Turkish Journal of Engineering
https://dergipark.org.tr/en/pub/tuje
e-ISSN 2587-1366

Sentiment analysis with ensemble and machine learning methods in multi-domain datasets

Muhammet Sinan Başarslan [*1], **Fatih Kayaalp**[2]

[1]*Istanbul Medeniyet University, Department of Computer Engineering, Türkiye*
[2]*Duzce University, Department of Computer Engineering, Türkiye*

Keywords
Ensemble Learning
Machine Learning
Sentiment Analysis
Text Representation

Research Article
DOI: 10.31127/tuje.1079698

Received: 26.02.2022
Accepted: 07.04.2022
Published: 26.04.2022

Abstract

The first place to get ideas on all the activities considered to occur in everyday life was the comments on the websites. This is an area that deals with these interpretations in the natural language processing, which is a sub-branch of artificial intelligence. Sentiment analysis studies, which is a task of natural language processing are carried out to give people an idea and even guide them with such comments. In this study, sentiment analysis was implemented on public user feedback on websites in two different areas. TripAdvisor dataset includes positive or negative user comments about hotels. And Rotten Tomatoes dataset includes positive (fresh) or negative (rotten) user comments about films. Sentiments analysis on datasets have been carried out by using Word2Vec word embedding model, which learns the vector representations of each word containing the positive or negative meaning of the sentences, and the Term Frequency Inverse Document Frequency text representation model with four machine learning methods (Naïve Bayes-NB, Support Vector Machines-SVM, Logistic Regression-LR, K-Nearest Neighbour-kNN) and two ensemble learning methods (Stacking, Majority Voting-MV). Accuracy and F-measure is used as a performance metric experiments. According to the results, Ensemble learning methods have shown better results than single machine learning algorithms. Among the overall approaches, MV outperformed Stacking.

1. Introduction

Software and hardware technologies are in constant evolution. As a result of this development, people's habits also change. Communication devices have been improved over time. Especially with the improvement of intelligent devices, the first place to be referred to, in many areas, especially in the health field, is the apps and websites on these devices.

The opinions of the people who post comments on these two websites can be seen from every device that has access to the internet and these opinions give an idea to the users. Naturally, before going to a movie or deciding on a hotel to stay at, it is important for people to visit these sites and see the experiences and comments of other people.

Positive (fresh) or negative (rotten) comments on a movie on the Rotten tomatoes site can be a guide for those who would watch a movie for the first time. Apart from the comments on the movie, information on the director, actor reviews, and cinema news are also included and it functions as a guide for the site users. Similarly, users who look for hotel and restaurant reviews on the TripAdvisor site before planning a trip can make their decisions on the hotels or restaurants they would prefer to visit in their trips, and even cancel their travels. It was seen that the comments on these websites pushed the communication to a process that guides people with the infrastructure of internet technologies. Companies that want to benefit from this process have supported the studies that extract sentiments from the comments using natural language processing techniques and this extracting process is a sub-branch of artificial intelligence. This process is not limited to the selection of movies or hotels and has turned into an area where people can get ideas and benefit from many features.

* **Corresponding Author**

*(msinanbasarslan@gmail.com) ORCID ID 0000-0002-7996-9169
(fatihkayaalp@duzce.edu.tr) ORCID ID 0000-0002-8752-3335

Cite this article

Başarslan, M. S., & Kayaalp, F. (2023). Sentiment analysis with ensemble and machine learning methods in multi-domain datasets. Turkish Journal of Engineering, 7(2), 141-148

In this study, a sentiment analysis study was carried out on the comments that were collected from two websites accessible to everyone and that were labeled with a sentiment class. After feature extraction with the help of a Word2Vec (W2V) model that learns the vector representations of the words in the comments and the Term frequency Inverse document frequency (TF-IDF) model based on text frequency, sentiment analysis was performed by creating comprehensive models with machine learning and collective learning methods. In comparing the performances of the models created for sentiment analysis, the training and test cluster distinctions were made in two ways as the holdout method, training-test sets, 80% -20%, and 70%-30%, respectively. Before creating classifier models for sentiment analysis, text pre-processing processes such as clearing numbers of comments, removing numbers and special characters, converting all letters to the lowercase, and stemming processes were carried out.

