International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V71I5P203

Volume 71 Issue 5, 18-26, May 2023
© 2023 Seventh Sense Research Group®

Original Article

Utilizing Machine Learning for Sentiment Analysis of IMDB Movie Review Data

Ubaid Mohamed Dahir[1], Faisal Kevin Alkindy[2]

[1]Faculty of Computing, Simad University, Mogadishu, Somalia.
[2]School of computer science, University Science Malaysia, Penang, Malaysia.

[1]Corresponding Author : engubaid@simad.edu.so

Received: 25 March 2023 Revised: 13 May 2023 Accepted: 18 May 2023 Published: 25 May 2023

Abstract *- In this study, we focus on sentiment analysis, an essential technique in the rapidly evolving field of text analytics. Ourapproach involves preprocessing the movie review text data using tokenization, lemmatization techniques, and feature extraction using Word of Bags and TF-IDF. We employ three popular machine learning methods, Logistic Regression, SVM, and Random Forest, to develop sentiment classification models. Our results show that logistic regression with the TF-IDF technique and default parameters outperforms the other models in terms of minimizing false positives, with an accuracy of 89.20%, a precision of 88.80%, recall of 89.80%, and an area under the receiver operating characteristics curve (AUC) of 89%. These findings have important implications for improving sentiment analysis and developing more accurate and effective text analytics tools, contributing to the novelty of the work in the journal fields.*

Keywords *- Bag of Words, Logistic regression, Movie review, Precision, Random forest, Sentiment analysis, SVM, TF-IDF.*

1. Introduction

The internet plays a vital role in establishing data connectivity between users and producers in this era of digitization. Global internet users are expected to reach 5.3 billion by 2023 [1]. The ease of having an internet connection also fosters the rapid and massive growth of numerous forms of data. Thus, the world has entered the era of big data. Such is the case in the film industry. Manypeople use social media platforms to express their sentiments or ideas about the movies they see and share them with other users, with text being one of the most common formats. Consumers utilize movie reviews to decide whether a film is worth the ticket price. Moreover, film production firms can use them as a marketing tooland a predictor of a film's financial success [2].

Text analytics is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform unstructured text data into normalized data suitable for analysis by using machine learning (ML) algorithms. One of the text analytics techniques, sentimentanalysis, can help reveal insights from movie review data. Sentiment analysis, or opinion mining, utilizes computer systems to identify, extract, and classify opinions expressed in vast amounts of text data. There are three types of sentiment analysis classification levels: document, sentence, and attribute [3]. The whole document will be used to analyze and figure out the sentiment at the document level. Then for the sentence level, it analyses the sentiment from each sentence, whether positive,negative, or neutral. The third level, the sentiment analysis, is done for both the document and sentencelevels. The attribute level is the most exact type of sentiment analysis, usually performed in review, comment and feedback.

Pang and Lee [4] made significant contributions that revolutionized the realm of opinion mining and sentiment analysis. Their groundbreaking work focused on classifying documents based on sentiment, particularly emphasizing movie reviews. This pivotal research laid the groundwork for subsequent investigations and sparked a wave of exploration into sentiment analysis techniques.

Taking inspiration from Pang and Lee's pioneering work, Maas et al. [5] put forth an innovative approach incorporating word vectors into sentiment analysis. Their study showcased the effectiveness of this technique in capturing word meanings and enhancing sentiment classification accuracy. This breakthrough introduced fresh possibilities by considering the contextual semantics of words. Expanding upon these advancements, Socher et al. [6] introduced deep recursive models that embraced the hierarchical structure of sentences. Their approach adeptly captured nuanced sentiment information at multiple levels, unveiling a more intricate understanding of sentiment analysis. This enabled researchers to delve deep into the subtleties of sentiment expression.

Kim [7]made significant strides in sentiment analysis by harnessing the formidable power of convolutional neural networks (CNNs). CNNs emerged as potent tools for capturing intricate sentence patterns through their work, leading to an unparalleled performance in sentiment classification tasks. This breakthrough unequivocally demonstrated the potential of deep learning models in extracting sentiment features.

In addition to analyzing sentiment at the word level, Zhang et al. [8] delved into the realm of character-level convolutional networks for sentiment analysis. Their research illuminated the value of considering character-level information, which helped unravel subtle nuances in sentiment expression. This approach seamlessly complemented traditional word-level analysis and elevated sentiment classification accuracy. Tang et al. [26] directed their focus towards capturing sequential dependencies within the text by introducing gated recurrent neural networks (RNNs) for sentiment analysis. By accounting for temporal dynamics and contextual cues in sentiment expression, their approach yielded improved sentiment analysis outcomes. This underscored the indispensability of contextual information in comprehending sentiments.

