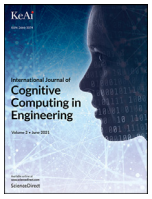
International Journal of Cognitive Computing in Engineering 2 (2021) 202–207

Contents lists available at ScienceDirect

International Journal of Cognitive Computing in Engineering

journal homepage: https://www.keaipublishing.com/en/journals/international-journal-of-cognitive-computing-in-engineering/



Developing Turkish sentiment analysis models using machine learning and e-commerce data

Murat Demircan[a], Adem Seller[b], Fatih Abut[c,*], Mehmet Fatih Akay[c]

[a] Department of Geomatics Engineering, Istanbul Technical University, Istanbul, Turkey
[b] Universal Yazılım A.Ş., Istanbul, Turkey
[c] Department of Computer Engineering, Çukurova University, Adana, Turkey

ARTICLE INFO

Keywords:
Sentiment analysis
Natural language processing
Machine learning
Classifier
Turkish

ABSTRACT

With the increment of Internet usage, there has been a significant increase in the access and interaction of users in social media, blogs, forums, and criticism sites recently. With social media, access to a large amount of data on various products, services, social and political events is provided. Important feedback about products and services can be obtained as a result of analyzing such data. This study aims to determine the sentiments expressed via texts on social media using machine learning methods. As a result of initial research, it is determined that the best case in which texts and emotions match was the product reviews and ratings used on e-commerce websites. Reviews on different products along with review scores from an e-commerce website have been converted into a table to be used in the machine learning-based sentiment analysis models. Reviews have been classified into three groups as positive, negative, and neutral using the review scores. Considering this claim, Turkish sentiment analysis models were developed using support vector machine (SVM), random forest (RF), decision tree (DT), logistic regression (LR), and k-nearest neighbors (KNN). Cross-validation results on independent test data taken from the same e-commerce website show that the SVM-based and RF-based sentiment analysis models outperform the other models. In more detail, there is no strict order between SVM-based and RF-based prediction models, but the results of the SVM-based and RF-based models, in general, are the highest or, in the worst case, similar if we compare them with the scores obtained by using the DT-based, LR-based, and KNN-based models. It can be concluded that SVM and RF are viable methods that can be used to classify product reviews into three groups as positive, negative, and neutral within acceptable error rates.

1. Introduction

Language is used for communication between people to understand each other. If the computer could understand the language of people, it would be an essential tool for communication. Natural Language Processing (NLP) is a subcategory of Linguistics and helps to analyze, understand, or reproduce the structure of natural languages. But the problem of NLP is not having fixed because of uncertainties in the natural languages. Encountered difficulties or problems in NLP studies are usually irregular and distorted texts, long sentences like paragraphs, and expression errors made in texts (Otter, Medina & Kalita, 2021; Torfi, Shirvani, Keneshloo, Tavaf & Fox, 2020).

The explosive growth of the digital industry, such as discussion platforms, e-commerce, product review websites, and social media, facilitates a continuous stream of thoughts and opinions. Growing a continuous stream of thoughts and opinions makes it challenging for companies to better understand customers' aggregate opinions and attitudes towards products. The explosion of Internet-generated content coupled with techniques like sentiment analysis provides opportunities for marketers to gain intelligence on consumers' attitudes towards their products (Rambocas & Pacheco, 2018). Sentiment analysis, also called opinion mining, is an NLP technique to detect positive or negative sentiment in text. Business companies generally use it to detect sentiment in social data, gage brand reputation, and understand the customer and their needs. Recently, as customers express their thoughts and feelings more openly than ever before with social media, sentiment analysis is continually becoming an essential tool to monitor and understand (Chakraborty, Bhattacharyya & Bag, 2020).

