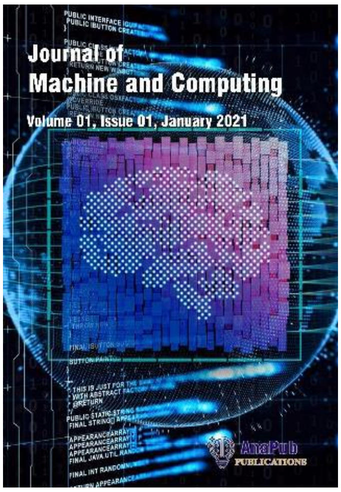
Journal Pre-proof

Sentiment Analysis of Product Reviews Using LSTM: A Comparative Evaluation with Machine Learning Algorithms Employing Bow and TF-IDF Techniques



Karthiga S, Sutha K, Pavithra V, Sakthivel S, Sowmya V and Sasidevi J

DOI: 10.53759/7669/jmc202505127

Reference: JMC202505127

Journal: Journal of Machine and Computing.

Received 16 April 2025

Revised form 28 May 2025

Accepted 10 June 2025

Please cite this article as: Karthiga S, Sutha K, Pavithra V, Sakthivel S, Sowmya V and Sasidevi J, "Sentiment Analysis of Product Reviews Using LSTM: A Comparative Evaluation with Machine Learning Algorithms Employing Bow and TF-IDF Techniques", Journal of Machine and Computing. (2025). Doi: https:// doi.org/10.53759/7669/jmc202505127.

Sentiment Analysis of Product Reviews Using LSTM: A Comparative Evaluation with Machine Learning Algorithms Employing Bow and Tf-Idf Techniques

[1]S. Karthiga, [2]K. Sutha, [3]V. Pavithra, [4]S. Sakthivel, [5]V. Sowmya, [6]J. Sasidevi

[1,2,3,5]Department of Computer Science and Applications, Faculty of Science and Humanities,
SRM Institute of Science and Technology, Ramapuram Campus, Chennai, India.
[4]Department of Computer Science and Engineering, School of Computing,
SRM Institute of Science and Technology, Tiruchirappalli Campus, Tamilnadu, India.
[6]Department of Computer Science and Business Systems, K. Ramakrishnan College of Engineering,
Samayapuram, Trichy.
[1]karthiga2191@gmail.com, [2]ksutha1986@gmail.com, [3]vpavithra.1989@gmail.com,
[4]sakthivel.solaimuthu84@gmail.com, [5]sowmydev@gmail.com, [6]sasisa[illegible]19@gmail.com

Correspondence should be addressed to V. Pavithra email id: vpavithra.1989@gmail.com

ABSTRACT

Sentiment analysis has become an invaluable tool in understanding consumer opinions in large datasets. This study explores sentiment analysis of the product review dataset applying different machine learning classification algorithms, specifically focusing on two primary feature extraction methods: (TF-IDF) and (BOW) A thorough comparison was conducted to assess the effectiveness of each method alone, as well as a novel hybrid technique that merges both TF-IDF and BOW. And compared with deep learning approach, our findings demonstrate that feature extraction technique significantly enhances classification performance. Among the tested algorithms, logistic regression with tfidf, bow exhibited even greater accuracy. Obtaining the most accurate results possible from the sentiment analysis is the primary objective of this endeavor. The first step in the process of analyzing and classifying the data is going to be the preprocessing of the data, followed by the extraction of features, then the categorization of sentiments via the use of machine learning algorithms, and lastly the assessment of the algorithms. The end findings indicate that the SVM classifier obtained an accuracy of 93%, the Naive Bayes classifier achieved an accuracy of 91%, the Logistic regression classifier got an accuracy of 94%, and the LSTM classifier earned an accuracy which was 93.58%. In future work may explore the integration of additional feature extraction methods with deep learning to refine and improve sentiment analysis models.