Moreover, Severyn and Moschitti [10] expanded the horizon of sentiment analysis to encompass social media data, particularly Twitter, by employing deep convolutional neural networks (DCNNs). Their research showcased the adaptability of DCNNs in extracting sentimental information from short, noisy texts. This served as a testament to the versatility of deep learning models across diverse data domains. Amrani et al. [11] proposed sentiment analysis by employing a hybrid approach to identify Amazon product reviews. They trained the data individually using random forest, support vector machine, and a combination ofthese models.

Based on Accuracy, Accuracy, Recall and F-measure, the hybrid technique comprised of Random Forest and Support Vector Machine yielded excellent results, with precision, recall, and F-measure values of 83.4%, 83.4%, and 83.4%, re- respectively. The Random Forest strategy improved performance in the case of small reviews, while the Support Vector Machine approach improved performance only in the instance of extensive reviews.

Another sentiment analysis study of movie review data was done by [12]. The study compared the effectiveness of the Naive Bayes and decision tree classification models. The sentiment analysis used IMDB movie review data of 250 positive and 250 negative reviews. Various assessment metrics were used to compare the model's performance, including the confusion matrix, precision, recall, f-measure, ROCcurve, and Area under the curve (AUC). The outcome revealed that the Naive Bayes classifier performed better, with an accuracy of 97%, but the accuracy of the decision tree was just 65%.

While several studies have been conducted on sentiment analysis using machine learning algorithms, there is still a research gap in understanding the effectiveness of different algorithms for sentiment analysis in the context of the film industry. Furthermore, there is a need for more accurate and efficient sentiment analysis tools to help film production firms understand audience sentiments towards movies, whichcan be helpful for marketing purposes and as a predictor of a film's financial success.

The sentiment analysis uses natural language processing and text analytics to determine whether the writer has a positive, negative, or neutral opinion about a specific issue [3], [12]. Therefore, the primary objective of this work is to predict audience sentiments using IMDB movie review data and several machine learning algorithms. Logistic Regression, Random Forest, and Support Vector Machine (SVM) classification models were developed to predict audience sentiment.

The rest of the paper is organized as follows. The problem is presented in Section 2, and our proposed solution is presented in Section 3. Evaluation criteria are presented in Section 4. Analysis and discussion are covered in Section 5. Finally, we conclude our work in Section 6.

2. Problem Statement

Today, data is highly beneficial for businesses to convert social media behavior into valuable business data. As with movie reviews, film production companies can comprehend how their films affect audiences. Themovie reviews of the audience can be used to predict the trend of movies. For a film production company to be competitive in the marketplace, the film production company needs to understand what type of film the audience prefers. However, sometimes the opinion or comment considered positive in one situation is also considered negative in another. In addition, each audience has unique ways of expressing their opinions, resulting in various comment styles.

About those issues, this project proposes two problems that must be addressed.

Movie review data is unstructured and still raw. The machine learning algorithm cannot understand it.Hence it needs to be preprocessed so that the algorithm can extract the feature information well.

1) Manually extracting information from huge movie review data is tough. Analyzing sentiment manually wastes time and is sensitive to subjectivity. Therefore, it requires technology capable of automating sentiment analysis.