Automatically analyzing the customer opinions on products allows brands to learn what makes customers happy or frustrated so that they can improve products and services to meet their customers' needs (Pang & Lee, 2008). Extracting each product review's sentiments helps marketers reach out to customers who need extra care, which will improve customer satisfaction and sales (Vyas & Uma, 2019). There is just too

* Corresponding author.
E-mail addresses: demircanmur@itu.edu.tr (M. Demircan), aseller@uni-yaz.com (A. Seller), fabut@cu.edu.tr (F. Abut), mfakay@cu.edu.tr (M.F. Akay).

https://doi.org/10.1016/j.ijcce.2021.11.003
Received 4 July 2021; Received in revised form 3 October 2021; Accepted 27 November 2021
Available online 29 November 2021

much business data to process manually. Tagging text by sentiment is highly subjective, influenced by personal experiences, thoughts, and beliefs. Using a centralized sentiment analysis system, companies can apply the same criteria to all their data, helping them improve accuracy and gain better insights (Pang & Lee, 2008). It was quite easy to follow the news or media in the past when the Internet was not a trend. However, it is now possible to follow all printed or visual media on the Internet. There are blogs, dictionaries, tweets from Twitter, and posts from every channel. Thousands of posts or tweets are published every second. At this point, sentiment analysis makes it possible to automatically determine people's sentiments and thoughts about specific topics. Sentiment analysis aims to identify the polarity and opinions given in the text, whereas the polarity can be categorized as positive, neutral, or negative.

There are two fundamental sentiment analysis approaches: lexicon-based and supervised learning methods. Many sentiment detection systems use lexicon-based methods, consisting of lists of words and the emotion they convey. One of the downsides of using lexicons to detect emotions/sentiments is that people express emotions in different ways in text. Some words express anger like bad or kill (e.g., your product is so bad or your customer support is killing me) and might also express happiness (e.g., this is a bad ass or you are killing it). The lexicon-based approach has the advantage of being simple. On the other hand, the supervised learning approach is more successful since it learns from samples of texts with known sentiment in the given domain without relying on specially compiled lexicons. The lexicon-based approach gets the polarities of the words or phrases in the text/document from a polarity lexicon to determine the document's semantic orientation. The distinctive aspect of lexicon-based approaches is that they do not involve any domain-specific learning. Supervised learning is a subcategory of machine learning where data is labeled with some measure of interest that we are trying to estimate or classify. Algorithms are trained on labeled datasets through machine learning methods to make classifications or predictions, uncovering key insights within data mining projects (Dhaoui, Webster & Tan, 2017).

This study aims to determine the sentiments expressed via texts on social media using supervised machine learning methods. Reviews on different products, along with review scores from an e-commerce website, have been converted into a table to be used in the machine learning-based sentiment analysis models. Reviews have been classified into three groups as positive, negative, and neutral using the review scores. Turkish sentiment analysis models were developed using support vector machine (SVM), random forest (RF), decision tree (DT), logistic regression (LR), and KNN. The contributions of this study can be summarized as follows:

- We develop new Turkish sentiment analysis models to classify product reviews into three groups as positive, negative, and neutral using various machine learning classifiers.
- By using five machine learning classifiers, including SVM, RF, DT, LR, and KNN, on the utilized dataset, this is one of the most comprehensive studies regarding the number of classifiers used in developing the Turkish sentiment analysis models.
- This study allows ranking of these applied methods in decomposing the sentiments expressed via review texts.

The rest of the paper is organized as follows. Section 2 outlines the related works. Section 3 introduces the dataset and evaluation methodology. Section 4 presents the results and discussion. Finally, Section 5 concludes the paper along with possible future works.

2. Related works

Sentiment analysis research has been active for the last ten years with increasing academic and commercial interest. (Naresh, 2021) used a Twitter API for creating a dataset that is a collection of client tweets of an airline. The created airline dataset included 1200 tweets, which were preprocessed and categorized as 560 positive, 362 negative, and 278 neutral tweets. KNN, SVM, and DT classifiers have been applied to find out the best classification performance. The accuracy scores of the model are 67.0%, 68.0%, and 80.0% for KNN, SVM, and DT, respectively. The precision scores of the model are 70.5%, 69%, and 81.4% for KNN, SVM, and DT, respectively. Finally, the recall scores of the model are 69.3%, 68.1%, and 81.4% for KNN, SVM, and DT, respectively.

(Kemaloğlu, Küçüksille & Özgünsür, 2021) conducted a study using different social media platform data such as Facebook, Twitter, Instagram, and Youtube. Only the Twitter dataset was used for training the Turkish sentiment model. The training dataset was tagged by a specialist and contained 28,189 different tagged comments that include 5712 positive, 11,567 negative, 11,247 neutrals tweets. Various classification models have been developed based on LR, RF, and Long Short-Term Memory (LSTM). The classification success rates were measured with precision, recall, accuracy, and f1-score metrics. The negative, positive, and neutral results have not been separately reported in the study. Instead, the average accuracy scores of the models have been presented. The best result was received with Word Indexing vectorization using the LSTM model with a success rate of 84.46%.