Our contribution to the literature covered in the present study;

The effect of text representation methods by frequency (TF-IDF) and by the prediction (W2V) on performance has been studied.

Creating a high-performance model in the analysis of sentiments, the impact of general learning methods and single machine learning methods for performance has been studied.

With the heterogeneous use of single machine learning methods in ensemble learning methods, the effect of the model on performance has been analyzed. The workflow process followed during the study can be seen in Figure 1.

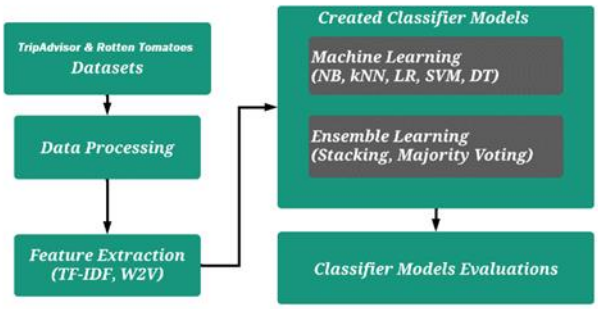

Figure 1. The flowchart of the study

As can be seen in Figure 1, after the data labeled with sentiment class were passed through the pre-processing process, separate text representation was created with both word frequency-based TF-IDF and predicted-based W2V methods. Then, models for machine learning, ensemble learning, and sentiment classes were created and the results of these models were obtained.

In this study, previous studies or similar studies with the sentiment analysis study on the hotel reviews on the TripAdvisor website and the criticisms of the movies, actors, and directors on the Rotten Tomatoes website are explained in the second section.

Information on the algorithms used in classifier models, ensemble learning algorithms, text representation methods, and the datasets used are presented in the third section. Performance metrics and the experimental results are presented in section four. Evaluations of the results and future works are presented in section five.

2. Related works

Reviews on websites such as TripAdvisor and Rotten Tomatoes are a big part of the sentiment analysis study. The ability to access these sites from any device and the comments made for movies or hotels give people an introductory idea about their preferences.

Sentiment analysis, which is the subject of classifying people's sentiments using natural language processing, and which is also a sub-branch of artificial intelligence, is examined in a multi-disciplinary fashion. In this section, machine learning models and interpretation, classification models will be analyzed. The authors collected reviews from the TripAdvisor site on five hotels in Aswan, Egypt, which received a total of 11.458 reviews. They used TF-IDF for text representation of these comments. NB yielded the highest accuracy Score among the models they created with SVM, NB, and Decision Tree (DT) classifiers for sentiment analysis [1].

The authors used a TripAdvisor dataset that It consists of approximately 250000 customer-provided reviews of 1850 hotels. They suggested feature extraction of these studies and Subjectivity Based Feature Extraction. They created a model through SVM, NB, and DT, and the SVM yielded the best result of 87.51% with the help of the method they suggested [2].

They collected a total of 2000 reviews from the TripAdvisor website, 500 positive and 500 negative reviews for training sets and 500 positives and 500 negative reviews for test sets. They created a model with SVM, NB, and DT. Through the machine learning method, they achieved an accuracy score of 87 % in hotel assessment classification with the help of SVM [3].

In sentiment analysis performed on the Rotten Tomatoes Analysis Dataset, the n-gram method (Bigram, Unigram, Trigram) and the combination of various n-grams were used for text representations. In the study, models were created in SVM, Maximum Entropy, and NB. The study notes that it yields better results for unigram, bigram, and trigram methods, but the score decreases when observed for four grams, five grams, six grams, and more [4].