3. Proposed Solution

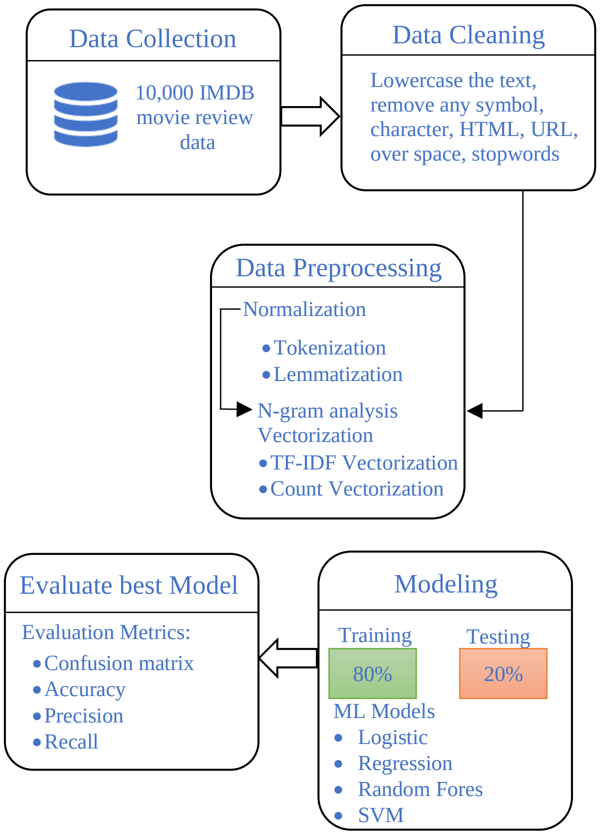


Fig. 1 Project framework

This project implements the CRISP-DM (CRoss Industry Standard Process for Data Mining) technique. The CRISP-DM process model aims to reduce the cost, accuracy, reproducibility, manageability, and speed of large data mining operations [13]. The CRISP-DM provides the six phases of the life cycle of a data mining project: business understanding, data understanding, data preparation, modelling, assessment, and deployment. However, due to the purpose of this project, the CRISP-DM model will only be implemented until the evaluation stage.

In this paper, there are two objectives:

1) To perform data preprocessing to extract review data information for machine learning model input.
2) To develop machine learning models for sentiment classification of movie review data.

The proposed solution is represented in the project framework in Figure 2 to achieve the objectives above. The sentiment analysis process was started by collecting the dataset. The dataset was an IMDB movie review that was collected from Kaggle. The dataset consists of two columns, the review column as the feature and the sentiment column as a target. The review features the audience's response to the movie, whilst the sentiment column contains the review's overall sentiment. Initially, the dataset had 50,000 movie reviews. However, due to the limitations of the machine, the original data was subset into 10000 moviereviews with a sentiment ratio of 50% positive and 50% negative reviews. The next stage is data cleaning. The review data needs to be cleaned to be preprocessed for the data preprocessing stage.

The cleaned data were normalized in the preprocessing stage using tokenization and lemmatization. Then n-gram analysis was conducted to see the contiguous n-words extracted from a text sequence. Afterwards, vectorization was done to convert the text data to numerical vectors so the machine could understand the data. Three machine learning models were used at the modelling stage: Logistic Regression, Random Forest, and SVM. After the training phase, the trained models were tested. After that, all models were evaluated with the evaluation metrics. If the models could not perform well, hyperparameter tuning would be employed. After that, the best model was chosen [14].

3.1. Data Cleaning

Data cleaning is every machine learning task's initial and most crucial stage. It is the process of converting unstructured data into standardized data that can be analyzed. Concerning text data, especially review data, there are several characters that the machine learningmodel may not require; therefore, it is crucial to clean these data into a machine-readable format. The Panda's Python library and Regex were used to clean the data in this project. Regex is a string containing a pattern that can match words corresponding to pattern b20. Therefore, Regex can remove words based on patterns.

The following describes the data cleansing procedure [15].

3.1.1. Lowercase the Text

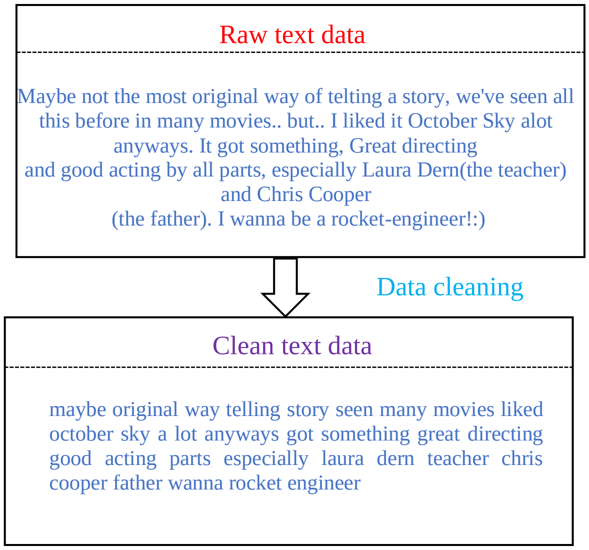
The review data consists of capital and lowercase characters. It is essential to lowercase all the characters, so there are no problems removing stop words since every character in stopwords is lowercase.

3.1.2. Remove Unicode Characters

Unicode is an international standard for character encoding that assigns a unique number to every character, regardless of language or script. Some reviews may contain emojis and other non-ASCII characters that are unreadable in ASCII format. Therefore, it must be removed from the data.

3.1.3. Remove HTM and URL

HTML and URL existed in the review data. *Beautiful Sop* Library was used to remove HTML. At the same time, the URL can be removed by applying Regex.