(Shehu, Tokat, Sharif & Uyaver, 2019) conducted Turkish sentiment analysis on the sent tweets. In this study, 13.000 Turkish tweets have been classified as either positive, negative, or neutral. Polarity lexicon and machine learning methods were used to predict the sentiment on Turkish tweets. Pre-processing, tokenization, and stemming stages were applied to tweets. Each word in a tweet has been matched within a dictionary that contains a bag of words in Turkish tweet data. The sentiment polarity of a tweet is estimated by matching the words with the ones in the dictionary. To classify a tweet as either positive, negative, or neutral; each tweet has been analyzed for its polarity using the SVM and Random Forest (RF) algorithms. RF performed better in classifying positive data. SVM performed better than the other algorithms in classifying negative and neutral data in most cases. While SVM achieved the best accuracy with 76.4% and 67.6%, RF achieved the best accuracy of 79.9% and 88.5% on two different raw datasets.

(Rumelli, Akkus, Kart & Isik, 2019) used a dataset containing 272.218 reviews from the hepsiburada.com e-commerce website. The under-sampling method has been used to avoid the underfitting problem. Test and train sets are built by randomly selecting positive and negative samples. 13.000 samples have been taken for each positive and negative from review texts. The ratings of review between 1 and 2 are labeled as negative, three as neutral, and 4–5 as positive. But neutral labeled reviews are discarded when performing the machine learning algorithms. The model results are summarized after 100-fold. Naive Bayesian (NB), RF, SVM, and k-Nearest Neighbor (kNN) algorithms have been applied to the same data to predict sentiments of reviews. The fastest running algorithms are NB and kNN, performed less than 1 min. SVM and RF algorithms needed more time to predict the same data. They achieved an accuracy of 0.73 and f1-score of 0.747 as the best results with the NB classifier. Other algorithms' accuracy, f1-score, and AUC metrics are all also about 0.73. They reported poor prediction results because of difficulties in the preprocessing stages of Turkish sentences and some misleading ratings of user reviews.

(Gezici & Yanıkoğlu, 2018) used Turkish movie review dataset to predict sentiment in text. Turkish movie review dataset has been extracted from the movie rating website called Beyazperde. They considered sentiment analysis as a binary classification problem. 4 or 5-star reviews are considered as positive reviews while 1 or 2-star reviews are considered negative. They excluded reviews with 3-stars from the study. The dataset includes a total of 5.331 positive and 5.330 negative movie reviews. The result of the study shows different metric results given with different approaches such as negative handling, using seed words, and booster words. The basic approach obtains 67.49% accuracy with the Naive Bayes classifier and 67.61% with the SVM classifier. In contrast, the best results are obtained with negation handling and seed words, achieving accuracies of 75.16% with the Naive Bayes and 73.70% with

the SVM classifiers, respectively. This study showed that considering seed or booster words does not improve accuracy. Our research doesn't use seed words or booster words. Using TF-IDF text classification, we managed to classify the relevance of the words. So, the more negative or positive relevant words are combined in the text, the more it is determined whether the text is positive, negative, or neutral.

(Haque, Saber & Shah, 2018) extracted Amazon product data reviews. The extracted data includes three categories from Amazon products: electronics reviews, cell phone and accessories reviews, and musical instrument product reviews, which consist of approximately 48.500 product reviews. As in previous study approaches, the study also considers 4 and 5-star ratings as positive, discards 3-star ratings because of neutral meaning, and considers other star ratings as negative. Linear SVM, Multinomial Naïve Bayes (NB), Stochastic Gradient Descent (SGD), RF, LR, and Decision tree (DT) classifiers have been used to predict the sentiments in the Amazon reviews dataset. They were able to achieve accuracy with the f1-measure, precision, and recall over 90%. Linear SVM provided the best classifying results.