Of the one hundred thousand views about the hotel, 70% is seperated into training sets, 10% is validation sets, and 20% is divided into test sets. GloVe built models with Bi-Directional Long Short-Term Memory (Bi-LSTM) and various Convolutional Neural Networks for emotion classification after text representations with FastText. The Bi-LSTM model with GloVe word embedding technique achieved the best performance with 73.73% test accuracy [5].

They used a movie dataset on Rotten Tomatoes consisting of 8000 polar movie reviews. They created models with RF, kNN, NB, and Bagging classifiers. In their study, the RF technique achieved the highest accuracy score of 95% [6].

Rotten Tomatoes Film Dataset created various methods such as NB, Instance-Based Learning, DT, SVM

in the classification study on comments. The kNN algorithm performed between 65% and 95% on average [7].

Using W2V technique and Machine Learning techniques, a Sentiment Analysis Model has been proposed to analyze the emotions of Egyptian students in the learning process with the pandemic. The word embedding process was then evaluated by NB, SVM and DT classification, and evaluated for precision, recall and accuracy [8].

During the COVID-19 pandemic from Twitter in 2020, English tweets were classified as positive or negative by applying the LR algorithm to them, using this method they achieved a classification accuracy of 78.5% [9].

In the study, MultinomialNB on Twitter datasets, Results were obtained with BernoulliNB, LR, Stochastic Gradient Descent (SGD). In the experimental results, BernoulliNB, LR and SGD classifier reached up to 75% accuracy [10].

In this study, three different ensemble Machine Learning models are proposed to classify the data of approximately 12 thousand tweets in the UK into three emotion tags. First, the stacking classifier gave the highest F1 score of 83.5%, while in the second model the voting classifier gave 83.3% and in the last model the bagging classifier gave 83.2% results [11].

As seen in these studies, experiments were carried out on text representations or classifying models for performance improvement in the models created in sentiment analysis studies. Similarly, in this study, models based on ensemble learning in which traditional and predictive text representation methods are used together with different classification models, are created and their effect on the classification is analyzed.

3. Methodology

In this section, text representations, machine and ensemble learning algorithms, and datasets will be discussed.

3.1. Datasets

In the study, comments from public hotel review sites, which are shared as open-source by those who collected from these sites, were used.

TripAdvisor hotel reviews dataset [12] consists of data from 20490 hotel reviews. Reviews with 1,2 stars are marked as negative, 3 stars as neutral, and 4,5 stars as positive. The dataset consists of two attributes, hotel reviews and ratings.

Rotten Tomatoes films and critic reviews dataset has movie reviews, labeled as 240000 fresh, and 240000 rotten [13]. Reviews of Gervais with 1,2 stars were marked as negative, ones with 3 stars as neutral and ones with 4,5 stars as positive. The dataset consists of two attributes, film reviews and rating.

Hotel and Movie datasets were preprocessed before classification, including removing punctuation and symbols, conversion of characters to lower case and stemming. Also, stop words have been removed. The python NLTK library is used for these operations.

3.2. Text representation

Methods used to represent the text in documents pave the way for successful results in classification. Text representation methods are divided into two methods as Frequency Based Representation and Prediction Based Representation. Frequency-based text representation, which is defined as the more traditional method, is based on the principle of identifying the words in documents and the frequency of these words. The following text representation methods, W2V and TF-IDF, were used in this study.

3.2.1. TF-IDF

TF-IDF is the weight factor calculated by the statistical method that shows the importance of a term in a document.

TF is the method used to calculate the weight of a term in a document. There are weight calculation methods such as binary, raw frequency, and logarithm normalization.

IDF attempts to figure out whether a word is a term and not a Stop Word by detecting the number of occurrences of the word in more than one document. It is calculated by dividing the Number of Documents Elapsed by the Period by the absolute value of the logarithm of the Number of Documents.

3.2.2. Word2Vec

Word2Vec is an unsupervised and prediction-based model that attempts to express words in vector space. It was invented in 2013 by Google researcher Tomas Mikolov and his team. There are 2 types of sub-methods: Continuous Bag of Words (CBOW) and Skip-Gram [12].

