2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) | 979-8-3503-4215-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISMSIT58785.2023.10304923

Sentiment Analysis Using Machine Learning Techniques on IMDB Dataset

Selen Nazlı BAŞA
Dept. Computer Engineering
Istanbul Medeniyet University
Istanbul, Turkey
mrsryanssnb@gmail.com

Muhammet Sinan BAŞARSLAN
Dept. Computer Engineering
Istanbul Medeniyet University
Istanbul, Turkey
0000-0002-7996-9169

***Abstract*— Artificial intelligence is utilized in various sectors due to the surge in data. The fundamental element of artificial intelligence is data, which is sourced from various channels, including websites. It involves analyzing emotions, opinions, and attitudes expressed in text data.Websites provide significant volumes of comment data, which is used to conduct sentiment analysis studies. Sentiment analysis is a natural language processing (NLP) task, which is a core focus of artificial intelligence research. Sentiment analysis was conducted on open-source data obtained from IMDb, a platform containing information about films as well as reviews of actors, directors, and movies. The data was preprocessed before word vectorization was performed with the frequency-based TF-IDF method, and then machine learning algorithms were employed for classification. SVM was found to have the highest accuracy of 90%.**

Keywords—Machine learning, sentiment analysis, text representation.

I. INTRODUCTION

Social media plays a significant and essential role in our daily lives. Consequently, it is necessary to enhance our accomplishments in this field. This study aims to enhance our competence to provide a distinct outlook on the models used to work with word strings. Additionally, it reports the differences between stages with statistical evidence to guide future research. During the research, conventional machine learning techniques will be applied to the IMDb database, commonly utilized in the field of study.

Sentiment analysis goes beyond simply categorizing data as positive or negative; it can also identify complex emotions. Negative comments, for instance, may contain not only dissatisfaction but also elements of racism, cyberbullying, or even terrorism. It is essential to accurately separate such data from positive data. In the future, companies can utilize our proposed model to enhance customer relationships.

If high accuracy results are achieved by the study's end, the model can be used to detect and rapidly close social media accounts engaged in automatic account spamming, permanent account closures, and illegal activities targeting general audience users on such platforms and websites. Large corporations can integrate the model into their systems for more accurate feedback. The study's flow is provided in Figure 1.

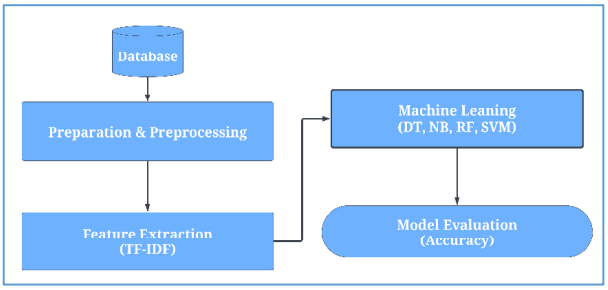

Fig. 1. Flowchart of the Study

Accordingly, the study is structured as follows: Section 2 presents the literature review. Section 3, methodology, describes the dataset and the preparatory phase for the classification of the text data. Section 4 describes the word representation method used in the study. Section 5 describes machine learning and the algorithms used in the study, while Section 7 describes the accuracy as a performance evaluation metric of the classification models built with these algorithms. Section 7 presents the experimental results. Finally, in Section 8, the study on IMDb is compared with previous studies and an overall evaluation of the study is presented.

II. LITERATURE REVIEW

Pang, Lee, and Vaithyanatham constructed a vector space model for pre-classifying IMDb reviews and conducted sentiment analysis using classification algorithms such as Naive Bayes (NB), Maximum Entropy, and Support Vector Machine (SVM). Among these, the SVM algorithm achieved the highest accuracy of 82.9% with unigrams [1]. Nikfarjam and Azadeh examined comments on medication that they collected from Twitter to conduct a study on sentiment analysis. They reported that the SVM algorithm outperformed other methods with an 82.1% accuracy [2].

There are several studies conducted on Turkish datasets. Nizam and Akın investigated the impact of data distribution in classes on the classification algorithm's success rate and found it to be significant. They achieved an accuracy rate of 72.33% using the SVM algorithm in their food sector study [3].

