Journal of
Computer & Electrical and Electronics Engineering Sciences
Original Article
DOI: 10.51271/JCEEES-0016

Sentiment analysis with machine learning for drug reviews

Muhammed Oğuzhan Bozkurt[1], Yağız Yaman[2], Fahrettin Horasan[2]

[1]*Department of Software Engineering, Faculty of Engineering, İstanbul Gedik University, İstanbul, Turkiye*
[2]*Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Kırıkkale University, Kırıkkale, Turkiye*

Cite this article: Bozkurt, M.O., Yaman, Y., & Horasan, F. (2024) Sentiment analysis with machine learning for drug reviews. *J Comp Electr Electron Eng Sci*, 2(2), 35-45.

Corresponding Author: Fahrettin Horasan, fhorasan@kku.edu.tr

Received: 29/08/2024 ◆ **Accepted**: 22/09/2024 ◆ **Published**: 29/10/2024

ABSTRACT

In the treatment of the diseases, the fact that individuals use drugs independently from doctors without appropriate consultation causes their health status to become worse than normal. This article aims to conduct a sentiment analysis over the comments of individuals about the drug in case they use drugs without consultation. Within the scope of this study, patients' comments about drugs were vectorized using Bow and TF-IDF algorithms, sentiment analysis was made, and the predicted sentiments were; it was evaluated with precision, recall, f1score, accuracy and AUC score. As a result of the evaluations, the most successful result was obtained in the TF-IDF method. This result is the result of the linear support vector classifier algorithm with an accuracy value of 93%.

Keywords: Drug, sentiment analysis, machine learning, natural language processing, smote

INTRODUCTION

One of the worldwide problems with the pandemic is the insufficient number of specialist doctors. This situation sometimes delays patients' access to appropriate diagnosis and treatment. It takes between 6 and 12 years for an average doctor to acquire the necessary qualifications, so the number of qualified doctors cannot be increased rapidly in a short time (Garg, 2021). It should be considered that telemedicine applications can accelerate patients' access to appropriate treatment in global health crises such as pandemics (Dinakaran et al., 2020).

In addition, clinical errors are quite common nowadays. According to the findings obtained from the studies, 200 thousand people in China and more than 100 thousand people in the USA are exposed to wrong treatment due to misdiagnosis and prescriptions (Garg, 2021). On the other hand, in the medical world, experts make more than 40% errors while writing prescriptions, mostly due to the limited knowledge of specialists in the treatment of the disease (Wittich et al., 2014). In addition, new drugs emerging every day led to the emergence of tests and new studies for clinical staff. This situation makes it increasingly difficult for doctors to determine the appropriate treatment and drug for the patient, depending on the indication and clinical history.

The rapid development of the web-based business industry has made product reviews an indispensable factor in purchasing. Individuals browse product reviews and alternative websites when deciding on the product to buy. Much of the past research has focused on ratings and recommendations on e-commerce sites and has rarely been studied in the field of medical care or clinical treatments. Recently, the use of online diagnostic websites has increased considerably. According to a survey conducted by the Pew American Research Center in 2013, approximately 60% of individuals have searched online for health-related topics this year, and approximately 35% of these individuals have received help from the web in diagnosing their disease (Fox et al., 2013). Considering these situations, it is vital to develop a drug recommendation system to assist experts and ensure that individuals reach the appropriate medicine for their disease.

Recommendation systems are systems designed to facilitate and accelerate users' access to the products they need (Jalili et al., 2018). In the drug recommendation system, drug recommendations are made to users using sentiment analysis and feature engineering. While sentiment analysis is the extraction of some emotional data such as opinions and attitudes from the relevant text (Kaur et al., 2017); feature engineering is optimizing existing data and extracting more features (Oyamada, 2019). Considering all these evaluations, it should not be overlooked that emotion analysis should be performed in the most accurate way before the recommendation system to be developed for patients to reach the appropriate treatment. This study aims to carry out the sentiment analysis to be performed with the highest possible accuracy.

In the continuation of the study, there is a literature review section in which the researches related to the subject are examined, the methodology section in which the methods applied within the scope of the study are explained and analyzed, the results section where the results of the applied methods are examined, and the discussion and solution sections in which the study is examined.

In the continuation of the study, there is a literature review section where the researches related to the subject are examined, the methodology section where the methods applied within the scope of the study are explained and examined, the results section where the results of the applied methods are examined, and the discussion and solution sections where the study is examined.

Literature Review

With the increase in the development of artificial intelligence, machine learning and deep learning methods have started to be applied in recommendation systems. Recommendation systems today it is frequently used in the travel industry, e-commerce sites, restaurant applications and TV series-film sites. However, in the field of drug recommendation medical expressions such as drug reviews, disease names, reactions, and synthetic names are complex and difficult to understand, leading to limited studies involving sentiment analysis in this area (Tekade et al., 2016).

In this article, in which a recommendation system that applies sentiment analysis technologies in drug reviews was created, a decision support platform was designed to assist patients in their drug selection. First, a rating of drugs was created by performing emotion analysis with drug reviews. Second, how useful drug reviews are for users, the patient's conditions, and the glossary sensitivity polarity of drug reviews are taken into account. These factors were then included in the recommendation system to list suitable drugs. Finally, hyper parameter optimization is performed for each algorithm to achieve higher performance (Hossain et al., 2020).

In this study, in which a recommendation system was developed using product images, Amazon Apparel database with clothing data was used. NLP technologies and CNN are used to predict similar products. The title of the product was chosen as the main attribute for NLP analysis and product recommendation. CNN is used to generate a feature vector from product images, and all other vectors are used to estimate this vector. By calculating the distances between the vectors of all products, the products with the least distance are recommended. VGG-16 architecture is used to extract features from images (Sharma et al., 2021).