Keywords:

Sentiment Analysis, Machine Learning, Deep Learning, LSTM, Feature Extraction, BOW, TF-IDF

1. INTRODUCTION

On a daily basis, millions of individuals post their reviews, thoughts, and assessments on movies and items on a variety of social networking websites such as Facebook and Twitter, as well as on e-commerce websites such as product and movie reviewing websites. It is possible that these evaluations and comments include some of the expectations that users have, which is something that is significant to business and marketing experts as well as researchers. The purpose of sentiment analysis is to examine a substantial quantity of data in order to ascertain the many emotions that are conveyed within it, whether they be good, negative, or neutral[1]. E-commerce refers to the online platform where individuals engage in buying and selling goods and services, as well as conducting financial transactions and exchanging information [2]. The advent of the e-commerce system has led to a shift in consumer behavior towards online purchasing, driven by customer evaluations and ratings. Consequently, it has become commonplace for individuals to assess product reviews prior to making a purchase in today's world. It will assist shoppers in purchasing high-quality products at reasonable prices. Implementing measures to mitigate cheating in the e-commerce system will be effective.

The comments may pertain to the product, the services provided by the shop, or the procedure of delivery. The abundance of reviews poses challenges in terms of readability and analysis. Feedback consist of two components: positive and negative reviews. The importance of customer reviews in driving sales for businesses is widely acknowledged. Sellers who possess an excellent reputation typically experience a significant surge in their sales volume [3]. In this day and age, people have a tendency to blindly accept the reviews that are accessible online and make an opinion about any movie even before they have seen it. There is an abundance of textual information on movies that can be found on websites such as Amazon, IMDb, and Rotten Tomatoes website. The scores that users give to films are predicted based on the reviews that are posted on IMDb. Researchers working in the area of machine learning have examined a variety of methods that may be used to carry out the operation with the best possible degree of precision. The purpose of this study is to demonstrate how a deep learning approach known as BERT may be used to identify fake movie reviews on IMDb. The BERT-base-uncased type is used in the work that is being suggested. This kind of model makes use of pandas, torch, and transformer, and it demonstrated an accuracy of 93% when applied to the IMDb dataset.

Sentiment analysis is the computer process of recognizing and classifying the emotional attitude conveyed by an consumer in a written text. Its applications in industry span a broad spectrum, ranging from predicting market trends by analyzing sentiment in news and blogs, to discerning consumer contentment and displeasure through the feedback.[4] Text mining is the extraction of significant and captivating information from unstructured text. This methodology has three stages: data pre-processing, feature extraction from the preprocessed data, and polarity determination using DL and ML techniques based on the extracted features.[5] Preprocessing encompasses various processes, including tokenization, stop word elimination, converting to lowercase, stemming, and eliminating numerals. The next step is extracting features. Various text features include count vectors, bag of words, TF- IDF, word embeddings and NLP based methods.[6]

Most of the researchers conducted an examination on the influence of pre-processing and extraction methods for the sentiment analysis using amazon review dataset[6].In this paper The study will investigate the impact of different approaches, such as TF-IDF and BOW, on the outcome.After using various pre-processing approaches, two types of features are retrieved from the reviews. Subsequently, different machine learning classification techniques are employed to determine which model is superior and. explored the implementation of LSTM and systematically compare it with different machine learning technique.

2. RELATED WORK

Apoorv Agarwal et al., [7] examines the influences of pre-processing. The tweets under consideration contain a plethora of symbols, unfamiliar terms, and abbreviations. The investigation involved the different pre-processing technique to clean the data. Researchers also explored the significance of slang phrases and spelling correction. In their experiment, they utilized an SVM classifier. With a previous recommendation state-of-the-art unigram model serving as our baseline, we report an overall increase of more than four percent for two classification tasks. These tasks include a binary classification of positive vs negative and a three-way rating of positive versus negative versus neutral statements. For all of these objectives, we provided a comprehensive series of experiments that were conducted using manually annotated data, which is a random sample of a stream of tweets. In this study, we studied two different types of models: tree kernel models and feature based models. We demonstrated that both of these models performed better than the unigram baseline approach. When it comes to our feature-based method, we do feature analysis, which demonstrates that the features that mix the prior polarity of words and their parts-of-speech tags are the most significant features. As a preliminary conclusion, we have determined that the analysis of sentiment for Twitter data is not significantly different from the study of sentiment for other types of content.

Rafat Habib Quraishi [8] employed ML and DL techniques, including SVM, LSTM, GRU for the sentiment analysis using IMDB dataset. The performance measurements indicated that deep learning based methods surpassed classical machine learning models in binary classification.