Maybe not the most original way of telting a story, we've seen all this before in many movies.. but.. I liked it October Sky alot anyways. It got something, Great directing and good acting by all parts, especially Laura Dern(the teacher) and Chris Cooper (the father). I wanna be a rocket-engineer!:)

maybe original way telling story seen many movies liked october sky a lot anyways got something great directing good acting parts especially laura dern teacher chris cooper father wanna rocket engineer

Fig. 2 The Comparison between raw text data and cleaned text data

3.1.4. Remove mention, hashtag, punctuation, and number

Mention, hashtag, and number could be present in the review data. While punctuation is typically used to compose a sentence, it can also emphasize a word or phrase. Therefore, each of these terms must be removed.

3.1.5. Space

During data sampling, some reviews have double spacing. Therefore, redundant spacing must be eliminated. In contrast, some reviews do not include a space after a stop or comma. In this situation, a space must be added so that there will be a space after the punctuation is removed.

3.1.6. Remove Stopwords

Stop words, such as" i"," me"," my," myself", etc., are typical words that often appear to be of little help in selecting materials that match a user's need [27]. NLTK library was utilized to remove stop words. In this situation, the stop words are set to English [17].

3.2. Text Normalization

3.2.1. Tokenization

Tokenization separates or splits textual data into smaller, meaningful components known as tokens [18], [19]. A text document comprises multiple components, including sentences that can be split into clauses, phrases, and individual words. In this study, a sentence was split into a list of words that can be used to reconstruct the text using the word tokenization technique. It is crucial to the text analytics process, especially for stemming and lemmatization procedures, which act on each word based on its stem and lemma. WhitespaceTokenizer from the nltk library was used to perform tokenization; it is a tokenizer that solely separates the words based on the white space characters.

3.2.2. Lemmatization

A lemma or root word is the standard form of a lexeme, where lexeme refers to the set of all forms with the same meaning, and lemma refers to the form chosen to represent the lexeme [19]. Lemmatization refers to locating a particular word's lemma or dictionary form. Lemmatization reduces a word's inflectional and occasionally derivationally related forms to its base form. The nltk library provides a robust lemmatization module that uses WordNet.

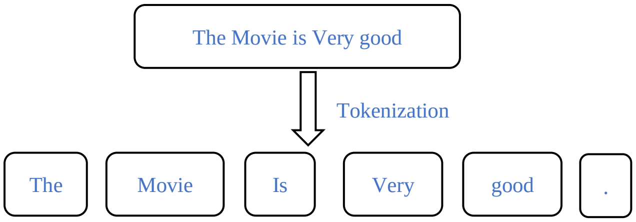


Fig. 3 Tokenization process

And the word's syntax and semantics, such as part of speech and context, to determine the root word or lemma [20].

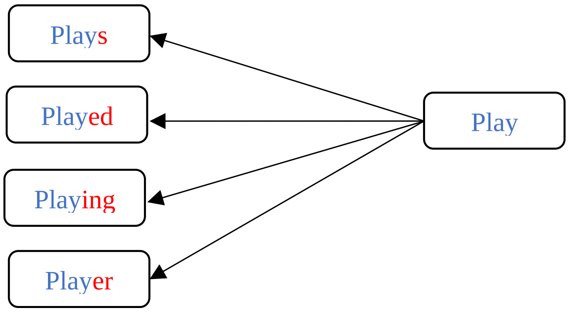


Fig. 4 Lemmatization process

3.2.3. N-Gram Analysis

N-grams are contiguous sequences of n-words extracted from a text sequence [19]. N-grams of sizes 1, 2, and 3 are called unigrams, bigrams, and trigrams, respectively. N-gram analysis is an essential pillar that is the foundation for constructing a standard bag of words model with relevant data. This analysis considers which words are most frequently seen in conjunction with other words in the dataset. This project defines unigram in the n-gram range for vectorization techniques. N-gram size will be added if the model cannot generalize properly.

3.2.4. Feature Extraction

Machine learning algorithms assume features in numeric vectors when attempting to discover patterns from data because the algorithm is essentially a mathematical process for optimizing and minimizing loss and error [19]. It is necessary to convert textual data to numeric vector space while working with it. The Vector Space Model is a mathematical model used to transform and present text data in the dimension form of numeric vectors. This project employed two feature extractiontechniques, which will be discussed below.