In another study, it was assumed that the recommended drug should be determined according to the patient's immunity. For example, if the patient's immunity is low, reliable drugs should be recommended at this point. The main purpose of the study is to protect a patient from infection by making use of the patient's clinical information. Risky situations such as the patient's weaknesses and allergies are evaluated and scored according to the effect of these conditions. The system calculates the risk level after comparing the risk factors, allowing doctors to easily make inferences about the patient's condition. As a result of the study, a web-based prototype system was also created that uses a decision support system that helps doctors choose primary care drugs. However, as in the previous study, it should be noted that this study did not design a recommendation system based on sentiment analysis and machine learning (Shimada et al., 2005).

In this article, in which vaccine hesitancy is examined and a solution to this problem is tried to be produced by sentiment analysis, articles related to different vaccine hesitations made in the last 11 years have been examined and analyzed. It is intended to develop the application of sentiment analysis on the most important literature findings (Alamoodi et al., 2021).

One of these studies presents GalenOWL, an online framework based on semantically augmented semantic web technologies to help professionals discover details about drugs. GalenOWL describes a framework that recommends medication for a patient based on the patient's infection, sensitivities, and drug interactions. However, this work requires advanced medical knowledge and many complex and memory-intensive queries. The recommendations generated in the system are based on complex queries that do not rely on sentiment analysis and machine learning, but mostly consider relationships between data (Doulaverakis et al., 2012).

In this study, in which the movie recommendation system developed using sentiment analysis is presented, cosine similarity is used for the developed recommendation system. The main purpose of the study is to present the content they need to end users through semi-structured data on the internet. As a result of the study, it was concluded that cosine similarity provides better and more efficient results for a recommendation system (Khatter et al., 2021).

In another study focusing on movie recommendations, sentiment analysis was performed using the LDA method, using user comments. The sentiment analysis has been used to develop a content-based recommendation system. In this study, BERT technique was used to train emotion classification models and LDA technique was used for modeling subjects (Zhang et al., 2022).

Zhang et al. (2014) has developed a cloud-assisted drug recommendation system (CADRE). According to patients' side effects, CADRE recommends the most relevant prescription drugs. This proposed framework was initially built on collaborative filtering techniques as indicated by the functional identification data of drugs. The model is then transformed into a cloud-powered approach that uses tensor decomposition to improve the quality of the drug recommendation experience.

In this study, in which a tourism recommendation system was developed, sentiment analysis of user comments on TripAdvisor was used to identify touristic places that might be of interest to tourists. After the comments are processed, they are clustered semantically and sentiment analysis is performed. These comments are then used to extract the attractions of the attractions. Finally, the user is recommended the most suitable tourist attractions. In addition, the system uses some contextual information such as time, location and weather to filter items and improve the quality of recommendations (Abbasi-Moud et al., 2021).

In another study, a universal drug recommendation system framework was designed and implemented, applying data mining technologies to the recommendation system. Drug recommendation system consists of database system module, data preparation module, recommendation model module, model evaluation and data visualization module. Based on the diagnosis data, support vector machine (SVM), BP neural network algorithm and ID3 decision tree algorithm were used for drug recommendation system. Although data mining methods were used in the study, sentiment analysis based on user comments was not performed (Bao et al., 2016).

In this study, in which a different approach to recommendation systems is presented, a hierarchical approach is presented to improve the performance of e-commerce recommendation systems. This approach, called DeepIDRS, has a two-level hierarchical structure. At the first level, a bidirectional encoder is used to represent the textual information of an item efficiently and accurately. At the second level, an attention-based recommendation system is used, which uses the item representations produced at the first level. In addition, the developed DeepIDRS approach has been compared with other existing approaches. As a result, it has been observed that this approach improves the performance of the recommendation system (Islek et al., 2022).

In this article, which examines the basic approaches used for sentiment analysis, important concepts related to sentiment analysis, the usage areas of sentiment analysis and the preprocessing process required for the realization of sentiment analysis are examined. In addition, models that can be used for machine learning are also examined (Kaur et al., 2017).

Pereira et al. (2020) presented a study aiming to compare these techniques by analyzing the filtering techniques commonly used in recommendation systems. In addition, criteria such as weighting, frequency, distance and similarity used in recommendation systems were also evaluated.

In another study, the reflections of people's professions on social media platforms were analyzed. Based on John Holland's theory of occupational choice, the study aims to classify occupation-related content on Twitter data of four different occupational groups. Long short-term memory (LSTM) and gated recurrent unit (GRU) models were used for classification. While the LSTM model yielded an accuracy score of 93.025%, the GRU model achieved 94.025% accuracy. These results suggest that social media posts may be consistent with individuals' occupations and that their occupations are reflected in their behavior on these platforms (Dağıstanlı et al., 2023).

METHODOLOGY

In this study, the drug review dataset (Drugs.com) available in the UCI ML repository was used for sentiment analysis. The fields in this data set; name of the drug (text), condition of the drug (text), review of the patient using the drug (text), given to the drug by the patient of rating score (numerical), date of the review (date), number of users who found the review helpful (numeric) consists of features. There are a total of 215063 samples in the data set.

Within the scope of the study, a three-stage study method was adopted for sentiment analysis based on user comments. These stages are respectively; data preparation, classification, and evaluation stages.

Data Preparation

In the data preparation phase; data preprocessing methods such as checking for empty samples, checking for meaningless samples, cleaning data, extracting features from rows and categorizing data were used.

Before checking the empty samples, meaningless samples were detected and the "condition" areas of these samples were cleared. At this stage, it was determined that there were 1171 samples containing meaningless expressions. As a result of checking the empty samples, it was determined that there were a total of 2365 empty values in the "condition" field, which represents the usage status of the drug. Instead of directly deleting the rows with these empty values, care was taken to recover as much data as possible. In this way, it is envisaged to increase the performance of the sentiment analysis to be performed.