CBOW and Skip-Gram models differ from each other in receiving input and output data. In the CBOW model, words that are not in the center of the window size Are taken as input and the words in the center are predicted to be output; In the Skip-Gram model, the words in the center are taken as input and the words that are not in the center are predicted to be output.

This process continues until the whole sentence is processed. This operation is applied to all of the sentences and the mapping operation is applied to the unlabeled data available at the beginning and it is then ready to be trained.

In the study, hyperparameter settings for vector size of 100 and 200, the window size of 5, the Sub-sampling Rate of 1e-3, Min-count of 5 were taken from the Word2Vec method. CBOW from W2V methods was used.

3.3. Machine learning

Machine learning is a sub-branch of computer science that was developed from studies of numerical learning and the recognition of models in artificial intelligence in 1959. Machine learning is a system that investigates the work of algorithms that can make predictions on data. The classifier algorithms Logistic LR, SVM, NB, and kNN were used in the study. It is seen in Figure 2. These Four algorithms are discussed in this section.

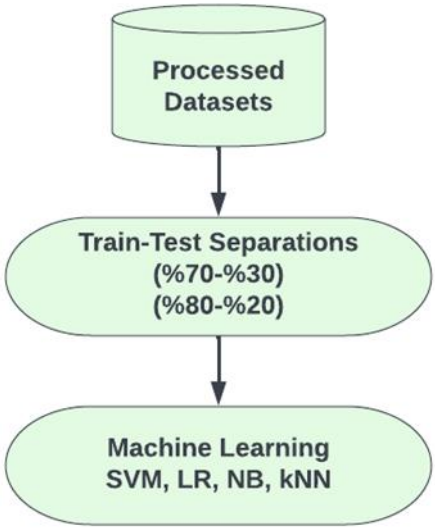

Figure 2. The flowchart of the Machine learning process

3.3.1. Naïve bayes

It is a lazy learning algorithm based on Bayes' theorem. It calculates all probabilities for each element in the data set and performs the classification based on the higher results [15].

3.3.2. Support vector machine

SVM is a supervised learning algorithm based on statistical learning theory. It is used to distinguish between two classes of data in the most appropriate way. The data set is divided into two, based on linear separation and non-separation [16].

3.3.3. K-Nearest neighbor

kNN is one of the easy-to-implement supervised learning algorithms. Although it is used to solve both classification and regression problems, it is mostly used in the solution of classification problems in the industry.

kNN algorithms were proposed by [17]. The algorithm is utilized by means of making use of data from a sample set whose classes are known. The distance of the new data to be included in the sample data set is calculated based on the existing data and the k number of close neighbors is checked. Generally, 3 types of distance functions are used for the calculation of distance [18]:

- "Euclidean" Distance
- Distance to "Manhattan"
- "Minkowski" is the Distance.

3.3.4. Logistic regression

It is a statistical method used to analyze a data set with one or more independent variables that determine the result. The result is measured by a binary variable [19].

3.4. Ensemble learning

Ensemble Learning is the combined use of machine algorithms of the same types (homogeneous Ensemble Learning) or different types (heterogeneous Ensemble Learning). It is based on the use of various models together to improve the performance achieved by a single algorithm [20].

Homogeneous or heterogeneous machine learning models have a variety of Ensemble Learning methods such as Bagging, Boosting, and Stacking, depending on the decision function (average, voting, etc.). In addition to these, two important variations have been developed one of these is the Voting, which complements Bagging, and the Blending, a subtype of Stacking. Although Voting is a subtype of Bagging and Blending is a subtype of Stacking, these techniques are frequently referred to as types of Ensemble Learning. It is seen in Figure 3.