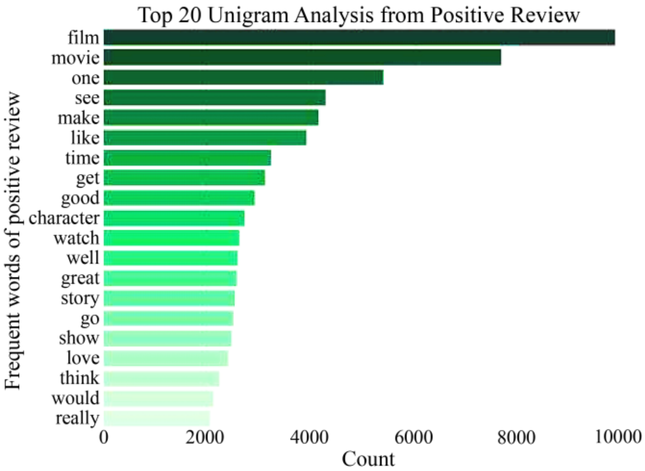


Fig. 5 Top 20 unigram analyses from a positive review

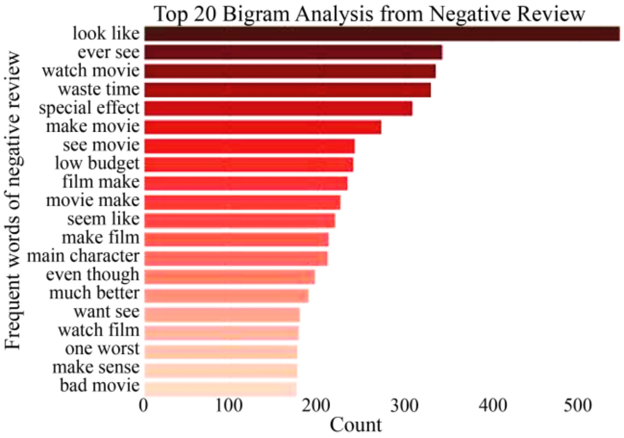


Fig. 6 Top 20 bigram analysis from a negative review

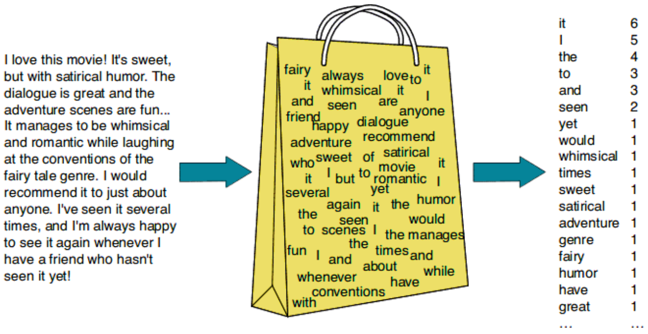
Bag of Words

The bag of words is a simple and effective technique for extracting text data into numerical vectors representing the frequency of each unique word in a document and sorting them in descending order [19], [21]. This technique can be performed by utilizing CountVectorizer to extract data as a feature that counts the occurrence of each word [28]. The Bag of Words method also allows anN-gram range parameter to accommodate n-grams as features.

TD-IDF

Term Frequency- I Inverse Document Frequency (TF-IDF) is a feature extraction techniquethat seeks to better identify the importance of a word inside a document by considering its relationship to other documents within the same corpus [19], [23]. This method examines the frequency at which a word appears in a document and the frequency at which the word appears in other documents in the corpus. The formula for TF-IDF is defined as follows:

$$tf\ idf\ (w, d, D) = tf\ (w, d) * idf\ (w, D) \qquad (1)$$


I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Fig. 7 The illustration of the Bag of Words process [14]

Here, f(w,d) is the frequency of the word w in document d, and f(w, D) is the word w in document D.

3.2.5. Machine Learning Algorithms

Logistic Regression

Logistic regression predicts the likelihood of an occurrence based on a given set of independent variables. The dependent variable is constrained between 0 and 1 because the outcome is a probability. Using the maximum likelihood, logistic regression fits a logistic function with an S-shaped curve. The formula for the logistic function is presented here.

The Logistic Regression model can show the association between the target variable and features in sentiment analysis. In addition, it is simple to implement, interpret, and cost-effective.

Random Forest

Random Forest is a nonparametric algorithm capable of processing linear and nonlinear data. It uses ensemble learning to enhance accuracy by combining many models to solve problems [24]. Random forest predicts a new instance formed by trees with little correlation. Several decision trees decision tree is generated using bootstrapped training data. Most accessible attributes are not considered at each decision tree split in a random forest. Decorating the trees reduces the standard deviation of the produced trees' means, boosting their effectiveness. Furthermore, this method is resistant to outliers. [27].

