated between the data point of that we want to assign its class and the K nearest points.

Bernoulli Naive Bayes (BNB) Classifier: NB classifier is a probability-based classification algorithm based on Bayes' theorem that makes use of Bayes Theorem during the training phase.

Gradient Boosting (GB): is a famous boosting algorithm using ensemble learning methods that enhance the results of training model sequentially that each model will enhance the previous one.

Decision Tree (DT): DT is one of the tree-based learning algorithms. It is a tree structure that performs classification by dividing the data set according to its characteristics.

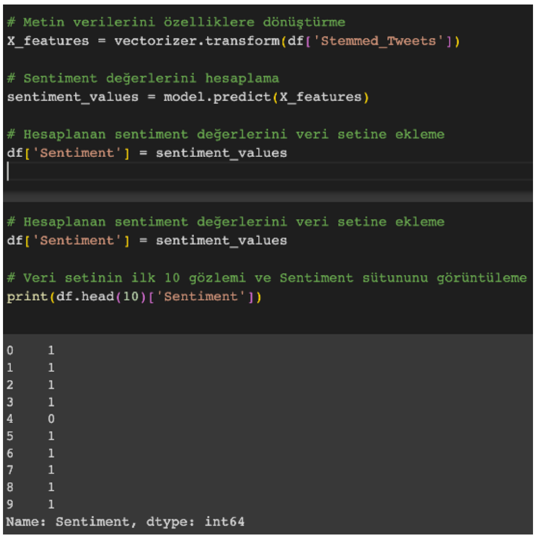
Support Vector Classifier (SVC): SVC classifier classify with the supervised learning method. It is considered as a powerful classification method that attempts to find a hyperplane with a maximum margin between different classes. It aims to have this line at the maximum distance for the points of both classes.

3.4. Model Performance and Evaluation

We performed model training and evaluation for each algorithm with the five-fold cross-validation method. Thus, we have increased the reliability of model performance and prevented problems such as overfitting. Using confusion matrix, we evaluated number of evaluation metrics such as accuracy, precision, recall and F1 score. By comparing our results, we tried to identify the most effective and successful algorithms. Using these methods, we aim to achieve successful results in SA in Turkish texts. We apply these methods to understand the effectiveness of different ML algorithms in SA and to classify emotional content in Turkish texts more accurately and effectively.

4. Experiments and Results

In this study, we evaluated different ML algorithms to perform SA in Turkish texts, and the results we obtained were quite remarkable. Experiments conducted on various tweets in our data set provided important insights into the effectiveness and accuracy rates of the algorithms used for SA.



```
# Metin verilerini özelliklere dönüştürme
X_features = vectorizer.transform(df['Stemmed_Tweets'])

# Sentiment değerlerini hesaplama
sentiment_values = model.predict(X_features)

# Hesaplanan sentiment değerlerini veri setine ekleme
df['Sentiment'] = sentiment_values
```

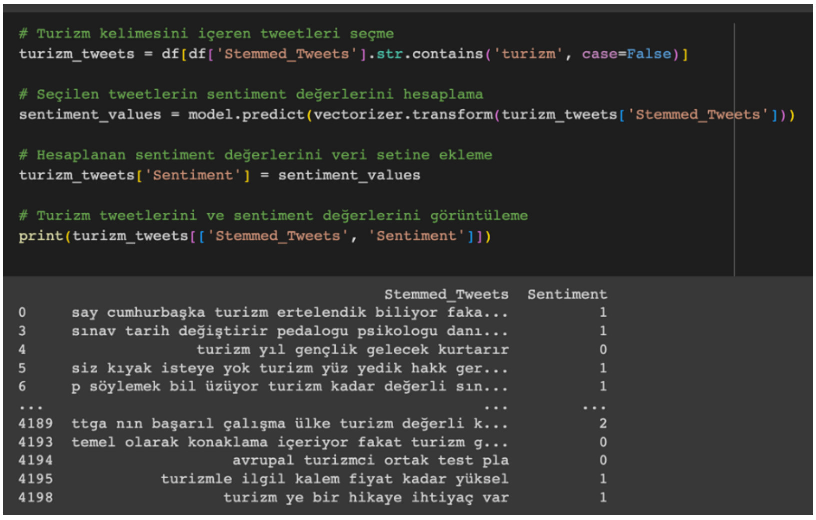
```
# Hesaplanan sentiment değerlerini veri setine ekleme
df['Sentiment'] = sentiment_values

# Veri setinin ilk 10 gözlemi ve Sentiment sütununu görüntüleme
print(df.head(10)['Sentiment'])
```

```
0    1
1    1
2    1
3    1
4    0
5    1
6    1
7    1
8    1
9    1
Name: Sentiment, dtype: int64
```