First of all, the records of 17 drugs that do not have a use case in the data set, that is, they cannot be recovered in any way, were removed from the data set. Then, 305 empty fields of the drugs with only 1 usage status were filled according to the current usage status. Finally, the use cases of the drugs were searched in the user reviews, and the empty field was tried to be filled by using the matching "condition" information in case a match was achieved. At this stage, a meticulous effort was made to fill in the blank data correctly. As a result of this process, 725 empty samples were filled. As a result, a total of 1030 data were recovered and 1318 rows containing empty samples were removed from the data set. As a result of checking the empty and meaningless samples, the number of new lines was determined as 213728. In order to improve the performance of text mining, it is extremely important to recover as much data as possible, since the "condition" field will be included in the vector to be created. In addition, the accuracy of text mining also depends on the relevance of the properties of the generated matrix to each other. For this reason, irrelevant data needs to be cleaned.

During the data cleaning phase in order to delete punctuation marks in the texts, deletion of numerical expressions, deletion of ineffective words and to obtain the roots of the words, lemmatization and stemming processes were performed. It has been observed that the vector created by lemmatization produces better results. Therefore, the study will continue with lemmatization.

Feature Extraction

After performing the data cleaning phase, feature extraction will be done using the "review" column, which contains user evaluations for sentiment analysis. The purpose of feature extraction here is to enable machine learning algorithms to work with text expressions. Each of the user ratings will be expressed as a numerical vector. This vector conversion will be done using bag of words and TF-IDF methods.

Bag of words is an algorithm used in natural language processing responsible for counting the number of all tokens in a review or document. A term or symbol can be called a word (unigram) or number of words, n-grams. A major disadvantage of the Bow model is that it creates a large matrix that is computationally costly to train (Pu et al., 2007). In this study (1.2) n-gram range was chosen.

TF-IDF is an algorithm used for statistical analyzes on text expressions in text mining and natural tooth processing. TF-IDF calculates the probability of a word being found in any text by weighting the words instead of counting them. Term frequency (TF) is the probability of the word being found in a text. Inverse document frequency (IDF) is calculated by taking the logarithm of the TF value and is inversely proportional to the TF. TF-IDF is obtained by multiplying TF and IDF values (Qaiser et al., 2018). In this study, the n-gram interval selected for TF-IDF was chosen as (1.2), just like in the bag of words algorithm.

Encoding of Data

As it is known, since many machine learning algorithms cannot learn and prediction on categorical data, these data must be encoding. Encoding of data can be done in different ways. Within the scope of this study, one-hot encoding method was preferred for encoding process. One-hot encoding provides binary representation of categorical data (Rodríguez et al., 2018).

As we mentioned before, the "condition" and "drugName" fields in the data set consist of text expressions. It is assumed that the inclusion of these fields in the machine learning models together with the "review" field will give more accurate results. After these fields were encoding with the one-hot encoding method, they were combined with the data set from which feature extraction was made through the "review" field and made ready for machine learning.

Train-Test Split

The data, created using bag of words, TF-IDF and one-hot encoding methods, is splited into 75% training and 25% testing. The same random state value was used to obtain the same set of random numbers when splited the data sets.

Synthetic Minority Oversampling Technique (SMOTE)

After the training-test splitting of the data, SMOTE was used to minimize the class imbalance problem. SMOTE is an oversampling technique that synthesizes existing data to balance the data. First, the difference between the examined feature vector and its nearest neighbor is taken. The result obtained is multiplied by a random number between 0 and 1. Finally, the result obtained from this multiplication is added to the examined feature vector and a synthetic new sample is formed (Feng et al., 2021).

In this study, in addition to the standard SMOTE method described above, ADASYN, SMOTE TOMEK and borderline SMOTE methods were also used. Unlike SMOTE, ADASYN provides more realistic data by adding random values to synthetic samples (He et al., 2008). In the SMOTE TOMEK method, TOMEK links are added to the synthetic data created with SMOTE (Wang et al., 2019). In the borderline SMOTE method, the samples located on the border lines of the clusters formed by the samples belonging to the minority class type are used to create synthetic samples (Chen et al., 2021). SMOTE methods are applied only to training data. The amounts of synthetic data generated by the SMOTE methods used in the Table 1 below can be examined.

Table 1. SMOTE methods

SMOTE	Class	Train (75%)	Test (25%)
No SMOTE	Negative	47991	15852
	Positive	112305	37580
	Total	160296	53432
Standart SMOTE	Negative	78613	15852
	Positive	112305	37580
	Total	190918	53432
ADASYN	Negative	78146	15852
	Positive	112305	37580
	Total	190451	53432
SMOTE TOMEK	Negative	78078	15852
	Positive	111770	37580
	Total	189848	53432
Borderline SMOTE	Negative	78613	15852
	Positive	112305	37580
	Total	190918	53432

SMOTE: Synthetic minority oversampling technique

As can be seen in the Table 1, the amount of synthetic samples produced in most methods is very close to each other. In the study, the standard SMOTE method, which is the fastest method, was preferred because both time and performance costs were considered and the methods yielded very close results.

Feature Scaling

Scale differences between features can prevent machine learning from being done properly. Some machine learning algorithms that use distance-based calculations require scaling of data to work more optimally. Here, while scaling, the scale differences between the features are minimized to obtain more accurate results. In order to minimize these differences, some standardization and normalization methods are used.

In this study, scaling problem was tried to be solved by using MaxAbsScaler, RobustScaler, StandartScaler, QuantileTransformer and normalizer methods. The most appropriate scaling algorithm was determined for the machine learning models and feature extraction method used, and the method that produced the best result was used.

In MaxAbsScaler, each property is scaled according to its maximum absolute value. Each property is scaled so that its maximum absolute value is "1". It does not shift or center the data, so that no sparsity is destroyed (Ahsan et al., 2021).

In RobustScaler, features are scaled using statistics based on outliers. When scaling, the median is removed and scales the data according to quantile intervals. The median and interquartile range can be used for scaling, as outliers will negatively affect the mean and variance (Ahsan et al., 2021).

StandardScaler is a scaling method in which the mean value of the variable distributions is converted to 0 and the standard deviation to 1, and the distribution is brought closer to the normal. It is performed by subtracting the average of the features and dividing by the standard deviation of the relevant feature. In this way, all observation units are reduced between -1 and 1.