In Sevindi conducted a study on Turkish film reviews in which several classification algorithms were employed to calculate emotional classes, including decision tree, k-nearest neighbor (KNN), NB, and SVM. The results showed that SVM performed the best [4]. Kaynar and Yıldız used NB, multilayer

979-8-3503-4215-4/23/$31.00 ©2023 IEEE

artificial neural networks (MLP), and SVM machine algorithms, along with TF-IDF for feature extraction, in their own movie interpretation study. At the conclusion of the study, it was observed that the artificial neural network outperformed with an 89.73% accuracy score [5].

Shaukat et. al., achieved an ACC performance of 86.67% using a multilayer perceptron model after employing Bag of Words (BoW) approach [6]. Mohaiminul and Sultana developed models using Multi Nominal Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), Support Vector Machines (SVM), Random Forest (RF), and Stochastic Gradient Descent (SGD) algorithms after applying Term Frequency-Inverse Document Frequency (TF-IDF). RF algorithm resulted in the highest accuracy of 83.66% [7]. Yenter and Verma obtained 89% accuracy in classification using TF-IDF [8]. Amulya et. al., achieved an accuracy rate of 88% using RNN after TF-IDF [9]. Misini et al. achieved the highest accuracy rate of 86.67% with MLP following BoW [10]. Yin et. al., obtained an accuracy rate of 87.90% with CNN using Keras embedding representation [11]. Basarslan and Kayaalp achieved an accuracy rate of 88.21% with BiLSTM using Keras embedding representation [12].

Rathee et. al., utilized several machine learning algorithms, including DT, KNN, LR, RF, and SVM following BoW. Among these, SVM yielded the highest accuracy of 76% [13]. Wang and Manning achieved an accuracy of 89.16% using SVM after bigram analysis [14]. Similarly, Narayanan and team obtained an accuracy of 88.80% using NB after ngram analysis [15].

III. METHODOLOGY

In this section, the dataset used in the study, preprocessing processes, machine learning algorithms and the criteria used in the performance evaluation of the models are explained.

A. Dataset

The The IMDb Movie Reviews dataset is widely used by AI researchers. With over 50,000 movie reviews, it includes 25,000 positive and 25,000 negative labeled reviews, both of which have an equal class distribution. IMDb [16], founded in 1990 by Col Needham, comprises a summary, genre, rating, and reviews of movies or TV series. Additionally, it provides actor biographies and information on their production histories. In the study dataset, solely the movie review feature was extracted from the website, which comprises 50,000 movie reviews and has a total of two features. Please see Table I for the features' descriptions.

TABLE I. IMDB MOVIE REVIEWS FEATURES

Feature	Description
Review	Texts containing opinions about the movie
Sentiment	Indicates the positive, negative or neutral emotion that the opinion wants to convey

Preparation procedures were also applied to the IMDb dataset and no characteristics were eliminated as both convey significant information and are essential components of the dataset.

Fig. 2. Words often used in IMDb Movie Reviews

Figure 2 shows the word cloud that occurs frequently in the dataset.

B. Preparation Stages

The preparation stages aimed to provide information about the processes that the dataset went through before sentiment analysis.

1) Stopwords

English [17], which started to be used as an inflected language from the Indo-European family of languages, contains a variety of words that do not add a great deal of meaning to the sentence, but rather are used to emphasize features such as place-direction, uniqueness-multiplicity. For example: 'a', 'the', 'nor'...

When we extract stopwords, we use a dictionary called stopwords corpus. The corpus shown in Figure 3 is a collection of stop words that can be thought of in the English language.

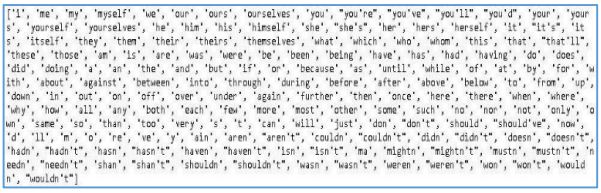

Fig. 3. Stopwords Corpus [18]

2) Stemming

Since English is an 'inflected' language, stemming is applied. In fact, we can say that stemming is the process of reducing the word to its root. Figure 4 shows example words derived from a word.

Fig. 4. Stemming [19]

As shown in Figure 4, the words "connected," "connect," and "connecting" stem from the root word "connect." By applying stemming, variations of the word can be removed, reducing the amount of data processing required for word processing algorithms.