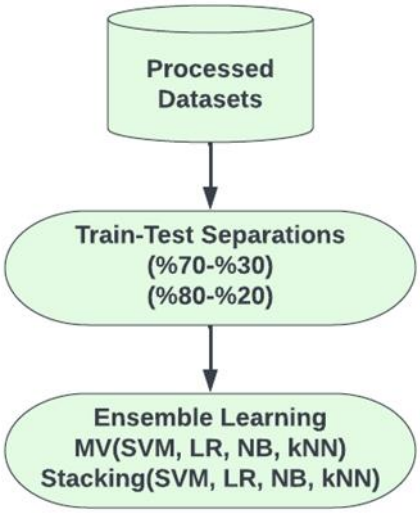

Figure 3. The flowchart of the Ensemble learning process

3.4.1. Voting

It is an effective and a simple ensemble algorithm used in classification processes. Minimum two sub-models are created and each sub-model determines the results of the model by combining the predictions through a vote that determines the prediction result by taking the average of the predictions [21]. In this study, majority voting was used.

3.4.1.1. Majority voting

In majority voting, every individual classifier vote for a class, and the majority wins. In statistical terms, the predicted target label of the ensemble is the mode of the distribution of individually predicted labels.

Here, we predict the class label $\hat{y}$ via majority (plurality) voting of each classifier C. This equation (1) is given.

$$\hat{y} = mode\{C_1(x), C_2(x), \dots \dots C_m(x)\} \quad (1)$$

Assuming that we combine three classifiers that classify a training sample as in the following equation (2):

$$\hat{y} = mode\{0,0,1\} = 0 \quad (2)$$

Via majority vote, we would classify the sample as "class 0." [22].

3.4.2. Stacking

Stacking is an extension of the Voting method used in classification processes. More than one sub-model may be chosen. It also allows the use of a different model for the best combination of predictions [23].

4. Experiments and results

In the experimental analysis, the two data sets were divided by using a test training separation method called holdout (80% -20% and 70% -30%). Python scikit-learn library was used in the experiments.

In the experimental analysis, NB, SVM, LR, and kNN classifiers were used. Stacking and Voting methods were used in ensemble classification. Experimental evaluations were carried out on a computer having a 4.1 GHz AMD RYZEN 7 2700 CPU with 32.00 GB RAM.

4.1. Performance metrics

The performance metric used to evaluate the predictive performance of sentiment classification models is the accuracy and F-measure. Accuracy is one of the most commonly used metrics. It is the ratio of true negatives (TN) and true positives (TP) to the total number of samples as given by Equation (3) [24].

$$ACC = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (3)$$

Abbreviations in Equation 1 show the number of TN, the number of TP, the number of false positives (FP), and the number of false negatives (FN).

The F-measure value is shown in Equation (4) [24].

$$F - measure\ (F) = \frac{2 * Precision\ * Sensitivity}{Precision\ + Sensitivity} \quad (4)$$

4.2. Experimental results

TF-IDF and W2V text representations in TripAdvisor and Rotten Tomatoes data sets, followed by 80-20% and 70-30% train-test separation, NB, kNN, LR, SVM machine learning methods, and Stacking and Voting ensemble learning methods using these methods together with sentiment classification models were created using Table 1, Table 2 show the accuracy results of the models obtained with TripAdvisor, Table 3 and Table 4 Rotten Tomatoes. Table 5 and Table 6 show TripAdvisor's F-measure result and Table 7 and Table 8 show the F-measure result for Rotten Tomatoes.

Table 1. Accuracy results of methods on TripAdvisor Dataset (80:20 training and test sets)

	NB	kNN	LR	SVM	Stacking	Voting
TF-IDF	83.1	80.2	82.8	83.3	83.6	**84.2**
W2V	83.5	81.6	83.4	83.6	84.1	**84.7**

When the Table 1 is analyzed, it is seen that SVM has the best results in an accuracy score among the single machine learning methods. When the ensemble learning methods are compared against single machine learning methods, it is seen that both ensemble learning methods have demonstrated better performance than single machine learning methods. Furthermore, Voting has demonstrated better results than stacking among ensemble learning methods.