QuantileTransformer scales features to follow a smooth or normal distribution. For a given attribute, this conversion tends to spread out the most frequent values. In addition, this method reduces the negative effects of outliers on learning (Ahsan et al., 2021).

Normalizer is a method of rescaling each sample that does not have a value of 0 and that has at least one component, independently of other samples, so that its norm is equal to 1.

The above methods were preferred for scaling the data used in the multinomial Naive Bayes and linear support vector classification models, which are used in the study and cannot produce reliable results without scaling. These methods have been tried in algorithms one by one and the most successful methods have been used.

Machine Learning

In this study, machine learning models to be applied to the vectors obtained by Bag of Words and TD-IDF; logistic regression, multinomial Naive Bayes, stochastic gradient descent, linear support vector classifier, perceptron and ridge classifier models. Tree-based classifiers are not preferred because it takes too much time to implement. Since there are approximately 210 thousand data in the data set, models that produce fast results were preferred.

Performance Metrics

Evaluation of the predictions was made by examining the precision (Prec), recall (Rec), f1score (F1), accuracy (Acc) and AUC (Powers, 2011) score. Abbreviations for the required values for the following formulas are: TP = Number of correctly predicted positive labeled samples, FP = Incorrectly predicted positive labeled samples, TN = Correctly predicted negative labeled samples, and FN = Incorrectly predicted negative labeled samples.

RESULTS

In this study, the data mining techniques described above were applied on the "review" feature in the data set, and the results were compared with 5 different feature scaling techniques, 4 different SMOTE methods and 6 different machine learning methods. While the classes have an uneven distribution, feature scaling techniques and machine learning algorithms are applied and results are obtained without applying SMOTE techniques. The max_features parameter of TF-IDF and bag of words techniques, which is described in the feature extraction section, has been selected as 100,000 and unlimited. Here, the purpose of limiting the max_features parameter is; is the measurement of the relationship between accuracy and time. Learning by including all the features, especially the linear support vector classifier algorithm, requires a lot of hardware power and time. Since a real-time system integration is planned in the future, it is aimed to minimize this time and hardware need by limiting the max_features parameter to 100,000, while making reasonable sacrifices in accuracy.

Some results were obtained from the multinomial Naive Bayes and linear support vector classifier algorithms, which were most affected by the feature scaling techniques described above, and in line with these results, the most successful feature scaling methods were used in the rest of the study. The results are shown in Table 2, 3.

In the TF-IDF method, we can see that for Multinomial Naive Bayes, the best results are obtained with the StandardScaler scaler. In addition, it is seen that MaxAbsScaler and QuantileTransformer scalers are feature scaling methods that give the best results on average for all algorithms.

Before applying the linear support vector classifier algorithm, the algorithm was executed by applying the QuantileTransformer scaler. It is seen that this feature scaling method gives the best results for the linear support vector classifier model with TF-IDF.

In the bag of words method, we can see that for multinomial Naive Bayes, the best results are obtained with the StandardScaler scaler. In addition, it is seen that MaxAbsScaler scaler is the feature scaling method that gives the best results on average for all algorithms.

Before applying the linear support vector classifier algorithm, the algorithm was executed by applying the MaxAbsScaler scaler. It is seen that this feature scaling method gives the best results for the linear support vector classifier model with Bag of Words.

Table 2. NO SMOTE/TF-IDF/max_feature = 100.000

Scale Methods	Model	Class	Precision	Recall	F1 score	Acc	AUC
MaxAbsScaler	MultinomialNB	Negative Positive	0.71 0.89	0.74 0.87	0.73 0.88	0.83	0.81
	LinearSVC	Negative Positive	0.86 0.94	0.87 0.94	0.86 0.94	0.92	0.90
RobustScaler	MultinomialNB	Negative Positive	0.66 0.83	0.59 0.87	0.62 0.85	0.79	0.73
	LinearSVC	Negative Positive	0.87 0.93	0.83 0.95	0.85 0.94	0.91	0.89
Normalizer	MultinomialNB	Negative Positive	0.79 0.79	0.41 0.95	0.54 0.87	0.79	0.68
	LinearSVC	Negative Positive	0.85 0.91	0.78 0.94	0.81 0.93	0.89	0.86
Quantile transformer	MultinomialNB	Negative Positive	0.70 0.89	0.76 0.86	0.73 0.88	0.83	0.81
	LinearSVC	Negative Positive	0.86 0.94	0.87 0.94	0.87 0.94	0.92	0.90
Standard scaler	MultinomialNB	Negative Positive	0.72 0.90	0.77 0.87	0.74 0.89	0.84	0.82
	LinearSVC	Negative Positive	0.79 0.93	0.83 0.91	0.81 0.92	0.89	0.87

SMOTE: Synthetic minority oversampling technique, Acc: Accuracy

Table 3. No SMOTE/BoW/max_feature = 100.000

Scale methods	Model	Class	Precision	Recall	F1 score	Acc	AUC
MaxAbsScaler	MultinomialNB	Negative Positive	0.72 0.89	0.74 0.88	0.73 0.88	0.84	0.81
	LinearSVC	Negative Positive	0.87 0.94	0.87 0.94	0.87 0.94	0.92	0.90
RobustScaler	MultinomialNB	Negative Positive	0.69 0.89	0.75 0.86	0.72 0.87	0.83	0.80
	LinearSVC	Negative Positive	0.86 0.94	0.87 0.94	0.87 0.94	0.92	0.91
Normalizer	MultinomialNB	Negative Positive	0.86 0.80	0.41 0.97	0.55 0.88	0.80	0.69
	LinearSVC	Negative Positive	0.84 0.91	0.78 0.94	0.81 0.93	0.89	0.86
Quantile transformer	MultinomialNB	Negative Positive	0.70 0.89	0.76 0.86	0.73 0.88	0.83	0.81
	LinearSVC	Negative Positive	0.86 0.94	0.87 0.94	0.87 0.94	0.92	0.91
StandardScaler	MultinomialNB	Negative Positive	0.71 0.90	0.76 0.87	0.73 0.88	0.84	0.81
	LinearSVC	Negative Positive	0.80 0.93	0.83 0.91	0.81 0.92	0.89	0.87

SMOTE: Synthetic minority oversampling technique, Acc: Accuracy

In line with the results obtained from both tables, StandardScaler scaler will be used for the Multinomial Naive Bayes algorithm in both methods (TF-IDF and BoW) in the machine learning applications to be made from now on. QuantileTransformer scaler will be used for linear support vector classifier algorithm while working in TF-IDF method, MaxAbsScaler scaler will be used for linear support vector classifier algorithm while operating in Bag of Words method. In this way, the algorithms will work with maximum efficiency.