3) Removal of Symbolic Markers

Since the study focuses on algorithmic differences, data cleaning was done through general steps. Factors that could be seen as 'noise' were removed during cleaning instead of

performing a detailed cleaning. However, incorporating more detailed steps during cleaning could positively impact the results.

IV. Feature Extraction Methods

NLP research does not analyze texts directly but rather utilizes various vectorization methods to extract meaning from textual data.

A. TF-IDF

Term frequency (TF) refers to the frequency with which term it is found in the document. It is calculated by dividing the number of occurrences of that term in the document by the total number of terms in the document, as shown in Equation (1).

$$TF\ (t,d) = \frac{number\ of\ t\ terms\ in\ d\ documents}{Total\ number\ of\ terms\ in\ document\ d} \quad (1)$$

IDF (inverse document frequency) is used to adjust the importance of term t in a document. This is necessary as DF (document frequency) is distorted due to the repeated use of stop words such as 'the' and similar words, resulting in false results about the significance of the term. To avoid this, the IDF reduces the importance of unimportant words used frequently in the document. Equation (2) [20] calculates the IDF.

$$IDF\ (t) = \log(\frac{N}{\text{IDF(t)}}) \quad (2)$$

The corpus, denoting the amount of words available for use in the study, is represented as N in equation (2). By applying the equations for term frequency (TF) and inverse document frequency (IDF) in conjunction, we can determine the significance of a given term, t, within the corpus for a given document, d. The calculation for TF-IDF is provided in Equation (3) [21].

$$TF - IDF = \text{TF(t, d)} * \text{IDF (t)} \quad (3)$$

V. Machine Learning

Machine learning pertains to algorithms that are statistically based and applied to construct predictive models using existing data. These algorithms are typically successful and have prompted the emergence and evolution of numerous machine learning models [21].

A. Decision Tree

The dataset's input space is divided into 'regions' by the Decision Tree (DT), which learns multiple decision rules. The classification process occurs in this manner [23]. Observe the DT Algorithm visualization illustrated in Figure 5.

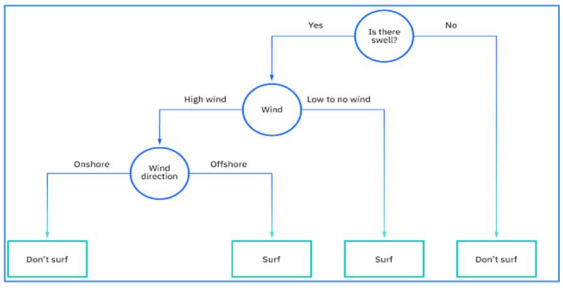

Fig. 5. Decision Tree Algorithm [23]

B. Random Forest

Random Forest (RF) is a machine learning algorithm that employs numerous decision trees on distinct sample sets to raise its accuracy and combat overfitting concerns. Figure 6 illustrates a graphic representation of the RF algorithm.

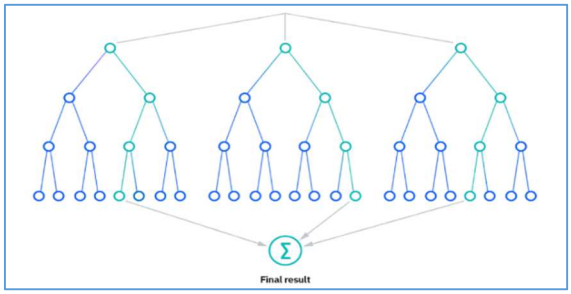

Fig. 6. Random Forest Algorithm [23]

C. Naïve Bayes

Derived in 1812 by Thomas Bayes with a simplistic method utilizing the conditional probability theorem (Bayes' Theorem), this algorithm computes the probability of each class and assigns it to the class with the highest probability [24].

D. Support Vector Machine

The Support Vector Machine (SVM) algorithm endeavors to create a hyperplane equidistant to both classes. It relies on the extreme examples from each class, termed Support Vectors, to establish this hyperplane. SVM's primary strength lies in its ability to generalize data effectively, as it learns via margin [24]. Refer to Figure 7 for a visualization of the SVM Algorithm.