Table 2. Accuracy results of methods on TripAdvisor Dataset (70:30 training and test sets)

	NB	kNN	LR	SVM	Stacking	Voting
TF-IDF	83.3	82.1	84.1	85	85.8	**86.2**
W2V	85.5	80.1	85.8	86.9	86.5	**87.3**

When the Table 2 is analyzed, it is seen that SVM has the best results in an accuracy score among the single machine learning methods with TF-IDF. However, the LR achieves the best Results with W2V. When the ensemble learning methods are compared against single machine learning methods, it is seen that Stacking is slightly behind LR and Voting is ahead of LR. Furthermore, Voting has demonstrated better results than Stacking among ensemble learning methods.

Table 3 and Table 4 show, the results obtained from sentiment classification models created using NB, kNN, LR, SVM, and all machine learning methods together in Stacking and Voting ensemble learning following TF-IDF and W2V text representations with 80%-20% and 70%-30% training/test sets on the Rotten Tomatoes dataset.

Table 3. Accuracy results of methods on Rotten Tomatoes Dataset (80:20 training and test sets)

	NB	kNN	LR	SVM	Stacking	Voting
TF-IDF	80.5	74.3	80.7	82.2	83.8	**84.5**
W2V	81.3	80.1	81.6	82.7	85.8	**86.5**

When the Table 3 is analyzed, it is seen that SVM has the best results in an accuracy score among the single machine learning methods with TF-IDF. However, SVM achieves the best Results with W2V. When the ensemble learning methods are compared against single machine Learning methods, it is seen that Stacking is behind SVM but Voting is ahead of SVM. Furthermore, Voting has demonstrated better results than stacking among ensemble learning methods.

Table 4. Accuracy results of methods on Rotten Tomatoes Dataset (70:30 training and test sets)

	NB	kNN	LR	SVM	Stacking	Voting
TF-IDF	81.5	76.8	81.7	83.8	83.9	**87.2**
W2V	81.6	80.6	81.9	87.8	88.5	**88.7**

When the Table 4 is analyzed, it is seen that SVM has the best results in an accuracy score among the single machine learning methods with TF-IDF. When the ensemble learning methods are compared against single machine learning methods, it is seen that both ensemble learning methods have demonstrated better performance than single machine learning methods. Furthermore, Voting has demonstrated better results than stacking among ensemble learning methods.

Table 5. F-measure results of methods on TripAdvisor Dataset (80:20 training and test sets)

	NB	kNN	LR	SVM	Stacking	Voting
TF-IDF	78.9	74.1	76.3	78.4	79.2	**79.5**
W2V	79.2	75.3	77.2	78	79.6	**80.1**

As can be seen in Table 5, Ensemble models gave better results. Although the results with the Community models were close to each other, Voting gave better results.

Table 6. F-measure results of methods on TripAdvisor Dataset (70:30 training and test sets)

	NB	kNN	LR	SVM	Stacking	Voting
TF-IDF	77.4	72.1	76.2	76.7	78.3	**78.6**
W2V	78.6	73.4	77.1	78.5	78.4	**79.4**

As seen in Table 6, Ensemble models gave better results than other machine learning models. Community models gave better results than Voting Stacking in itself.

Table 7. F-measure results of methods on Rotten Tomatoes Dataset (80:20 training and test sets)

	NB	kNN	LR	SVM	Stacking	Voting
TF-IDF	82.4	81.3	84.5	85.7	87.3	**87.5**
W2V	83.6	82.5	85.3	86.3	87.4	**87.8**

As seen in Table 7, the Voting Ensemble model gave the best results. It was seen that the models created after W2V gave better results than the models created with TF-IDF.

Table 8. F-measure results of methods on Rotten Tomatoes Dataset (70:30 training and test sets)

	NB	kNN	LR	SVM	Stacking	Voting
TF-IDF	82.5	81.2	83.2	84.6	85.3	**85.6**
W2V	83.4	80.1	84.2	85.4	86.2	**86.5**

As seen in Table 8, Ensemble models gave better results. Ensemble models gave better results than Voting Stacking in itself.