The results obtained when the data set has uneven class distribution without applying SMOTE techniques are shown in Tables 4-7.

Table 4. No SMOTE/TF-IDF/max_feature no limit

Model	Class	Precision	Recall	F1 score	Acc	AUC
LogisticRegression	Negative Positive	0.85 0.88	0.69 0.95	0.76 0.91	0.87	0.82
Perceptron (MaxAbsScaler)	Negative Positive	0.90 0.94	0.86 0.96	0.88 0.95	0.93	0.91
RidgeClassifier	Negative Positive	0.92 0.93	0.84 0.97	0.88 0.95	0.93	0.90
MultinomialNB (StandarScaler)	Negative Positive	0.84 0.91	0.77 0.94	0.80 0.92	0.89	0.85
SGDClassifier	Negative Positive	0.85 0.87	0.66 0.95	0.74 0.91	0.86	0.92
LinearSVC (StandartScaler)	Negative Positive	0.77 0.90	0.77 0.91	0.77 0.90	0.87	0.84

SMOTE: Synthetic minority oversampling technique, Acc: Accuracy

Table 5. No SMOTE/BoW/max_feature no limit

Model	Class	Precision	Recall	F1 score	Acc	AUC
LogisticRegression	Negative Positive	0.90 0.94	0.87 0.96	0.88 0.95	0.93	0.91
Perceptron	Negative Positive	0.89 0.95	0.88 0.95	0.88 0.95	0.93	0.92
RidgeClassifier	Negative Positive	0.86 0.94	0.86 0.94	0.86 0.94	0.92	0.90
MultinomialNB (StandarScaler)	Negative Positive	0.85 0.90	0.77 0.94	0.80 0.92	0.89	0.85
SGDClassifier	Negative Positive	0.90 0.95	0.88 0.96	0.89 0.95	0.93	0.97
LinearSVC (StandartScaler)	Negative Positive	0.74 0.91	0.78 0.88	0.76 0.90	0.85	0.83

SMOTE: Synthetic minority oversampling technique, Acc: Accuracy

Table 6. No SMOTE/TF-IDF/max_feature = 100.000

Model	Class	Precision	Recall	F1 score	Acc	AUC
LogisticRegression	Negative Positive	0.82 0.88	0.70 0.94	0.76 0.91	0.87	0.82
Perceptron (MaxAbsScaler)	Negative Positive	0.86 0.94	0.85 0.94	0.85 0.94	0.91	0.90
RidgeClassifier	Negative Positive	0.86 0.91	0.78 0.95	0.82 0.93	0.90	0.86
MultinomialNB (StandarScaler)	Negative Positive	0.72 0.90	0.77 0.87	0.74 0.89	0.84	0.82
SGDClassifier	Negative Positive	0.83 0.87	0.68 0.94	0.75 0.91	0.86	0.92
LinearSVC (StandartScaler)	Negative Positive	0.79 0.93	0.83 0.91	0.81 0.92	0.89	0.87

SMOTE: Synthetic minority oversampling technique, Acc: Accuracy

Table 7. No SMOTE/ BoW/max_feature = 100.000

Model	Class	Precision	Recall	F1 score	Acc	AUC
LogisticRegression	Negative Positive	0.87 0.93	0.82 0.95	0.84 0.94	0.91	0.89
Perceptron	Negative Positive	0.85 0.94	0.87 0.94	0.86 0.94	0.92	0.90
RidgeClassifier	Negative Positive	0.80 0.92	0.82 0.91	0.81 0.92	0.88	0.87
MultinomialNB (StandarScaler)	Negative Positive	0.71 0.90	0.76 0.87	0.73 0.88	0.84	0.81
SGDClassifier	Negative Positive	0.88 0.92	0.81 0.95	0.84 0.94	0.91	0.95
LinearSVC (StandartScaler)	Negative Positive	0.79 0.93	0.83 0.91	0.81 0.92	0.89	0.87

SMOTE: Synthetic minority oversampling technique, Acc: Accuracy

Table 8, 9 were created to observe the SMOTE techniques described in the study and the effect of each technique on the algorithms used. The max_features value has been chosen as 100,000 because of the time it takes to create each SMOTE technique and the high hardware requirement and long waiting times that may occur due to the number of features. Comparing between SMOTE techniques, the most successful SMOTE technique was chosen and results were obtained.

As can be seen from the tables, it is seen that the results of almost all techniques and algorithms are very close. For this reason, it would be the most logical choice to choose the Standard SMOTE technique in terms of time and cost. The SMOTE mentioned in the next results will be the Standard SMOTE technique.

In this section, the following Tables 10-13 were created by using the SMOTE technique, that is, by making the class distribution in the data set more balanced compared to the past.

It is seen in the tables that the linear support vector classifier model gives the best results in almost all results for both methods and the max_features parameter. For the TF-IDF method and the max_features parameter value of 100,000, the best algorithm was the linear support vector classifier algorithm with an Accuracy value of 92%. For the bag of words method and the max_features parameter value of 100,000, the best algorithm was the linear support vector classifier algorithm with an accuracy value of 92%. While the TF-IDF method and max_features parameter had no limit, the best algorithm was the linear support vector classifier algorithm with an Accuracy value of 93%. While bag of words method and max_features parameter had no limit, the best algorithm was the SGD Classifier algorithm with an Accuracy value of 93%.