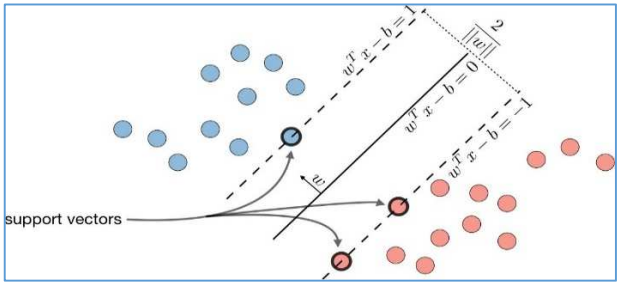

Fig. 7. SVM Algorithm [24]

VI. Performance Metric

Since the IMDb film review dataset is balanced, the accuracy score suffices as an evaluation metric. The Confusion Matrix table in Table II displays actual and predicted values for a classification problem [24].

TABLE II. Confusion Matrix

		Actual value		
		Positive	**Negative**	**Total**
Estimated value	**Positive**	True Positive (TP)	False Positive (FP)	TPos
	Negative	False Negative (FN)	True Negative (TN)	TNeg
	Total	Pos	Neg	M

Accuracy is provided in Equation 4.

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + F_N + T_N} \quad (4)$$

VII. Experimental Results

After conducting a study on the IMDb dataset, this section presents the experimental results obtained. The dataset was divided into an 80%-20% hold-out separation for training and testing. Sentiment analysis models were generated using four different machine learning methods, which utilized TF-IDF for word representation extraction. The machine learning methods applied were DT, NB, SVM, and RF. The performance outcomes for each method are provided in Table III.

TABLE III. Accuracy Values of ML Algorithms on IMDb Movie Reviews Dataset

DT	RF	NB	SVM
0.72	0.86	0.86	0.90

According to Table III, the model produced by SVM using TF-IDF on the IMDb Movie Reviews dataset achieved the highest accuracy score. While SVM performed the best, NB and RF had comparable results and were tied for second place, while DT was the least accurate.

VIII. Results and Discussion

In this study, we utilized TF-IDF to extract word representations after splitting the IMDb dataset into 80%-20% training and testing sets. We then applied machine learning methods for sentiment analysis and tested the resulting models. Our models employed DT, NB, SVM, and RF machine learning methods, with SVM showing the most promising outcome. NB and RF performed similarly well to SVM. Table IV presents a comparison of our study's best result with those of previous IMDb dataset studies.

TABLE IV. Previous Studies on the IMDB Dataset.

Models	Text Representation	Models	Accuracy (%)
[1]	Unigram	SVM	82.9
[2]	Conditional random fields (CRF) based on their own proposed text representation	SVM	82.1
[3]	None	SVM	72.33
[4]	Chi-square with BoW	SVM	77
[5]	TF-IDF	ANN	89.73
[6]	BoW	SVM	86.67
[7]	TF-IDF	RF	83.66
[8]	TF-IDF	CNN-LSTM	89
[9]	TF-IDF	RNN	88
[10]	BoW	MLP	86.67
[11]	Keras embedding	CNN	87.90
[12]	Keras embedding	BiLSTM	88.21
[13]	BoW	SVM	76
[14]	Bigram	SVM	89.16
[15]	N-gram	NB	88.80
The present	TF-IDF	SVM	**90**

Models for sentiment analysis were created within the scope of this study, using TF-IDF for text representation. Table 3 demonstrates that the sentiment analysis study, utilizing the traditional frequency-based method like TF-IDF, is competitive with the existing literature. It is important not to overlook TF-IDF when working with or prior to embarking on embedding methods, such as Word2Vec, GloVe, or transfer learning-based methods like BERT and XLNET. In future research, there are plans to create hybrid models using transfer learning techniques for text representation.

References

[1] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques," Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[2] Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., & Gonzalez, G. "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," Journal of the American Medical Informatics Association ,2015, pp ocu041.

[3] Nizam, Hatice, and Saliha Sıla Akın. "Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması," XIX. Türkiye'de İnternet Konferansı,2014.

[4] Türkmen, Ali Caner, and Ali Taylan Cemgil. "Political interest and tendency prediction from microblog data," 2014 22nd Signal Processing and Communications Applications Conference (SIU), IEEE, 2014.

[5] Kaynar, O., Görmez, Y., Yıldız, M., Albayrak, A. "Makine öğrenmesi yöntemleri ile Duygu Analizi." In International Artificial Intelligence and Data Processing Symposium (IDAP'16), 2016.