Figure 1. Calculation of Sentiment Values of the Data

The code in Figure1 uses a vectorizer to convert text data from a data set into features and calculates sentiment values using these features. It then adds the calculated sentiment values to the df data set and then prints the first 10 observations of the data set and the "Sentiment" column on the screen.



```
# Turizm kelimesini içeren tweetleri seçme
turizm_tweets = df[df['Stemmed_Tweets'].str.contains('turizm', case=False)]

# Seçilen tweetlerin sentiment değerlerini hesaplama
sentiment_values = model.predict(vectorizer.transform(turizm_tweets['Stemmed_Tweets']))

# Hesaplanan sentiment değerlerini veri setine ekleme
turizm_tweets['Sentiment'] = sentiment_values

# Turizm tweetlerini ve sentiment değerlerini görüntüleme
print(turizm_tweets[['Stemmed_Tweets', 'Sentiment']])
```

```
                                        Stemmed_Tweets  Sentiment
0     say cumhurbaşka turizm ertelendik biliyor faka...          1
3     sınav tarih değiştirir pedalogu psikologu danı...          1
4                   turizm yıl gençlik gelecek kurtarır          0
5     siz kıyak isteye yok turizm yüz yedik hakk ger...          1
6     p söylemek bil üzüyor turizm kadar değerli sın...          1
...                                                 ...        ...
4189  ttga nın başarıl çalışma ülke turizm değerli k...          2
4193  temel olarak konaklama içeriyor fakat turizm g...          0
4194                     avrupal turizmci ortak test pla          0
4195              turizmle ilgil kalem fiyat kadar yüksel          1
4198                   turizm ye bir hikaye ihtiyaç var          1
```

Figure 2. Sentiment Values of the Data Containing the Word ‘turizm’

The code in Figure 2 selects tweets that contain the word tourism; the number of tweets that contain the word ‘turizm’ is 840, calculates the sentiment values of those tweets, adds these values to the tourism tweets sub data set, and finally displays the tourism tweets and sentiment values. The experiments that have been included in the study are applied to the sub data set of size 840.

The confusion matrix allows us to evaluate the performance of the classification model in more detail. However, based solely on the results of this matrix, it is difficult to determine with certainty how good the performance of the model is. The confusion matrix can be used to understand how the model performs in certain classes, but it needs to be considered in conjunction with other performance metrics in order to fully evaluate performance. To evaluate the confusion matrices in more details, we calculated the performance metrics of each class, such as precision, recall, and F1 score. We can also evaluate these metrics on a class-by-class basis to understand the performance differences between classes.

According to the results of the experiments shown in Table1, Logistic Regression (LR) and Support Vector Classifier (SVC) were found to have the highest accuracy rates of SA in Turkish texts (0.62). Random Forest (RF) and Bernoulli Naïve Bayes (BNB) accuracies are almost similar to Logistic Regression (LR) and Support Vector Classifier (SVC) with accuracy rate of 0.61. However, the K-Nearest Neighbours (KNN) algorithm achieved a slightly lower accuracy rate with 0.48 compared to other methods. To be able to decide which algorithm(s) work better than the others, a confusion matrix analysis was performed for all experiments. Table2 shows the confusion matrices for RF, LR, KNN, BNB, GB, DT, and SVC respectively. The confusion matrix here is a multi-class model of size 3X3, Negative class, Positive class, and Neutral class. Using the confusion matrix for each experiment we calculated Accuracy, Precision, Recall, and F1-score evaluation metrics for each class as shown in Table 3. From Table 3 we can see that the best achieved result is for BNB’s Neutral class with 0.75 accuracy rate followed by GB’s Neutral class with 0.74 accuracy rate. For the Positive class, the best achieved accuracy is for SVC with 0.74 followed by LR with 0.71. For the Negative class, LR and DT have the highest accuracy rate with 0.54 and 0.45, respectively.