DISCUSSION

Although the results obtained from both methods are good, it can be developed by using different methods to create a real-time recommendation system or the existing methods can be improved. Other smote methods such as K-Means smote and SVM smote, which were not used in this study, can be used to minimize the class imbalance problem in the data set. The "usefulCount" property in the dataset can be included in machine learning as an independent variable. In order for bag of words and TF-IDF methods to produce more meaningful word groups, the most appropriate value for the max_features parameter can be determined. In addition to the feature extraction methods used in the study, Word2vec and different feature extraction methods can be used. Also feature extraction can be performed manually. In addition to the machine learning algorithms used in the study, community-based machine learning algorithms can be used. The results can be made even better by making improvements on existing machine learning algorithms.

In this study, sentiment analysis was performed by classification method over categorical text data. The aim is to show the methodology that makes the best use of each data and feature in the data set when performing the sentiment analysis.

Table 8. SMOTE/TF-IDF/max_feature=100.000

SMOTE	Model	Class	Precision	Recall	F1 score	Acc	AUC
Standart smote	LogisticRegression	Negative Positive	0.77 0.91	0.79 0.90	0.78 0.91	0.87	0.85
	Perceptron (MaxAbsScaler)	Negative Positive	0.85 0.94	0.87 0.93	0.86 0.94	0.91	0.90
	RidgeClassifier	Negative Positive	0.82 0.93	0.84 0.92	0.83 0.93	0.90	0.88
	MultinomialNB (StandartScaler)	Negative Positive	0.72 0.90	0.76 0.88	0.74 0.89	0.84	0.82
	SGDClassifier	Negative Positive	0.77 0.90	0.77 0.90	0.77 0.90	0.86	0.92
	LinearSVC (Quantile transformer)	Negative Positive	0.86 0.94	0.87 0.94	0.87 0.94	0.92	0.90
Adasyn	LogisticRegression	Negative Positive	0.77 0.91	0.80 0.90	0.78 0.91	0.87	0.85
	Perceptron (MaxAbsScaler)	Negative Positive	0.85 0.94	0.87 0.94	0.86 0.94	0.92	0.90
	RidgeClassifier	Negative Positive	0.82 0.93	0.85 0.92	0.83 0.93	0.90	0.88
	MultinomialNB (StandartScaler)	Negative Positive	0.72 0.90	0.78 0.87	0.75 0.89	0.84	0.83
	SGDClassifier	Negative Positive	0.76 0.91	0.78 0.90	0.77 0.90	0.86	0.92
	LinearSVC (Quantile transformer)	Negative Positive	0.86 0.94	0.87 0.94	0.87 0.94	0.92	0.91
SMOTE-tomek	LogisticRegression	Negative Positive	0.78 0.91	0.79 0.90	0.78 0.91	0.87	0.85
	Perceptron (MaxAbsScaler)	Negative Positive	0.85 0.94	0.86 0.94	0.85 0.94	0.91	0.90
	RidgeClassifier	Negative Positive	0.82 0.93	0.84 0.92	0.83 0.93	0.90	0.88
	MultinomialNB (StandartScaler)	Negative Positive	0.72 0.90	0.76 0.88	0.74 0.89	0.84	0.82
	SGDClassifier	Negative Positive	0.76 0.90	0.77 0.90	0.77 0.90	0.86	0.92
	LinearSVC (Quantile transformer)	Negative Positive	0.86 0.94	0.87 0.94	0.86 0.94	0.92	0.90
Borderline	LogisticRegression	Negative Positive	0.77 0.91	0.80 0.90	0.78 0.91	0.87	0.84
	Perceptron (MaxAbsScaler	Negative Positive	0.85 0.94	0.86 0.94	0.86 0.94	0.92	0.90
	RidgeClassifier	Negative Positive	0.82 0.93	0.84 0.92	0.83 0.93	0.90	0.88
	MultinomialNB (StandartScaler)	Negative Positive	0.72 0.90	0.77 0.87	0.74 0.89	0.84	0.82
	SGDClassifier	Negative Positive	0.76 0.91	0.78 0.90	0.77 0.90	0.86	0.92
	LinearSVC (Quantile transformer)	Negative Positive	0.86 0.94	0.87 0.94	0.86 0.94	0.92	0.90

SMOTE: Synthetic minority oversampling technique, Acc: Accuracy

CONCLUSION

User comments on various systems we use in our daily life have an important role in choosing the service we want to receive. In this study, in which sentiment analysis was performed on user comments, logistic regression, perceptron, ridge classifier, multinomial Naive Bayes, SGD classifier and linear support vector classifier machine learning algorithms were applied to the data obtained by bag of words and TF-IDF methods. Obtained results were evaluated with precision, recall, F1 score, accuracy and AUC score metrics. The best result for the bag of words method is the SGD classifier algorithm with an accuracy value of 93%. The best result for the TF-IDF method was the linear support vector classifier algorithm with an accuracy value of 93%. In both methods, there are machine learning algorithms that are very close to the best algorithms and have the same accuracy value. However, when looking at all the evaluation metrics, the best algorithms were selected.