Support Vector Machine (SVM)

The SVM technique is one of the most common and effective machine learning algorithms [25]. SVM is used to map the input variable to an n-dimensional feature space. The SVM transforms data from a low- dimensional space to a higher-dimensional space. To prevent overfitting, SVM constructs a hyperplane that divides the feature space by class given labelled training data. SVM calculates the association between high-dimensional data without converting it, saving time and money.

3.3. Evaluation Criteria

The effectiveness of the proposed framework for sentiment analysis is determined by computing several evaluation metrics used to evaluate the classifier's performance. The proposed evaluation metrics are as follows [19],[27].

3.3.1. Confusion Matrix

A confusion matrix is the best way to summarize the performance of classifiers. It can reveal the strengths and shortcomings of a model. The column of the confusion matrix depicts the total number of observations in thepredicted class, whereas the row displays the total number of observations in the actual class. On the sloping side of the Table, false-negatives and false positives are shown.

3.3.2. Accuracy

Accuracy is the number of sentiment types properly classified. Because the class labels in this dataset are balanced, accuracy is suitable.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (2)$$

3.3.3. Precision

It refers to the proportion of correctly classified positive samples relative to the number of predicted positive samples. A higher precision indicates fewer false positives, whereas a lower precision indicates more false positives. The formula can depict precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (3)$$

3.3.4. Recall

The proportion of Positive samples correctly identified as Positive relative to the total number of Positive samples. A higher recall indicates fewer false negatives, while a lower recall indicates more. The formula can depict recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (4)$$

3.3.5. ROC-AUC

The ROC curve is a line graph showing a False Positive Rate on the x-axis and a True Positive Rate on the y-axis between 0 and 1. A good model's curve will be located at the upper left corner of the plot. The AUC is the Area Under the ROC Curve, which assesses the probability that the model would randomly select a positive class over a negative class. An AUC value of 1 reflects the perfect model.

4. Analysis and Discussion

In this paper, three machine learning methods were chosen to develop sentiment classification models for a movie review data set. Multiple experiments were conducted on the dataset to evaluate the efficacy of the proposed models. The experiment used a MacBook Pro 13" with a Jupyter Lab platform, utilizing an Apple M1 processor with an 8-core CPU, 8-core GPU, 16-core Neural Engine, 16GB of RAM, and a 512GB solid-state drive. The proposed work utilized Kera's package with Python 3.8 and TensorFlow as the backend. Each machine learning model was trained with two feature extraction techniques, Bag of Words and TF-IDF. The first experiment involved training the models with the default parameters, while the second experiment utilized Random Search CV for hyperparameter tuning. Detailed explanations of each experiment are presented below.

4.1. Performance of Models with Default Parameter

The performance of models using the Bag of Words and TF-IDF vectorization techniques with default parameter values was compared regarding evaluation metrics (Table 1). The results showed that all machine learning models performed well in classifying the positive and negative sentiment of the movie review dataset. Generally, the TF-IDF-based models outperformed the Bag of Words techniques. Logistic regression with TF-IDF vectorization had the highest accuracy and precision, corresponding to 89.200% and 88.800%, respectively. SVM with TF-IDF vectorization has the highest recall of 90.500%. The AUC score was identical for both models, with a value of 89%. While Random Forest had the lowest performance compared to logistic regression and SVM, it did not overfit the data.

Tables 2 and 3 illustrate the confusion matrix of logistic regression and SVM utilizing TF-IDF as the default parameter. SVM is less effective than the logistic regression—model at minimizing false positives. In sentiment analysis, the negative sentiment holds more significance, and it helps film production studios identify areas for improvement. Since logistic regression with TF-IDF showed higher accuracy and precision than SVM with TFIDF, it was selected as the optimal model for sentiment classification with the default parameters.

4.2. Performance of Models with Hyperparameter Tuning

In the second experiment, hyperparameter tuning was per- formed to improve the performance of the models. The performance of models using the Bag of Words and TF-IDF vectorization techniques with tuned parameters is compared in Table 4. Overall, the performance of the models using default parameter values in the first experiment does not differ significantly from the performance of the tuned models. SVM with the TF-IDF technique outperformed all othermodels with (89.200%) accuracy, (88.500%) precision, (90.200%) recall, and (89.30%) AUC. Moreover, the Performance of SVM using Bag of Word is slightly enhanced compared to its performance employing the default value. After hyperparameter tuning, logistic regression with a Bag of words performed better than TF-IDF. However, the performance of the random forest was marginally diminished compared to using the default values.