F1- scores measures the harmonic mean of the Precision and Recall, Table3 shows that the F1-score for Positive mood class is higher than the F1-score for both the Negative and the Neutral classes for all experiments except KNN. The best achieved F1-score is for BNB’s Positive class with 0.72, followed by SVC’s Positive class with 0.71.

Table 1. Summary Table

Experiments	Accuracy	Precision	Recall	F1 score
Random Forest (RF)	0.61	0.62	0.61	0.61
Logistic Regression (LR)	0.62	0.62	0.62	0.62
K-nearest Neighbors (KNN)	0.48	0.48	0.48	0.48
Bernoulli Naïve Bayes (BNB)	0.61	0.62	0.61	0.61
Gradient Boosting (GB)	0.55	0.58	0.55	0.53
Decision Tree (DT)	0.57	0.57	0.57	0.57
Support Vector Classifier (SVC)	0.62	0.63	0.62	0.61

Table 2. Confusion Matrix for All Experiments

Experiment	Negative	Positive	Neutral
Random Forest (RF)			
Negative	101	32	98
Positive	25	204	83
Neutral	43	39	215
Logistic Regression (LR)			
Negative	125	44	62
Positive	44	218	50
Neutral	67	49	181
K-nearest Neighbors (KNN)			
Negative	92	68	71
Positive	61	150	101
Neutral	66	65	166
Bernoulli Naïve Bayes (BNB)			
Negative	81	31	119
Positive	29	215	68
Neutral	36	37	224
Gradient Boosting (GB)			
Negative	62	31	138
Positive	16	179	117
Neutral	27	49	221
Decision Tree (DT)			
Negative	105	38	88
Positive	44	202	66
Neutral	58	65	174
Support Vector Classifier (SVC)			
Negative	87	46	98
Positive	22	231	59
Neutral	27	59	211

5. Discussion and Conclusion

In this study, we conducted a series of experiments using different ML algorithms and text features for sentiment analysis. Table 1 represents a summary of our experiments and shows that Logistic Regression (LR) and Support Vector Classifier (SVC) have the highest F1 score with 0.62 followed by Random Forest (RF) and Bernoulli Naïve Bayes (BNB) with 0.61 F1 score. Table3 represents Accuracy, Precision, Recall, and F1 score performance metrics for each class separately; Positive class, Negative class, and Neutral class for the algorithms conducted in the study. Table3 shows that the best achieved F1 score result is for BNB algorithm for the positive class with 0.72 followed by SVC with 0.71 for the positive mood as well. Positive class's lowest F1 score is for KNN with 0.50 followed by GB with 0.63. For the Negative class, the best achieved F1 score was for LR with 0.54 followed by RF with 0.51, and the lowest performance was for GB with 0.37 F1 score followed by KNN with 0.41. For the Neutral class, the highest F1 score was for SVC and BNB with 0.64 followed by RF with 0.62, and the lowest F1 score was for KNN with 0.52 followed by DT with 0.56. SVC, BNB, and DT worked better with the Positive mood tweets, while LR worked better with the Negative mood tweets, BNB, GB, and SVC worked better for the Neutral mood tweets.

To discuss the experiments in detail, we firstly used the RF algorithm; this algorithm classifies text data by combining many decision trees after converting them into vector space. From Table 3, we observed that the RF model has one of the lowest accuracies in the study for the Negative class with 0.34. However, it achieved one of the highest accuracies as show in Table1 and achieved a moderate performance for the Positive class with 0.53 and Neutral class with 0.45 as shown in Table3. This can be attributed to the class imbalance in the data set and the differences in the characteristics of different classes. For example, in the confusion matrix in Table2, the RF model, Positive class (204) appears to have a moderate level compared with other algorithms. However, we can see that there are also incorrect predictions for Negative class and Neutral class as well.