Ratings and reviews left by customers are becoming more significant since they are likely to play a significant part in the process of selling and purchasing a product. Reviews from consumers also give first-hand feedback that comes straight from the customers themselves; this may be beneficial to sellers as well, since it can help them improve future sales. By analyzing the evaluations, one might become aware of the likely factors that led to the success or failure of a product. Consequently, the purpose of this article is to demonstrate the sentiment analysis of the reviews in order to get a deeper comprehension of the sentiments that were conveyed by the consumers. The mobile phones, which are quite popular and are used by a large number of people, were selected as the product, and Amazon was selected as the digital seller for this particular research. In the beginning, this effort started with the preprocessing of the data. Following the completion of the data pretreatment step, the Bow and n-grams word embedding techniques were used to represent the clean reviews in vector form. Subsequently, the features were produced. Finally, the performance of supervised machine learning classifiers such Decision Tree, Naive Bayes, Random Forest, and SVM was experimentally tested using accuracy, recall, f1-score, and precision. These metrics were used to evaluate the effectiveness of the classifiers. According to the findings of the empirical study, the Random Forest Classifier has the highest level of performance, with an accuracy rate of 97.48%. The feature extraction approaches mentioned in reference [9] included TF, TF-IDF, Global Vectors (GloVe), and word2vec. TF-IDF utilizes count of word to ascertain the significance of words in relation to a specific document. GloVe measures likelihood of two words appearing together, while word2vec identifies significant connections between them.The output of each technique results in a matrix that represents all as vectors.

Soni and Kirti Mathur [10] model is dependent on the combination of numerous embeddings that are processed by an attention encoder and then fed into an LSTM framework. In order to extract contextual information, our method involves combining the embeddings of Paragraph2vec, ELMo, and BERT. Additionally, FastText is used in an effective manner in order to seize syntactic properties. Following that, these embeddings were combined with the embeddings that were acquired from the attention encoder, which resulted in the whole embeddings being formed. In order to speculate on the ultimate categorization, an LSTM model is used. The Twitter Sentiment140 dataset as well as the Twitter US Airline Sentiment dataset were used in the conducting of our studies. A number of well-known models, including LSTM, Bi-directional LSTM, BERT, and Att-Coder, were used to assess and compare the performance of our specific fusion model. In terms of performance, the results of the tests make it abundantly evident that our technique offers superior outcomes than the baseline models. The LSTM model achieved an accuracy of 87% while evaluating online reviews in the Hindi language. A sentiment analysis was conducted using an LSTM model that incorporated an attention encoder Jitendra. In the study conducted by [11], different ML techniques were utilized: Support Vector Machine(SVM), NB and Maximum Entropy , for sentiment classification.Machine learning classifiers were trained using both unigrams and weighted unigrams. The experimental result was assessed based on its correctness. SVM algorithm attained accuracy of 81%, surpassing all other approaches. In contrast to typical machine learning classification algorithms, LSTM has demonstrated its effectiveness in achieving high accuracy for emotion classification [12].

Neogi et al. [13] conducted a study where they gathered around 20,000 tweets for sentiment analysis. The models employed were BOW and TfIdf. The investigation exposed that the BOW method outperformed the other technique.Different ML techniques were utilized . Among these algorithms, random forest achieved the greatest accuracy in classification. N-grams and TfIdf were contrasted as feature extraction methods for sentiment analysis by Das et al. [14]. Classification techniques included k-nearest neighbors ,SVM, RF, Multinomial NB, Decision Tree and, LR, TfIdf was found to significantly increase feature extraction when compared to the other two feature extraction techniques. The RF obtained the greatest accuracy values (93.8%) while using TfIdf. In their study, Xiao et al. [15] opted to employ LSTM technique with various datasets,The accuracy rates for the LSTM model was 89.85% and the model proved to be successful.

Gaur et al. [16] employed the NaiveBaiye using TF-IDF to classify the Twitter review. Based on accuracy, recall, and precision performance criteria, the proposed model outcomes demonstrated enhanced accuracy (84.44%) and precision.

3. METHODOLOGY

The research commenced by utilizing the Amazon review dataset and implemented various preprocessing techniques to adequately prepare the data for analysis. The preprocessing stages encompassed addressing missing values, standardizing the data, and partitioning the data into relevant subsets according to the analysis criteria. Subsequently, we employed both BoW and TF-IDF approaches to extract features. After performing feature extraction, we utilized various machine learning classification algorithms to analyze the data and assessed their performance using four performance criteria. Ultimately, we evaluated the efficacy of these conventional machine learning methods in comparison to a deep learning strategy.