[6] Z. Shaukat, A. A. Zulfiqar, C. Xiao, M. Azeem, and T. Mahmood, "Sentiment analysis on IMDB using lexicon and neural networks," SN Appl. Sci., vol. 2, no. 2, p. 148, Feb. 2020, doi: 10.1007/s42452-019-1926-x.

[7] M. Mohaiminul and N. Sultana, "Comparative Study on Machine Learning Algorithms for Sentiment Classification," Int. J. Comput. Appl., vol. 182, no. 21, pp. 1–7, 2018, doi: 10.5120/ijca2018917961.

[8] Alec Yenter and Abhishek Verma, Deep CNN-LSTM with Combined Kernels from Multiple Branches for IMDb Review Sentiment Analysis, 2017.

[9] K. Amulya, S. B. Swathi, P. Kamakshi, and Y. Bhavani, "Sentiment Analysis on IMDB Movie Reviews using Machine Learning and Deep Learning Algorithms," in 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Jan. 2022, pp. 814–819, doi: 10.1109/ICSSIT53264.2022.9716550.

[10] Misini, A., Kadriu, A., & Canhasi, E. (2023, June). Albanian Authorship Attribution Model. In 2023 12th Mediterranean Conference on Embedded Computing (MECO) (pp. 1-5). IEEE.

[11] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative Study of CNN and RNN for Natural Language Processing," 2017, [Online]. Available: http://arxiv.org/abs/1702.01923.

[12] M. S. Basarslan and F. Kayaalp, "Sentiment Analysis with Various Deep Learning Models on Movie Reviews," 2022 International Conference on Artificial Intelligence of Things (ICAIoT), Istanbul, Turkey, 2022, pp. 1-5, doi: 10.1109/ICAIoT57170.2022.10121745.

[13] N. Rathee, N. Joshi and J. Kaur, "Sentiment analysis using machine learning techniques on Python," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 779-785

[14] S. Wang & C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012- Proceedings of the Conference, vol. 2, no. 2, pp. 90–94, 2012.

[15] Narayanan, V., Arora, I., Bhatia, A. (2013). Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model. In: Yin, H., et al. Intelligent Data Engineering and Automated Learning – IDEAL 2013. IDEAL 2013. Lecture Notes in Computer Science, vol 8206. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41278-3_24

[16] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” ACL-HLT 2011 - Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol., vol. 1, pp. 142–150, 2011

[17] F. Bal and F. Kayaalp, "A Novel Deep Learning-Based Hybrid Method for the Determination of Productivity of Agricultural Products: Apple Case Study," in IEEE Access, vol. 11, pp. 7808-7821, 2023, doi: 10.1109/ACCESS.2023.3238570.

[18] M. S. Başarslan and F. Kayaalp , "Sentiment analysis with ensemble and machine learning methods in multi-domain datasets", Turkish Journal of Engineering, vol. 7, no. 2, pp. 141-148, Apr. 2023, doi:10.31127/tuje.1079698

[19] Z. Turgut, G Akgun, “Occupancy detection for energy efficiency of smart buildings using machine learning,” 19th International Conference on Sustainable Energy Technologies,2022.

[20] S. Gulmez, A. G. Kakisim and I. Sogukpinar, "Analysis of the Dynamic Features on Ransomware Detection Using Deep Learning-based Methods," 2023 11th International Symposium on Digital Forensics and Security (ISDFS), Chattanooga, TN, USA, 2023, pp. 1-6, doi: 10.1109/ISDFS58141.2023.10131862.

[21] F. Kayaalp, M. S. Basarslan and K. Polat, "A hybrid classification example in describing chronic kidney disease," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, 2018, pp. 1-4, doi: 10.1109/EBBT.2018.8391444.

[22] Kakisim, A.G. Enhancing attributed network embedding via enriched attribute representations. Appl Intell 52, 1566–1580 (2022). https://doi.org/10.1007/s10489-021-02498-w

[23] Kakisim, A. G., Gulmez, S., & Sogukpinar, I. (2022). Sequential opcode embedding-based malware detection method. Computers & Electrical Engineering, 98, 107703.

[24] Millionschik, M., Cohen-Nissan, A., Rousso, R., & Hamo, Y. 2022 IEEE Signal Processing Cup: Synthetic Speech Attribution.