Table 3. Performance Metrics for Negative, Positive, and Neutral Classes

Experiment	Accuracy	Precision	Recall	F1 score
Random Forest (RF)				
Negative Class	0.34	0.60	0.44	0.51
Positive Class	0.53	0.74	0.65	0.70
Neutral Class	0.45	0.54	0.72	0.62
Logistic Regression (LR)				
Negative Class	0.54	0.53	0.54	0.54
Positive Class	0.70	0.70	0.70	0.70
Neutral Class	0.61	0.62	0.61	0.61
K-nearest Neighbors (KNN)				
Negative Class	0.40	0.42	0.40	0.41
Positive Class	0.48	0.53	0.48	0.50
Neutral Class	0.56	0.49	0.56	0.52
Bernoulli Naïve Bayes (BNB)				
Negative Class	0.35	0.56	0.35	0.43
Positive Class	0.69	0.76	0.69	0.72
Neutral Class	0.75	0.55	0.75	0.64
Gradient Boosting (GB)				
Negative Class	0.27	0.59	0.27	0.37
Positive Class	0.57	0.69	0.57	0.63
Neutral Class	0.74	0.46	0.74	0.57
Decision Tree (DT)				
Negative Class	0.45	0.51	0.45	0.48
Positive Class	0.65	0.66	0.65	0.66
Neutral Class	0.59	0.53	0.59	0.56
Support Vector Classifier (SVC)				
Negative Class	0.38	0.64	0.38	0.47
Positive Class	0.74	0.69	0.74	0.71
Neutral Class	0.71	0.57	0.71	0.64

Next, we used the LR algorithm. This algorithm classifies text data with a linear model. The best achieved accuracy among all experiments goes for this algorithm with 0.62 shared with SVC (Table1). From the detailed analysis for each class shown in Table3, we observed that the LR model has the second highest accuracy for the Positive class with 0.70, and the second highest accuracy also for the negative class with 0.54, and the fourth highest accuracy for the Neutral class with 0.61. From the statistics, we can see that this algorithm works better with classes that carry emotions; this may be due to the class distribution and the fact that the model has a linear classification capability.

We also tried the KNN algorithm. This algorithm classifies a new data point based on the majority of its closest neighbours. In our experimental results, we observed that the KNN model has lower accuracy values for all classes compared with the first other algorithms. The accuracy value for the positive class is not only considered to be the lowest with 0.50, but it is also far away from the next one, which is 0.63. This may be due to the fact that the KNN algorithm is sensitive to class balance and similarities of text data.

We continued in evaluated the sentiment analysis task using BNB, GB, DT, and SVM machine learning algorithms in our experiments. We observed that each algorithm yielded different performance results for different classes as shown in Table3. Algorithms such as RF, SVC, RF, and BNB, all achieved reasonable accuracy rates in the test data set. In particular, the LR models have emerged as an effective option to achieve a high level of success in sentiment analysis. KNN, on the other hand, was the model with the lowest accuracy rate. Some algorithms showed a sensitivity to certain classes compared with the other two classes; SVC showed a sensitivity toward Negative class of 0.38 compared with 0.74 and 0.71 for the positive and neutral classes, respectively and BG showed the same sensitivity with 0.27 for the negative class vs. 0.57 and 0.74 for the positive and neutral classes. However, accuracy alone may not be enough to fully understand the performance of a model. In some cases, factors such as class imbalance must be considered. That's why it's important to evaluate along with other metrics. For example, if the classes are unbalanced and the majority of correctly classified samples belong to the majority class, the accuracy rate can be misleading. In this case, other metrics such as precision, recall, or F1 score must also be taken into account.

Confusion matrix analysis (Table3) for KNN shows that 0.48 of Positive moods and 0.40 of Negative moods were correctly classified. The GB algorithm, on the other hand, shows that 0.57 of Positive moods and 0.27 of Negative moods were correctly classified. As a result, it was found that deep learning-based algorithms such as LR and RF were the most effective options in sentiment analysis in Turkish texts. However, traditional algorithms such as SVC also has a reasonable accuracy rate and may be preferable, especially for fast classification. KNN and GB, on the other hand, are alternatives that can be used in some cases and can be used for a comparison purposes. Our findings make an important contribution to future studies aimed at identifying the most appropriate algorithms for sentiment analysis in Turkish texts and to better understand emotional content. Confusion matrix analysis also helped us to evaluate the emotion classification performance of the algorithms in more detail for each mood.