3.1. Dataset description

This project leverages a dataset obtained from Kaggle.com, focusing on Amazon product reviews. The dataset encompasses a substantial collection of over 34,660 reviews contributed by customers across diverse product categories, including electronics, home furniture, and various other commodities. Beyond customer reviews, the dataset incorporates crucial elements such as product ratings and a diverse set of additional information. Comprising a total of 21 features, the dataset includes comprehensive details ranging from product specifications to star ratings provided by customers, encompassing a holistic perspective of customer feedback and product attributes.

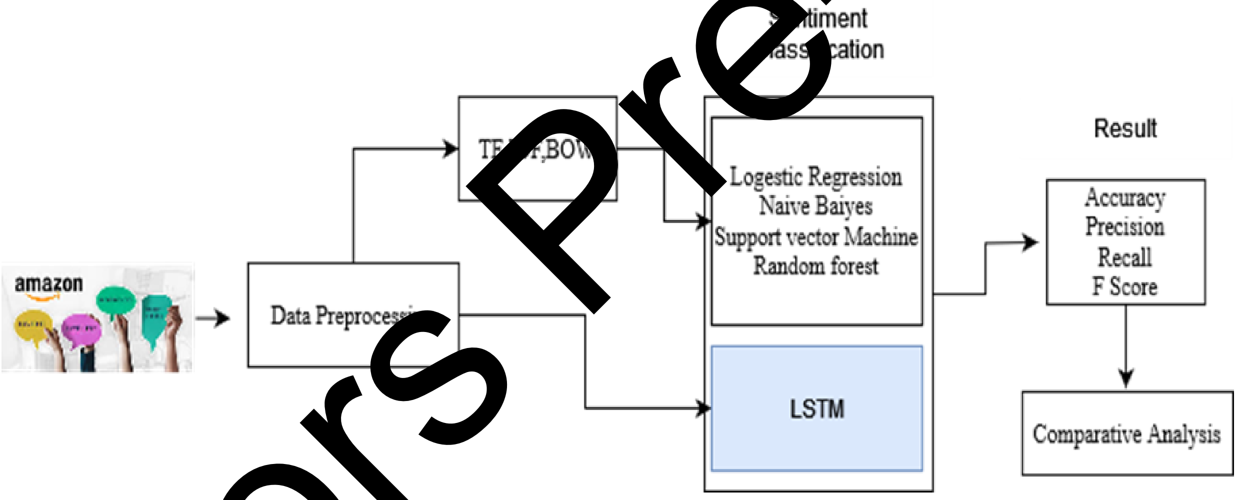


Figure 1. Methodology of sentiment analysis.

3.2. Preprocessing

The study involved the elimination of stop words, and various data preprocessing techniques, including stemming, tokenizing, and lemmatization, were applied to the Amazon dataset. Due to the potential informality and noise in user reviews, the data required thorough cleaning and transformation to ensure it adheres to a format understandable by the classification model.

3.3. Tokenization

It is the procedure of dissecting a text, such as a paragraph or a sentence, into separate words or "tokens." Tokens serve as the fundamental components of language, and the process of tokenization aids computers in comprehending and manipulating human language by dividing it into manageable segments.An illustration of tokenization may be shown by breaking down the statement "I love chocolates" into three distinct tokens: "I," "love," and "chocolates."

3.4. Normalization

Multiple activities are executed concurrently in order to accomplish normalization The method entails converting the text to either uppercase or lowercase, eliminating punctuation, and translating numerals into their respective nouns. This enhances the consistency of preprocessing applied to the document.

3.5. Stemming

Stemming is a technique to obtain the base form of words by removing affixes. It is akin to pruning a tree's branches down to its main stems. For instance, the root of the words speaking, speaks, and speak is speak. Lemmatization is advisable when the significance of the word is crucial for analysis. It is the process of classifying different inflected forms of a word into a unified group. Lemmatization improves the accuracy and efficiency of technologies like chatbots and search engine queries by combining words with similar meanings into a single term. It refers to the procedure of simplifying a phrase to its fundamental form., which is called a lemma. For example, the verb "speaking" might be identified as "speak". 3.2.5 Stop Words removal refer to a collection of frequently encountered terms in any given language that have little semantic value in sentences. These terms are ubiquitous in the grammatical structure of all languages. Each language possesses an own collection of stop words. Some examples of English stop words include "the," "she," "us," "we," "her," and "himself." We have employed manual data cleansing techniques in conjunction with regular expressions in Natural Language Processing (NLP) to remove any unwanted artifacts or disturbances. The noise removal process is executed with meticulous care to ensure the elimination of a limited number of rows in the dataset, which may result in reduced accuracy. The regular expression employed for data cleansing effectively eliminated superfluous white spaces and organized the data into appropriate columns.