In this study, we tried to make our work more efficient by choosing and combining the best ideas from related works. Our system used a large dataset from the hepsiburada.com e-commerce website to give efficient results and make better decisions. We approached the preprocessing step differently from the other studies on Turkish sentiment analysis and considered the usage of negative words in Turkish sentences. Previous studies show that while training the dataset reviews, whole text data is included in their preprocessing approach. In our study, however, only sentences with less than 1000 characters were included in the training because long sentences could make losing the sentiment written in text. To allow the computer to understand and classify any text, words should be broken down. In our research, while we are analyzing the reviews, it is concluded that people usually comment their emotions at the beginning of their sentences. So, training the dataset based on this approach makes our model more accurate.

3. Methodology

3.1. Dataset

The dataset containing reviews and review scores of products from the hepsiburada.com e-commerce site has been used in this study. Hepsiburada is one of the largest e-commerce sites in Turkey. When extracting the data from the hepsiburada.com website, various kinds of product contents are reviewed. It is assured that reviews are combined from many different categories of product reviews. To manage the review's emotions, we also extracted scores of the reviews. Reviews are the features, and review scores are the labels of our dataset. The utilized dataset includes more than 250.000 reviews of several different products. 4 and 5-star ratings are labeled as positive, 3-star ratings are considered neutral, and other (i.e., 1 and 2-star) ratings are considered negative. The dataset has been simplified to overcome the imbalanced data problem. Since the number of positively labeled reviews is higher than the negatively labeled ones, the positive prediction rate is always higher when such a dataset is given to an ML model. So, the under-sampling method has been used to avoid over-fitting and under-sampling problems. The dataset includes 19.392 positive, 7.517 negative, 4.848 neutral reviews after applying the preprocessing step.

3.2. Word and text classification techniques

Bag of Words (BoW) approach is one of the most used text representation methods. There are two classification methods used for text classification: Count Vectorizer and Term Frequency (TF). Count Vectorizer class is used for word-counting purposes. TF is used for text summarization and classification that calculates the number of times the word appears in the document. Inverse Document Frequency (IDF) method understands whether the word is a term (i.e., stop words) or a repeated word by looking at the number of occurrences of the selected word in more than one document (Haque et al., 2018).

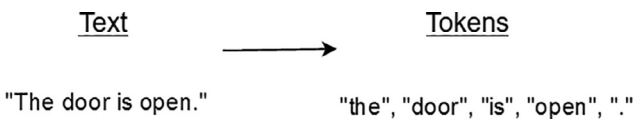


Fig. 1. Word Tokenization.

TF-IDF is another topic of text classification. TF-IDF is calculated with the weight factor of words by using the statistical method. It measures the relevance on the contrary of Count Vectorization. The more the word appears in the document, the less valuable that word becomes. For instance, some frequently used words in the sentences such as "and", "or", "the" can be discounted to analyze the meaning of a sentence. The less the term is repeated, the greater the IDF value we obtain. Eq. (1) shows the TF-IDF equation, where $tf_{i,j}$ is the number of occurrences of i in j, df_i is the number of documents containing i, and N is the total number of documents.

$$w_{i,j} = tf_{i,j} \; x \; \log\left(\frac{N}{df_i}\right) \tag{1}$$

In the preprocessing step, reviews were examined whether to add as feature data. Sentences with less than 1000 characters were included in train data as long articles could make losing the sentiment written in text. It has been observed that negative comments include sharp terms such as "sevmedim (I didn't like)", "iğrenç (disgusting)", "nefret (hate)" and positive comments include "bayıldım (loved it)", "güzel (nice, beautiful)", "muhteşem (wonderful, amazing)".

To get the computer to understand any text, it is needed to break that word down so that our machine can understand. That is the subject of tokenization in NLP. Tokenization is used to separate a piece of text into smaller units called tokens. Tokenization can be broadly classified into three types such as word, character, and sub-word tokenization. Tokenization is the most important step while modeling text data. Tokenization is performed to obtain tokens, as illustrated in Fig. 1. Then, the tokens are used to prepare a vocabulary. Traditional NLP approaches such as Count Vectorizer and TF-IDF use vocabulary as features. Each word in the vocabulary is treated as a unique feature.

Term Document Matrix is tracking the term frequency for each term by each document. We split the training dataset into a train and validation dataset to evaluate the result and apply cross-validation.

3.3. Methods for classification of text data

Five machine learning classifiers, including SVM, RF, DT, LR, and KNN, have been applied on the utilized dataset to group reviews into positive, negative, and neutral reviews.