Table 1. Comparison of classification models using the Bag of words and RF-IDDF vectorization techniques, the default parameter

Model	Logistic Regression		Random Forest		SVM	
	Bag of Words	TF-IDF	Bag of Words	TF-IDF	Bag of Words	TF-IDF
Accuracy	87.20%	89.20%	85.20%	86.00%	86.40%	89.10%
Precision	87.40%	88.80%	86.80%	87.70%	84.40%	88.00%
Recall	87.10%	89.80%	82.90%	83.60%	89.40%	90.50%
AUC	87.00%	89.00%	85.00%	86.00%	86.00%	89.00%

Table 2. Confusion matrix of logistic regression utilizing

True Labeled	SVM -TF-IDF		
	Positive	877	123
	Negative	95	905
		Negative	Positive
	Predicted Label		

TF-IDF withthe default parameter

Table 3. Confusion matrix SVM utilizes TF-IDF with thedefault parameter

True Labeled	Logistic Regression-TF-IDF		
	Positive	887	113
	Negative	102	989
		Negative	Positive
	Predicted Label		

Table 4. Comparison of classification models using the Bag ofWords and TF-IDF vectorization techniques with tuned parameters

Model	Logistic Regession		Random Forest		SVM	
	Bag of Words	TF-IDF	Bag of Words	TF-IDF	Bag of Words	TF-IDF
Accuracy	88.40%	81.80%	84.40%	85.10%	87.60%	89.20%
Precision	88.00%	79.10%	81.50%	82.50%	87.10%	88.50%
Recall	89.10%	86.40%	89.00%	89.10%	88.20%	90.20%
AUC	88.50%	81.80%	84.40%	85.10%	87.60%	89.30%

Table 5. confusion matrix of logistic regression utilizing TF-IDFwith the default parameter

True Labeled	Logistic Regression-TF-IDF		
	Positive	887	122
	Negative	109	891
		Negative	Positive
	Predicted Label		

Table 6. Confusion matrix of SV utilizing TF-IDFwith the default parameter

True Labeled	**Logistic Regression-TF-IDF**		
	Positive	883	117
	Negative	98	902
		Negative	Positive
	Predicted Label		

As shown by the confusion matrix in Tables 5 and 6, even though the performance of logistic regression with Bag of Words is superior to TF-IDF in the second experiment, it could not surpass the performance of logistic regression with TF-IDF from the first experiment. In the second experiment, the SVM model with the TF-IDF technique can minimize false positives more effectively than the SVM model with the TF-IDF technique in the first experiment.

The results of both experiments demonstrate that logistic regression utilizing the TF-IDF technique with the default parameter minimizes false positives more effectively than other models. As a result, it is selectedas the best model for classifying the sentiment from movie review data.

5. Conclusion

The data in the actual world is frequently massive, unstructured, and tedious. Text analytics and sentiment analysis have been applied to such unstructured data in this project to classify user reviews as positive or negative. The analysis was conducted on the IMDB movie review dataset collected from Kaggle. Two major issues were addressed: machine learning models cannot be trained with unstructured data, and manually extracting attributes from vast data is impossible.

Therefore, this project had two objectives: to preprocess the raw data and to create machine learning classifiers. For preprocessing, numerous strategies were utilized, such as removing Unicode characters, stop words, HTML, URLs, etc. In addition, the texts were normalized using tokenization and lemmatization techniques. The preprocessing phase was concluded by doing an N-Gram analysis, thus achieving the first objective.

Two feature extraction methods, Bag of Words and TF-IDF, were employed to convert the text data into numerical form. Then, three machine learning modelswere trained and evaluated, including Logistic Regression, SVM, and Random Forest. Two experimental setups were employed, with each modeltrained using the default and tuned parameters. Several evaluation criteria were utilized to evaluate the models, including the Confusion Matrix, Accuracy, Recall, Precision, and AUC score. The results demonstrated that Logistic Regression employing the TF-IDF approach with default parameters outperformed other models because it could effectively minimize false positives.

References

[1] Internet Users Worldwide 2023 - Statista, 2022. [Online]. Available: https://www.statista.com/statistics/1190263/internet- users-worldwide/

[2] Jacob R. Pentheny, "*The Influence of Movie Reviews on Consumers*," Honors Theses and Capstones.[Google Scholar] [Publisher Link]

[3] P. G. Preethi, V. Uma, and Ajith Kumar, "Temporal Sentiment Analysis and Causal Rules Extraction from Tweets for Event Prediction," *Procedia Computer Science*, vol. 48, pp. 84–89, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[4] Bo Pang, and Lillian Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. (1–2), pp. 1-135, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[5] Andrew L. Maas et al., "Learning Word Vectors for Sentiment Analysis," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142-150, 2011. [Google Scholar] [Publisher Link]