4.3. Literature comparison

The results, in which we compared the method with which we obtained the best results in the sentiment analysis study on the TripAdvisor dataset, and the other studies, are given in Table 9.

Table 9. Accuracy results comparison of models on Tridadvisor dataset

References	Method	ACC	F measure
25	TF-IDF	0.82	-
26	BOW	0.82	-
Our Voting Model	TF-IDF	**86.2**	**79.5**
	W2V	**87.3**	**80.1**

No study has been found in the literature on the open source Gervious's Rotten Tomatoes dataset [13]. The dataset was compared with other studies with a common source. These comparison results are given in Table 10.

Table 10. Accuracy results comparison of models on Rotten Tomatoes dataset

References	Method	ACC	F measure
27	n-gram	-	80.7
Our Voting Model	TF-IDF	**87.2**	**87.5**
	W2V	**88.7**	**87.8**

As can be seen in Table 9 and Table 10, even if the word representation methods are common, using machine learning algorithms together with ensemble models instead of using them alone increases the classification performance.

5. Discussion and conclusion

Sentiment classification study was carried out following the performance of text representation (TF-IDF) and word embedding methods (W2V) after text processing on public data sets of TripAdvisor and Rotten Tomatoes. Using the machine learning methods analyzed in this study together with ensemble learning algorithms instead of using them alone is the main contribution of the paper to the sentiment classification process. In this context, a sentiment classification study was carried out with the help of four different machine learning algorithms, namely NB, kNN, SVM, and LR, and two ensemble learning algorithms, namely Stacking and Voting. Training/test sets are used in the separation of 80%- 20% and 70%- 30% on both datasets in the holdout method. The experimental results were evaluated by measuring the accuracy and F-measure performance metric.

Instead of using single machine learning classifiers, using them together in ensemble learning methods has demonstrated better results. Voting, which is one of the ensemble methods, has been observed to yield better results in all experiments compared to Stacking.

When TF-IDF and W2V results of the experiments are evaluated, it is seen that W2V has outperformed TF-IDF. At 70%-30% and 80%-20% test-train separation, W2V gave better results than the TF-IDF.

Based on the experiments carried out as part of this study, it has been observed that more successful results are obtained in the models that are created using single classification algorithms together for ensemble learning. The performances of ensemble learning algorithms for larger data sets from different domains are aimed to be analyzed in future works.

We are planning to examine the performances of Bert in particular, FastText, Glove, Bert (even RoBerta, DistilBert etc., which are derivatives of Bert), together with single and ensemble ML methods, in future studies.

Author contributions

Fatih Kayaalp: Defining the methodology, evaluations of the results and draft editing **Muhammet Sinan Başarslan:** Preprocessing the dataset, data analysis, experiments and evaluations, manuscript draft preparation