The SVM (Qi, Silvestrov & Nazir, 2017) has been used for a wide range of applications to solve classification problems. It is a classifier that separates the data with hyper-plane. In 2-dimensional spaces, the hyperplane is a straight line that maximizes the margin between the two classes. The key idea of SVM is to find line separators in the search space for separating various groups. SVM method's mathematical formulation uses the values of cost (C), epsilon (ε), gamma (γ), and the type of kernel function. We conducted a grid search to optimize the performance of SVM. As a result, we acquired the best values of parameters "C" as 4, "ε" as 0.001, "γ" as scale, and the linear kernel function.

The RF (Biau, 2012) belongs to the class of ensemble algorithms. The forest part indicates the collection of decision trees, whereas the random part indicates taking a subset of the input data and create bootstrapped dataset. The bootstrapped dataset contains only 2/3 of the input data. Some of these data is duplicated, and some are not included in the training. The excluded dataset can be used for the testing process. RF classifier uses more than one criteria to optimize the performance of a model. Hyperparameters of the RF classifier include the number of trees

in the forest (n_estimators), the maximum number of features considered for splitting a node (max_features), the maximum number of levels in each decision tree (max_depth), the minimum number of data points placed in a node before the node is split (min_samples_split), the minimum number of data points allowed in a leaf node (min_samples_leaf), and method for sampling data points with or without replacement (bootstrap). We used the RandomizedSearchCV method from the sklearn library. After seeding the range of hyperparameters to the RandomizedSearchCV method, we set n_iter=100 to try 100 different combinations and fold number to 3 for cross-validation. In this way, settings will cover wider search space and will reduce the chances of overfitting. As a result, we acquired the best values of "n_estimators" as "200″, "max_features" as "sqrt", "max_depth" as none, "min_samples_split" as 2, "min_samples_leaf" as "2″, and "bootstrap" as false.

As the name implies, the DT (Patel & Prajapati, 2018) uses a tree-like model of decisions. DT implementation can be performed without scaling the data, and it could implicitly perform variable screening or feature selection. However, DT has the drawback of suffering from overfitting problems that can lead to low accuracy in classification problems. The DT classifier uses two criteria to decide how to split the data with the Gini index and Information Gain.

The LR (Widodo & Handoyo, 2017) belongs to the group of linear classifiers and is somewhat similar to polynomial and linear regression. LR is fast and relatively uncomplicated, and it is convenient to interpret the results. Although it is essentially a method for binary classification, it can also be applied to multiclass problems. LR uses different optimization functions such as lbfgs, liblinear, and different parameters such as "C", "penalty", "dual". We acquired the best values of "C" as 1.0, "penalty" as l2, and "dual" as false.

Finally, the KNN (Zhang, 2016) solves both classification and regression problems. It indicates that similar values exist in close proximity, calculating the similarities between the input sample datasets. Getting the best result of the KNN classifier requires choosing the optimum number of neighbors in predictions. The Elbow method is used to find the best KNN model by trying different neighbor parameters ranging from 1 to 30 to determine the best KNN model.

The performance of the models has been evaluated using three different metrics, including precision, recall, and f1-score. The equations of precision, recall, and f1-score are given in (2), (3), and (4), respectively.

$$precision = \frac{tp}{tp + fp} \tag{2}$$

$$recall = \frac{tp}{tp + fn} \tag{3}$$

$$f1 - score = 2 \cdot \frac{1}{\frac{1}{recall} + \frac{1}{precision}} = 2 \cdot \frac{precision * recall}{precision + recall} \tag{4}$$

In Eqs. 1 through 3, tp, fp, and fn are the number of true positive, false positive, and false negative estimates. Higher precision means fewer false positives, while a lower precision means more false positives. Recall calculates the sensitivity of a classifier, i.e., how much positive data it returns. Higher recall means fewer false negatives. Recall is the ratio of the number of instances accurately classified to the total number of predicted instances. Combining precision and recall produces a single metric known as the f1-score measure, which is the weighted harmonic mean of precision and recall (Haque et al., 2018).

4. Results and discussion

Table 1 through Table 5 show precision, recall, and f1-scores of SVM-based, RF-based, DT-based, LR-based, and KNN-based models.