Another important aspect that might be taken in consideration in understanding the result properly is the characteristics of the data set, and the nature of the algorithms; further work and optimization can be done to improve performance. The results we obtained in the sentiment analysis were quite satisfactory. In the classification task performed on the different machine learning algorithms used in the analysis, it was observed that all models performed at acceptable levels except KNN.

As a result, the accuracy rate alone allows us to make an assessment based on a model, but other factors and metrics must also be considered. However, there are also some challenges encountered in the sentiment analysis process. For example, some data points may show ambiguity due to multiple interpretations of certain emotional expressions. These uncertainties can create difficulty in accurate classification and increase the likelihood of errors in results. For example, situations where positive emotional expressions are more common than negative expressions. This can make it difficult to accurately classify the model's bias due to imbalance and underrepresented classes. In addition, language features such as variability in language, idioms, puns, and irony can also affect correct classification.

It is important to experiment with feature engineering strategies to increase accuracy. In addition, the accuracy rate should be evaluated depending on the requirements of our data set and the purpose of our analysis. When similar literature related to the field of Sentiment Analysis is examined, it is seen that the first study was conducted by Pang et al. [16], and movie reviews in the Internet Movie Database archive were used as a data set. In their study, they created the vector space models required for classification by using feature extraction methods such as unigram, bigram, Part of Speech (POS) on the relevant data set. Tokcaer [2] performed the classification process using machine learning algorithms such as Naive Bayes, Maximum Entropy and SVM on the data set obtained as a result of vector space models. As a result of the findings, the best result in the classification of sentiment analysis was obtained by the SVM machine learning method with an accuracy rate of 82.9% on the unigram data set. In addition, O'Connor et al. [17] applied sentiment analysis to comments on twitter and health-related forums to investigate patients' negative thoughts about the side effects of medications. In practice, the machine learning-based ADRMine method, which uses conditional random fields, was used to extract the concepts in the field of medicine, which were also put forward by them. 6279 and 1784 comments from the health site DailyStrength and Twitter were used as a data set. When the results were examined, it was observed that the ADRMine method gave a higher success rate than SVM and MetaMap methods, which are classifiers used in the field of health, with 82.1%.

As a conclusion, a 62% accuracy rate is not a sufficient metric to evaluate the success of an analysis. Further studies can be done to better understand and improve the analysis results. To improve any model, it's important to experiment with different algorithms, further review the data set, and evaluate other performance metrics. A 62% accuracy rate can be a starting point, but it's important to evaluate other factors to better understand and refine your analysis and model. A 62% accuracy rate is an acceptable result based on the purpose of our analysis and the context.

As a result, our experiments have shown that machine learning models can be used effectively in the field of sentiment analysis. These models provide a valuable tool for classifying text data and understanding emotional content. However, it is important to consider challenges such as ambiguities and language characteristics. In the future, we aim to obtain more precise results with experiments and model improvements with larger and more diverse data sets. Negation is one of the most important concepts that affects the accuracy of the model, in this study “I liked” and “I don't like” are both classified as positive sentiment while after using negation the second sentence “I don't like” should be classified part of the negative class. Another enhancement that is suggested for future work is to use lemmatization instead of stemming in the text preprocessing step, stemming sometimes is harsh and affect the sentiment of the word, so using lemmatization may lead to enhancement in the model performance.

For future work also, studying SA in Turkish data sets from Twitter and compare it with data sets from other domain will be interesting for the reason that writing reviews for hotels, hospitals, restaurants, etc is different than writing tweets. Tweets are normally shorter and classifying them has its own challenge, while reviews are more detailed and contains direct content related to certain good/service and the goal behind writing the review is either to give the opinion or might be used as a complain. Alawi and Bozkurt [18] and Cam et al. [19] focused on data sets from Twitter, while Inan [20] and Alzoubi et al. [21] focused on data sets from reviews. In Alawi and Bozkurt [18], the conventional machine learning model SVM achieved an accuracy of 0.8805 and an F1-Score of 0.8348, and in Cam et al. [19], SVM and Multilayer Perceptron classifier achieved 0.89 and 0.88 accuracy rates. In Inan [20], the logistic regression method was the most successful classification algorithm, with an accuracy rate of 0.92, and in Alzoubi et al. [21] the best achieved accuracy in traditional techniques was 78% accuracy for the Support Vector Machine. Comparing SA classification techniques for data sets collected from Twitter with other data sets resources might give some insights for research in the domain.