Table 9. SMOTE/BoW/max_feature=100.000

SMOTE	Model	Class	Precision	Recall	F1 score	Acc	AUC
Standart smote	LogisticRegression	Negative Positive	0.83 0.93	0.85 0.92	0.84 0.93	0.90	0.89
	Perceptron	Negative Positive	0.84 0.95	0.87 0.93	0.86 0.94	0.91	0.90
	RidgeClassifier	Negative Positive	0.78 0.93	0.84 0.90	0.81 0.91	0.88	0.87
	MultinomialNB (StandartScaler)	Negative Positive	0.71 0.89	0.75 0.87	0.73 0.88	0.84	0.81
	SGDClassifier	Negative Positive	0.82 0.94	0.87 0.92	0.84 0.93	0.90	0.95
	LinearSVC (MaxAbsScaler)	Negative Positive	0.86 0.95	0.87 0.94	0.86 0.94	0.92	0.91
Adasyn	LogisticRegression	Negative Positive	0.83 0.93	0.85 0.93	0.84 0.93	0.90	0.89
	Perceptron	Negative Positive	0.72 0.92	0.82 0.87	0.77 0.89	0.85	0.84
	RidgeClassifier	Negative Positive	0.78 0.93	0.84 0.90	0.81 0.91	0.88	0.87
	MultinomialNB (StandartScaler)	Negative Positive	0.71 0.89	0.75 0.87	0.73 0.88	0.84	0.81
	SGDClassifier	Negative Positive	0.84 0.93	0.84 0.93	0.84 0.93	0.90	0.95
	LinearSVC (MaxAbsScaler)	Negative Positive	0.85 0.95	0.87 0.94	0.86 0.94	0.92	0.91
SMOTE-tomek	LogisticRegression	Negative Positive	0.83 0.93	0.84 0.93	0.84 0.93	0.90	0.88
	Perceptron	Negative Positive	0.73 0.92	0.82 0.87	0.77 0.90	0.86	0.85
	RidgeClassifier	Negative Positive	0.78 0.93	0.84 0.90	0.81 0.91	0.88	0.87
	MultinomialNB (StandartScaler)	Negative Positive	0.71 0.89	0.75 0.87	0.73 0.88	0.84	0.81
	SGDClassifier	Negative Positive	0.81 0.95	0.88 0.94	0.84 0.93	0.90	0.95
	LinearSVC (MaxAbsScaler)	Negative Positive	0.85 0.95	0.87 0.94	0.86 0.94	0.92	0.91
Borderline	LogisticRegression	Negative Positive	0.83 0.94	0.85 0.93	0.84 0.93	0.90	0.89
	Perceptron	Negative Positive	0.72 0.92	0.83 0.86	0.77 0.89	0.85	0.85
	RidgeClassifier	Negative Positive	0.78 0.93	0.84 0.90	0.81 0.91	0.88	0.87
	MultinomialNB (StandartScaler)	Negative Positive	0.71 0.89	0.75 0.87	0.73 0.88	0.84	0.81
	SGDClassifier	Negative Positive	0.83 0.94	0.85 0.93	0.84 0.93	0.90	0.95
	LinearSVC (MaxAbsScaler)	Negative Positive	0.85 0.95	0.87 0.94	0.86 0.94	0.92	0.91

SMOTE: Synthetic minority oversampling technique, Acc: Accuracy

Table 10. SMOTE/TF-IDF/max_feature no limit

Model	Class	Precision	Recall	F1 score	Acc	AUC
LogisticRegression	Negative Positive	0.80 0.91	0.79 0.92	0.80 0.91	0.88	0.85
Perceptron (MaxAbsScaler)	Negative Positive	0.90 0.94	0.86 0.96	0.88 0.95	0.93	0.91
RidgeClassifier	Negative Positive	0.90 0.95	0.87 0.96	0.88 0.95	0.93	0.91
MultinomialNB (StandarScaler)	Negative Positive	0.86 0.90	0.75 0.95	0.80 0.92	0.89	0.85
SGDClassifier	Negative Positive	0.78 0.90	0.76 0.91	0.77 0.90	0.87	0.92
LinearSVC (Quantile transformer)	Negative Positive	0.91 0.95	0.87 0.96	0.89 0.95	0.93	0.92

SMOTE: Synthetic minority oversampling technique, Acc: Accuracy

Table 11. SMOTE/BoW/max_feature no limit

Model	Class	Precision	Recall	F1 score	Acc	AUC
LogisticRegression	Negative Positive	0.87 0.95	0.89 0.94	0.88 0.95	0.93	0.92
Perceptron	Negative Positive	0.86 0.95	0.89 0.96	0.88 0.95	0.93	0.92
RidgeClassifier	Negative Positive	0.83 0.95	0.88 0.92	0.85 0.93	0.91	0.90
MultinomialNB (StandarScaler)	Negative Positive	0.85 0.90	0.76 0.94	0.80 0.92	0.89	0.85
SGDClassifier	Negative Positive	0.88 0.95	0.88 0.95	0.88 0.95	0.93	0.96
LinearSVC (MaxAbsScaler)	Negative Positive	0.87 0.95	0.88 0.94	0.87 0.95	0.92	0.91

Acc: Accuracy

Table 12. SMOTE/TF-IDF/max_feature = 100.000

Model	Class	Precision	Recall	F1 score	Acc	1
LogisticRegression	Negative Positive	0.77 0.91	0.79 0.90	0.78 0.91	0.87	0.85
Perceptron (MaxAbsScaler)	Negative Positive	0.85 0.94	0.87 0.93	0.86 0.94	0.91	0.90
RidgeClassifier	Negative Positive	0.82 0.93	0.84 0.92	0.83 0.93	0.90	0.88
MultinomialNB (StandarScaler)	Negative Positive	0.72 0.90	0.76 0.88	0.74 0.89	0.84	0.82
SGDClassifier	Negative Positive	0.77 0.90	0.77 0.90	0.77 0.90	0.86	0.92
LinearSVC (Quantile transformer)	Negative Positive	0.86 0.94	0.87 0.94	0.87 0.94	0.92	0.90

SMOTE: Synthetic minority oversampling technique, Acc: Accuracy

Table 13. SMOTE/BoW /max_feature = 100.000

Model	Class	Precision	Recall	F1 score	Acc	AUC
LogisticRegression	Negative Positive	0.83 0.93	0.85 0.92	0.84 0.93	0.90	0.89
Perceptron	Negative Positive	0.84 0.95	0.87 0.93	0.86 0.94	0.91	0.90
RidgeClassifier	Negative Positive	0.78 0.93	0.84 0.90	0.81 0.91	0.88	0.87
MultinomialNB (StandartScaler)	Negative Positive	0.71 0.89	0.75 0.87	0.73 0.88	0.84	0.81
SGDClassifier	Negative Positive	0.82 0.94	0.87 0.92	0.84 0.93	0.90	0.95
LinearSVC (MaxAbsScaler)	Negative Positive	0.86 0.95	0.87 0.94	0.86 0.94	0.92	0.91

SMOTE: Synthetic minority oversampling technique, Acc: Accuracy

ETHICAL DECLARATIONS

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Financial Disclosure

The authors declared that this study has received no financial support.