There is no strict order between SVM-based and RF-based prediction models in terms of precision and recall values. The SVM-based model's f1-scores, however, are consistently higher than the ones of the RF-based model. On the other hand, the precision, recall, and f1-scores of the SVM-based and RF-based models, in general, are the highest or, in the worst case, similar if we compare them with the scores obtained by using the DT-based, LR-based, and KNN-based models. The only exception is that the SVM-based model's precision value for neutral text is lower than the precision score obtained by using the LR-based model. Similar observation also applies to the precision and recall scores of the RF-based model. In most cases, the precision and recall values of the RF-based model are higher than those obtained by using DT-based, LR-based, and KNN-based models. However, there are some exceptions. In more detail, the RF-based model's precision value for positive text is lower than the precision scores obtained by using the DT-based and KNN-based models. Similarly, the RF-based model's recall value for negative text is lower than the recall scores obtained by using the DT-based, LR-based, and KNN-based models. Furthermore, the RF-based model's recall values for neutral and negative texts are also lower than the corresponding recall scores obtained by using the DT-based model.

Table 1
Precision, recall, f1-score of tuned SVM-based model.

	Precision	Recall	f1-score
Negative	0.87	0.87	0.87
Neutral	0.81	0.72	0.77
Positive	0.93	0.95	0.94

Table 2
Precision, recall, f1-score of tuned RF-based model.

	Precision	Recall	f1-score
Negative	0.90	0.73	0.80
Neutral	1.00	0.48	0.64
Positive	0.82	0.99	0.90

Table 3
Precision, recall, and f1-score of DT-based model.

	Precision	Recall	f1-score
Negative	0.81	0.79	0.80
Neutral	0.77	0.73	0.75
Positive	0.91	0.93	0.92

Table 4
Precision, recall, and f1-score of LR-based model.

	Precision	Recall	f1-score
Negative	0.84	0.79	0.82
Neutral	0.87	0.38	0.53
Positive	0.84	0.97	0.90

Table 5
Precision, recall, and f1-score of KNN-based model.

	Precision	Recall	f1-score
Negative	0.68	0.83	0.75
Neutral	0.65	0.65	0.65
Positive	0.91	0.83	0.87

The performance of the Turkish sentiment analysis model was also manually tested in sentences and verbs as well as in words. The test results are illustrated in Table 6 and Table 7. All manual test results were correct as expected.

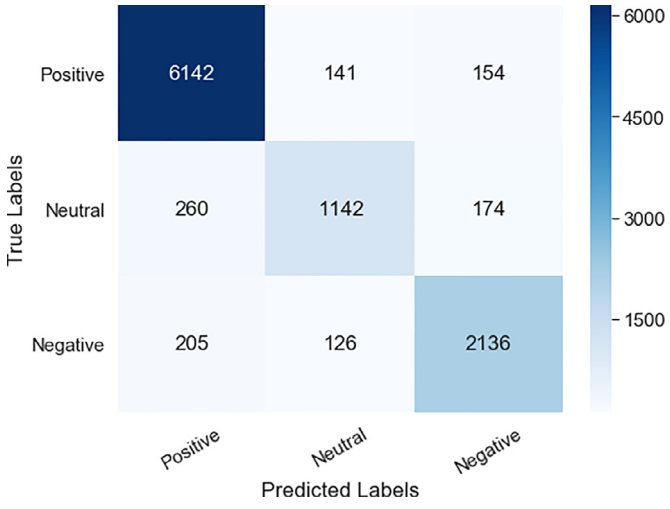


Fig. 2. Confusion matrix results of the tuned SVM-based model.

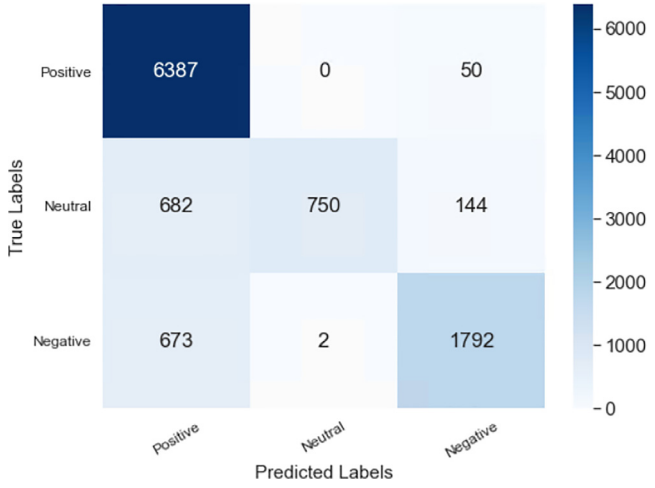


Fig. 3. Confusion matrix results of the tuned RF-based model.