References

[1] Tokcaer, S. (2021). Türkçe metinlerde duygu analizi. Yaşar University E-Dergisi, 16(63), 1514-1534.

[2] Kaynar, O., Görmez, Y., Yıldız, M., & Albayrak, A. (2016, September). Makine öğrenmesi yöntemleri ile Duygu Analizi. In International Artificial Intelligence and Data Processing Symposium (IDAP'16) (Vol. 17, No. 18, pp. 17-18).

[3] Danisman, T., & Alpkocak, A. (2008, April). Feeler: Emotion classification of text using vector space model. In AISB 2008 convention communication, interaction and social intelligence (Vol. 1, p. 53). T.

[4] Medler, D. A., Arnoldussen, A., Binder, J.R., & Seidenberg, M.S. (2005). The Wisconsin Perceptual Attribute Ratings Database. http://www.neuro.mcw.edu/ratings/

[5] Alpkoçak, A., Tocoglu, M. A., Çelikten, A., & Aygün, İ. (2019). Türkçe metinlerde duygu analizi için farklı makine öğrenmesi yöntemlerinin karşılaştırılması. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 21(63), 719-725.

[6] Türkmenoğlu, C. (2016). Türkçe metinlerde duygu analizi (Doctoral dissertation, Fen Bilimleri Enstitüsü).

[7] Kozareva, Z., Navarro, B., Vázquez, S., & Montoyo, A. (2007, June). UA-ZBSA: a headline emotion classification through web information. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007) (pp. 334-337).

[8] Mohammad, S. (2012, June). Portable features for classifying emotional text. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 587-591).

[9] Chaffar, S., & Inkpen, D. (2011). Using a heterogeneous dataset for emotion analysis in text. In Advances in Artificial Intelligence: 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, St. John’s, Canada, May 25-27, 2011. Proceedings 24 (pp. 62-67). Springer Berlin Heidelberg.

[10] Boynukalın, Z. (2012). Emotion analysis of Turkish texts by using machine learning methods (Master's thesis, Middle East Technical University).

[11] Kaya, M., Fidan, G., & Toroslu, I. H. (2012, December). Sentiment analysis of Turkish political news. In 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (Vol. 1, pp. 174-180). IEEE.

[12] Oflazer, K. (1994). Two-level description of Turkish morphology. Literary and linguistic computing, 9(2), 137-148.

[13] Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007, January). Analysis of affect expressed through the evolving language of online communication. In Proceedings of the 12th international conference on Intelligent user interfaces (pp. 278-281).

[14] Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012, September). Harnessing twitter" big data" for automatic emotion identification. In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing (pp. 587-592). IEEE.

[15] U. Erogul. Sentiment analysis in Turkish. Master’s thesis, Middle East Technical University, 2009.

[16] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.

[17] O’Connor, K., Pimpalkhute, P., Nikfarjam, A., Ginn, R., Smith, K. L., & Gonzalez, G. (2014). Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. In AMIA annual symposium proceedings (Vol. 2014, p. 924). American Medical Informatics Association.

[18] Alawi, A. B., & Bozkurt, F. (2024). A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data. Decision Analytics Journal, 11, 100473.

[19] Cam, H., Cam, A. V., Demirel, U., & Ahmed, S. (2024). Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers. Heliyon, 10(1).

[20] İnan, H. E. Comparison of Machine Learning Algorithms for Classification of Hotel Reviews: Sentiment Analysis of TripAdvisor Reviews. GSI Journals Serie A: Advancements in Tourism Recreation and Sports Sciences, 7(1), 111-122.

[21] Alzoubi, Y. I., Topcu, A. E., & Erkaya, A. E. (2023). Machine learning-based text classification comparison: Turkish language context. Applied Sciences, 13(16), 9428.