Author Contributions

All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Abbasi Moud, Z., Vahdat-Nejad, H., & Sadri, J. (2021). Tourism recommendation system based on semantic clustering and sentiment analysis. *Expert Syst App*, 167,114324.
2. Ahsan, M.M., Mahmud, M.P., Saha, P.K., Gupta, K.D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, 9(3),52.
3. Alamoodi, A.H., Zaidan, B.B., Al Masawa, M., Taresh, S.M., Noman, S., Ahmaro, I.Y., ... & Salahaldin, A. (2021). Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy. *Comput Biol Med*, 139,104957.
4. Bao, Y., & Jiang, X. (2016). An intelligent medicine recommender system framework. In 2016 IEEE 11th conference on industrial electronics and applications (ICIEA) (pp. 1383-1388). IEEE.
5. Chen, Y., Chang, R., & Guo, J. (2021). Effects of data augmentation method borderline-SMOTE on emotion recognition of EEG signals based on convolutional neural network. *IEEE Access*, 9,47491-47502.
6. Dağıstanlı, Ö., Erbay, H., Kör, H., & Yurttakal, A. H. (2023). Reflection of people's professions on social media platforms. *Neural Comput Applications*, 35(7),5575-5586.
7. Dinakaran, D., Manjunatha, N., Kumar, C.N., & Math, S.B. (2021). Telemedicine practice guidelines of India, 2020: Implications and challenges. *Indian J Psychiatry*, 63(1),97-101.
8. Doulaverakis, C., Nikolaidis, G., Kleontas, A., & Kompatsiaris, I. (2012). GalenOWL: ontology-based drug recommendations discovery. *J Biomed Semantics*, 3,1-9.
9. Feng, S., Keung, J., Yu, X., Xiao, Y., & Zhang, M. (2021). Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction. *Informat Software Technol*, 139,106662.
10. Fox, S. (2013). Health online 2013. *Pew Intern Am Life Project.*
11. Garg, S. (2021). Drug recommendation system based on sentiment analysis of drug reviews using machine learning. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 175-181). IEEE.
12. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE İnternational Joint Conference on Neural Networks (pp. 1322-1328). IEEE.
13. Hossain, M. D., Azam, M. S., Ali, M. J., & Sabit, H. (2020). Drugs rating generation and recommendation from sentiment analysis of drug reviews using machine learning. In 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE) (pp. 1-6). IEEE.
14. Islek, I., & Oguducu, S.G. (2022). A hierarchical recommendation system for e-commerce using online user reviews. *Electronic Commer Res Appl*, 52,101131.
15. Jalili, M., Ahmadian, S., Izadi, M., Moradi, P., & Salehi, M. (2018). Evaluating collaborative filtering recommender algorithms: a survey. *IEEE Access*, 6,74003-74024.
16. Kaur, H., & Mangat, V. (2017). A survey of sentiment analysis techniques. In 2017 International Conference on I-SMAC. (pp. 921-925). IEEE.
17. Kaur, H., & Mangat, V. (2017). A survey of sentiment analysis techniques. In 2017 International Conference on I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC) (pp. 921-925). IEEE.
18. Khatter, H., Goel, N., Gupta, N., & Gulati, M. (2021). Movie recommendation system using cosine similarity with sentiment analysis. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 597-603). IEEE.
19. Oyamada, M. (2019). Extracting feature engineering knowledge from data science notebooks. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 6172-6173). IEEE.
20. Pereira, J.G., Tiwari, S., & Ajoy, S. (2020). A survey on filtering techniques for recommendation system. In 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC) (pp. 1-6). IEEE.
21. Powers, D.M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. ArXiv Preprint ArXiv:2010.16061.
22. Pu, W., Liu, N., Yan, S., Yan, J., Xie, K., & Chen, Z. (2007). Local word bag model for text categorization. In Seventh IEEE international conference on data mining (ICDM 2007) (pp. 625-630). IEEE.
23. Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *Int J Comput Appl*, 181(1),25-29.
24. Rodríguez, P., Bautista, M.A., Gonzalez, J., & Escalera, S. (2018). Beyond one-hot encoding: lower dimensional target embedding. *Image Vision Computing*, 75,21-31.
25. Sharma, A.K., Bajpai, B., Adhvaryu, R., Pankajkumar, S.D., Gordhanbhai, P.P., & Kumar, A. (2023). An efficient approach of product recommendation system using NLP technique. *Mater Today Proceed*, 80,3730-3743.
26. Shimada, K., Takada, H., Mitsuyama, S., Ban, H., Matsuo, H., Otake, H., ... & Kaku, M. (2005). Drug-recommendation system for patients with infectious diseases. In AMIA Annual Symposium Proceedings (Vol. 2005, p. 1112). American Medical Informatics Association.
27. Tekade, T.N., & Emmanuel, M. (2016). Probabilistic aspect mining approach for interpretation and evaluation of drug reviews. In 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES) (pp. 1471-1476). IEEE.

28. Wang, Z.H.E., Wu, C., Zheng, K., Niu, X., & Wang, X. (2019). SMOTETomek-based resampling for personality recognition. *IEEE Access*, 7,129678-129689.

29. Wittich, C. M., Burkle, C. M., & Lanier, W. L. (2014). Medication errors: an overview for clinicians. In Mayo Clinic Proceedings (Vol. 89, No. 8, pp. 1116-1125). Elsevier.

30. Zhang, Y., & Zhang, L. (2022). Movie recommendation algorithm based on sentiment analysis and LDA. *Procedia Comput Sci*, 199,871-878.

31. Zhang, Y., Zhang, D., Hassan, M.M., Alamri, A., & Peng, L. (2015). CADRE: Cloud-assisted drug recommendation service for online pharmacies. *Mobile Networks Appl*, 20,348-355.