Table 6
Results of Turkish sentiment analysis model used in sentences and verbs.

Sentence/Reviews	Sentiment
bu nasıl bir şey ya (What kind of thing is that? Seriously!)	Negative
şuna bakar mısın (Could you look at that?)	Negative
beğenmedim (I didn't like it)	Negative
hoşlanmadım (I disliked it)	Negative
sevdim (I liked it)	Positive
iğrenç ama yine de sevdim (disgusting but I liked it anyway)	Positive
idare eder, ne iyi ne kötü (it's okay, neither good or bad)	Neutral

Table 7
Results of Turkish sentiment analysis model used in words.

Words	Sentiment
nefret (hate)	Negative
iyi bir şey (a good thing)	Positive
sıkıcı (boring)	Negative
muhteşem (gorgeous)	Positive
çekici (attractive)	Positive
şaka (joke)	Positive
efsane (legend)	Positive
çok güzel (very beautiful)	Positive
çok güzel değil (not so good)	Negative
boş (empty)	Negative
iğrenç (disgusting)	Negative

Using cross-validation with SVM method to tune hyper-parameters takes ca. 2 h and 36 min. The training time of the RF-based model lasts ca. 2 h 46 min. In contrast, training the DT-based, LR-based, and KNN-based models requires much shorter times ranging from 1 to 9 s, respectively.

The confusion matrix results of the SVM-based model are given in Fig. 2. It is observed that overall, out of 6437 positively labeled reviews, the classifier correctly predicted 6142 reviews as positive, whereas 260 and 205 reviews were misclassified as neutral and negative, respectively. Out of 1576 actual neutrally labeled reviews, the classifier correctly predicted 1142 reviews as neutral, whereas 260 and 174 reviews were misclassified as positive and negative, respectively. Out of 2467 negatively labeled reviews, the classifier correctly predicted 2136 reviews as negative, whereas 205 and 126 reviews were misclassified as positive and neutral, respectively.

Similarly, the confusion matrix results of the RF-based model are given in Fig. 3. It is observed that overall, out of 6437 positively labeled reviews, the classifier correctly predicted 6387 reviews as positive, whereas only 50 reviews were misclassified as negative. Out of 1576 neutrally labeled reviews, the classifier correctly predicted 750 reviews as neutral, whereas 682 and 144 reviews were misclassified as positive and negative, respectively. Out of 2467 negatively labeled reviews, the classifier correctly predicted 1792 reviews as negative, whereas 673 and 2 reviews were misclassified as positive and neutral, respectively.

As previously mentioned in Section 2, (Rumelli et al., 2019) collected 221.071 positively and 13.012 negatively labeled product reviews from the hepsiburada.com e-commerce website as in our study. The neutrally labeled data was discarded in the Rumelli study. The under-sampling method was used to avoid the underfitting problem. In total, 13.000 labeled samples have been taken from each positive and negative reviews. The proposed best-performed KNN-based model produced an f1-score of appr. 0.73 after 100-fold cross-validation. In our study, we trained negative, positive, and neutral reviews separately, also using KNN. The f1-scores of our KNN-based model are meassured as 0.75, 0.65, and 0.87 for negative, neutral, and positive reviews, respectively, which are, on average, slightly more accurate than the ones reported in (Rumelli et al., 2019). Building our model with SVM and RBF increases the classification performance even more. In these cases, the f1-scores of our SVM-based model are calculated as 0.87, 0.77, and 0.94 for negative, neutral, and positive reviews, respectively. Finally, the f1-scores of our RF-based model are calculated as 0.80, 0.64, and 0.90 for negative, neutral, and positive reviews, respectively.

5. Conclusion and future work

Sentiment analysis helps businesses to process vast amounts of data efficiently and cost-effectively. Sentiment analysis is critical because it allows firms to quickly understand the overall opinions of their customers. By automatically sorting the sentiment behind reviews, social media conversations, and more, one can make faster and more accurate decisions. Sentiment analysis empowers all kinds of markets and competitive analysis. Whether exploring a new market or anticipating future trends, sentiment analysis can make all the difference. Sentiment analysis helps governments and companies make better-informed decisions while saving time and money.