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Mostafa, L. (2020). Machine learning-based sentiment analysis for analyzing the travelers reviews on Egyptian hotels. In Joint European-US Workshop on Applications of Invariance in Computer Vision. Springer, Cham, 405-413.
2. Dehkharghani, R., Yanikoglu, B., Tapucu, D., & Saygin, Y. (2012). Adaptation and Use of Subjectivity Lexicons for Domain Dependent Sentiment Classification. IEEE 12th International Conference on Data Mining Workshops, 10 December, Washington, 669–673.
3. Raut, V. B., & Londhe, D. D. (2014). Opinion Mining and Summarization of Hotel Reviews. International Conference on Computational Intelligence and Communication Networks, November, Bhopal, 556–559.
4. Tiwari, P., Mishra, B. K., Kumar, S., & Kumar, V. (2017). Implementation of n-gram methodology for rotten tomatoes review dataset sentiment analysis. International Journal of Knowledge Discovery in Bioinformatics (IJKDB), 7(1),30–41.
5. Zhou, Y. (2019). Sentiment Classification with Deep Neural Networks. Master's Thesis. Tampere University. Finland.
6. Sahu, T. P., & Ahuja, S. (2016). Sentiment analysis of movie reviews: A study on feature selection and classification algorithms. International Conference on Microelectronics, Computing, and Communications (MicroCom), 23-25 January, Durgapur, 1–6.
7. Oswin, H. R., Virginia, G., & Antonius, R. C. (2016). Sentiment Classification of Film Reviews Using IB1. 7th International Conference on Intelligent Systems, Modelling, and Simulation (ISMS), 23-25 January, Bangkok 78–82.
8. Mostafa, L. (2021). Egyptian Student Sentiment Analysis Using Word2vec During the Coronavirus (Covid-19) Pandemic. In: Hassanien A.E., Slowik A., Snášel V., El-Deeb H., Tolba F.M. (eds) Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020. AISI 2020. Advances in Intelligent Systems and Computing, vol 1261. Springer, Cham.
9. Machuca, C. R., Gallardo, C., & Toasa, R. M. (2021, February). Twitter sentiment analysis on coronavirus: Machine learning approach. In Journal of Physics: Conference Series (Vol. 1828, No. 1, p. 012104). IOP Publishing.
10. U. A. Siddiqua, T. Ahsan, & A. N. Chy, (2016). Combining a rule-based classifier with ensemble of feature sets and machine learning techniques for sentiment analysis on microblog. in 2016 19th International Conference on Computer and Information Technology (ICCIT), 2016, 304– 309.
11. Rahman, M., & Islam, M. N. (2022). Exploring the performance of ensemble machine learning classifiers for sentiment analysis of covid-19 tweets. In Sentimental Analysis and Deep Learning (pp. 383-396). Springer, Singapore.
12. Alam, M. H., Ryu, W. J., & Lee, S. (2016). Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. Information Sciences, 339, 206–223.
13. Gervais, N. (2019). Rotten Tomatoes Dataset. rotten-tomatoes-dataset (Access Date:21.02.2020).
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing systems 3111-3119.
15. Basarslan, M. S., & Kayaalp, F. (2020). Sentiment analysis with machine learning methods on social media. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 9(3),5-15.
16. Bakay, M. S., & Ağbulut, Ü. (2021). Electricity production-based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms. Journal of Cleaner Production, 285, 125324.
17. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1),21–27.
18. Basarslan, M. S., Bakir, H., & Yücedağ, İ. (2019, April). Fuzzy logic and correlation-based hybrid classification on hepatitis disease data set. In The International Conference on Artificial Intelligence and Applied Mathematics in Engineering (pp. 787-800). Springer, Cham.
19. Indulkar, Y., & Patil, A. (2021). Comparative Study of Machine Learning Algorithms for Twitter Sentiment Analysis. 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 295–299.
20. Zhou, Z. H. (2012). Ensemble methods: foundations and algorithms. CRC press.
21. Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and systems magazine, 6(3), 21-45.
22. Tao, F., Jiang, L., & Li, C. (2021). Differential evolution-based weighted soft majority voting for crowdsourcing. Engineering Applications of Artificial Intelligence, 106, 104474.
23. Battiti, R., & Colla, A. M. (1994). Democracy in neural nets: Voting schemes for classification. Neural Networks, 7(4), 691-707.
24. Canli, H., & Toklu, S. (2021). Deep Learning-Based Mobile Application Design for Smart Parking. IEEE Access, 9, 61171-61183.

25. Mahima, K. T. Y., Ginige, T. N. D. S., & De Zoysa, K. (2021). Evaluation of Sentiment Analysis based on AutoML and Traditional Approaches. Evaluation, 12(2).
26. Assyafah, H. B., Yulianti, D. T., & Kom, S. (2021). Analisis Dataset menggunakan Sentiment Analysis (Studi Kasus Pada Tripadvisor). Jurnal STRATEGI-Jurnal Maranatha, 3(2), 320-331.
27. Frangidis, P., Georgiou, K., Papadopoulos, S. (2020). Sentiment Analysis on Movie Scripts and Reviews. In: Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology, vol 583. Springer, Cham. https://doi.org/10.1007/978-3-030-49161-1_36