[6] Richard Socher et al., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631-1642, 2013. [Google Scholar] [Publisher Link]

[7] Joosung Yoon, and Hyeoncheol Kim, "Multi-Channel Lexicon Integrated CNN-Bilstm Models for Sentiment Analysis," *Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017)*, pp. 244-253, 2017. [Google Scholar] [Publisher Link]

[8] Xiang Zhang, Junbo Zhao, and Yann LeCun, "Character-Level Convolutional Networks for Text Classification," *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, vol. 1, pp. 649-657, 2015. [Google Scholar] [Publisher Link]

[9] Afreen Jaha et al., "Text Sentiment Analysis Using Naïve Baye's Classifier," *International Journal of Computer Trends and Technology*, vol. 68, no. 4, pp. 261-265, 2020. [CrossRef] [Publisher Link]

[10] Aliaksei Severyn, and Alessandro Moschitti, "UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification," *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 464-469, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[11] Yassine Al Amrani, Mohamed Lazaar, and Kamal Eddine El Kadiri, "Random Forest and Support Vector Machine Based Hybrid Approach to Sentiment Analysis," *Procedia Computer Science*, vol. 127, pp. 511–520, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[12] B. Lakshmi Devi et al., "Sentiment Analysis on Movie Reviews," *Emerging Research in Data Engineering Systems and Computer Communications*, vol. 1054, pp. 321-328, 2020. [CrossRef] [Publisher Link]

[13] Rüdiger Wirth, and Jochen Hipp, "Crisp-Dm: Towards a Standard Process Modell for Data Mining," *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, vol. 1, pp. 29–39, 2013. [Google Scholar] [Publisher Link]

[14] Michael Fitzgerald, *Introducing Regular Expressions*, O'Reilly Media, Inc., 2012. [Google Scholar] [Publisher Link]

[15] Irfan Alghani Khalid, Cleaning Text Data with Python, 2020. [Online]. Available: https://towardsdatascience.com/cleaning-text-data-with-python- b69b47b97b76

[16] Gunisetti Tirupathi Rao, and Dr. Rajendra Gupta, "An Approach of Clustering and Analysis of Unstructured Data," *SSRG International Journal of Computer Science and Engineering*, vol. 6, no. 11, pp. 64-69, 2019. [CrossRef] [Publisher Link]

[17] Alberto Fernández et al., *Learning From Imbalanced Data Sets*, Springer, 2019. [Google Scholar] [Publisher Link]

[18] Tarek Kanan et al., "A Review of Natural Language Processing and Machine Learning Tools Used to Analyze Arabic Social Media," *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pp. 622-628, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Dipanjan Sarkar, *Text Analytics with Python*, Apress, 2019. [Google Scholar] [Publisher Link]

[20] Okan Ozturkmenoglu, and Adil Alpkocak, "Comparison of Different Lemmatization Approaches for Information Retrieval on Turkish Text Collection," *2012 International Symposium on Innovations in Intelligent Systems and Applications*, pp. 1-5, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[21] Koushik kumar, NLP: Bag of Words and TF-IDF Explained!, 2021. [Online]. Available: https://koushik1102.medium.com/nlp-bag-of-words-and-tf-idfexplained-fd1f49dce7c4

[22] Prafulla Mohapatra et al., "Sentiment Classification of Movie Review and Twitter Data Using Machine Learning," *International Journal of Computer and Organization Trends,* vol. 9, no. 3, pp. 1-8, 2019. [CrossRef] [Publisher Link]

[23] M. Borcan, TF-IDF Explained and Python Sklearn Implementation, 2020. [Online]. Available: https://towardsdatascience.com/tf-idf-explainedand-python-sklearn-implementation-b020c5e83275

[24] M. Sheykhmousa et al., "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6308-6325, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Sourish Ghosh, Anasuya Dasgupta, and Aleena Swetapadma, "A Study on Support Vector Machine Based Linear and Nonlinear Pattern Classification," *2019 International Conference on Intelligent Sustainable Systems* (ICISS), pp. 24-28, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[26] Duyu Tang, Bing Qin, and Ting Liu, "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification," *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1422-1432, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[27] Christopher D. Manning, and Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval*, Cambridge University Press, 2008. [Google Scholar] [Publisher Link]

[28] S. Arafin Mahtab, N. Islam, and M. Mahfuzur Rahaman, "Sentiment Analysis on Bangladesh Cricket with Support Vector Machine," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP),* pp. 1-4, 2018. [CrossRef] [Google Scholar] [Publisher Link]