In this study, models for predicting the sentiments expressed via texts on social media have been proposed using supervised machine learning methods. Particularly, SVM, RF, DT, LR, and KNN classifiers have been used to classify the reviews as either positive, negative, or neutral. The sentiment analysis prediction results made on the test review data were observed. The results show that there is no strict order between SVM-based and RF-based sentiment analysis models prediction models, but their results, in general, outperform the other models built based on DT, LR, and KNN. It can be concluded that SVM and RF are viable methods that can be used to classify product reviews into three groups as positive, negative, and neutral within acceptable error rates.

As future works, after acquiring more data to our training dataset, we plan to split the dataset into a fixed number of k-folds to improve the performance of the sentiment analysis model. Transferring labeled specific words with polarity would increase into the training dataset, leading to better accuracy results. New words would be investigated to insert into the training dataset. The future investigation would involve every Turkish abbreviation word to the dataset to improve accuracy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research, 13*, 1063–1095.

Chakraborty, K., Bhattacharyya, S., & Bag, R. (2020). "A survey of sentiment analysis from social media data". *In Proc.of IEEE Transactions on Computational Social Systems, 7*(2), 450–464. 10.1109/TCSS.2019.2956957.

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: Lexicon versus machine learning. *Journal of Consumer Marketing, 34*(6), 480–488. 10.1108/JCM-03-2017-2141.

Gezici, G., & Yanıkoğlu, B. (2018). *Sentiment analysis in turkish*. Cham: Springer Turkish natural language processingpp. 255-271. 10.1007/978-3-319-90165-7_12.

Haque, T.U., .Saber, N.N., .& Shah, F.M. (.2018). Sentiment analysis on large scale Amazon product reviews. In Proc. of IEEE International Conference on Innovative Research and Development, 1-6. https://doi.org/10.1109/ICIRD.2018.8376299

Kemaloğlu, N., Küçüksille, E. U., & Özgünsür, M. E. (2021). Turkish sentiment analysis on social media. *Sakarya University Journal of Science, 25*(3), 629–638. 10.16984/-saufenbilder.872227.

Naresh, A. (2021). Recommender system for sentiment analysis using machine learning models. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(10), 583–588.

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). "A survey of the usages of deep learning for natural language processing". *In Proc. of IEEE Transactions on Neural Networks and Learning Systems, 32*(2), 604–624. 10.1109/TNNLS.2020.2979670.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*, 1–135 1-2. 10.1561/1500000011.

Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering, 6*(10), 74–78.

Qi, X., Silvestrov, S., & Nazir, T. (2017). Data classification with support vector machine and generalized support vector machine. In *Proc. of AIP Conference*, 2020, 1798 (1), 020126. AIP Publishing LLC.

Rambocas, M., & Pacheco, B. G. (2018). Online sentiment analysis in marketing research: A review. *Journal of Research in Interactive Marketing, 12*(2), 146–163. 10.1108/JRIM-05-2017-0030.

Rumelli, M., Akkus, D., Kart, O., & Isik, Z. (2019). Sentiment analysis in turkish text with machine learning algorithms. In *Proc. of Innovations in Intelligent Systems and Applications Conference*. 10.1109/ASYU48272.2019.8946436.

Shehu, H. A., Tokat, S., Sharif, M. H., & Uyaver, S. (2019). Sentiment analysis of Turkish Twitter data. *Proc. of AIP Conference, 2183*(1), Article 080004. 10.1063/1.5136197.

Torfi, A., Shirvani, R.A., .Keneshloo, Y., Tavaf, N., & Fox, E.A. (.2020). Natural language processing advancements by deep learning: A survey. arXiv preprint arXiv:2003.01200.

Vyas, V., & Uma, V. (2019). *Approaches to sentiment analysis on product reviews. sentiment analysis and knowledge discovery in contemporary business* (pp. 15–30). IGI Global.

Widodo, A., & Handoyo, S. (2017). The classification performance using logistic regression and support vector machine (SVM). *Journal of Theoretical & Applied Information Technology, 95*(19), 5184–5193.

Zhang, Z. (2016). Introduction to machine learning: K-nearest neighbors. *Annals of translational medicine, 4*(11), 218–225.