3.6. Feature Extraction Techniques (TF-IDF and Bag of words)

The acronym TF in TF-IDF stands for term frequency, Term frequency is a metric that quantifies the frequency of a term's occurrence in a text, indicating that the term is more significant than other terms in the document.. Words possessing a high TF value hold significant importance within manuscripts. Conversely, the document frequency (DF) indicates the frequency of occurrence of a particular word in the collection of documents. The program determines the frequency of the word over numerous texts, rather than just one document. Words having a high DF value lack significance as they are frequently seen in all documents. The IDF is to quantify the significance of terms across all publications. The high IDF values indicate in equation 1-3 the presence of uncommon terms in all papers, leading to a rise in their significance[17].

$$\mathrm{TF} = \frac{(\text{Number of Times term } t \text{ present in a document})}{(\text{Total number of terms in the document})} \tag{1}$$

$$\mathrm{IDF} = \frac{(\text{ Total Number of document })}{(\text{number of terms } t \text{ in the document})} \tag{2}$$

$$\textit{TF-IDF} = TF(t) * IDF(t) \tag{3}$$

The BoW model is a simple representation utilized in NLP. A text is an unstructured assemblage of its constituent words, devoid of any consideration for syntax or even the sequence of words. During the process of text classification the weight assigned to a word in a document is determined by its frequency inside that document as well as its frequency across other publications.

3.7. Classification Algorithms

Naïve Bayes is a type of generative learning algorithm that seeks to mimic the distribution of inputs in a certain class or category. Unlike discriminative classifiers such as logistic regression, it does not gather information about the crucial features that distinguish between classes.. It is extensively employed in tasks such as text classification, spam filtering, and recommendation systems.

Logistic Regression approach is commonly employed for classification and is classified as a Generalized Linear Model. Logistic regression is a statistical method used to represent the chances that describe the outcome of an experiment[22]. This strategy is also known as Maximum Entropy..

In high or infinite dimensional space SVM creates a hyperplane or a collection of hyperplanes,that is located at the maximum distance from the closest training data points in each class achieves a high level of separation.. This is because a larger margin generally leads to a smaller generalization error for the classifier. It demonstrates efficacy in spaces with a large number of dimensions and exhibits varying behavior depending on the specific mathematical functions, referred to as the kernel. Kernel functions like sigmoid, polynomial, RBF, and linear are frequently used in SVM classifiers. The number 82 is encapsulated between square brackets.[22]

3.8. LSTM

LSTM networks are an extension of Recurrent Neural networks (RNNs) specifically created to effectively learn and capture the patterns and relationships in sequential or temporal data, including their long-term dependencies, with greater accuracy compared to traditional RNNs.

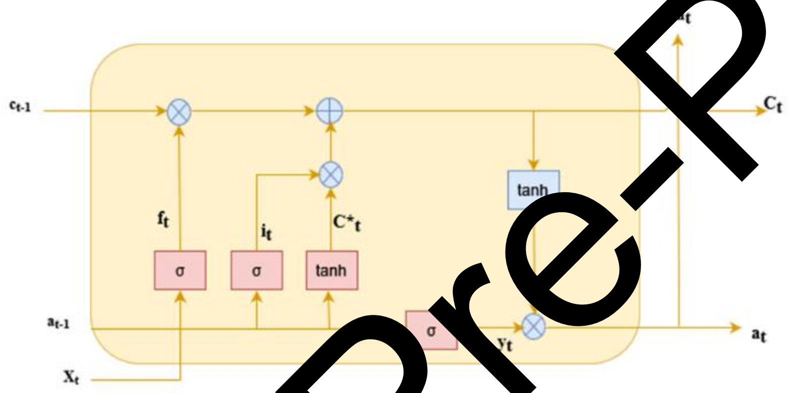


Figure 2. LSTM Neural Network

The LSTM model consists of three gates: Input ,forget and output gate, in addition to the cell memory. The data to be updated and saved in the memory cell is dictated by the gate input. The forget gate is responsible for evaluating the suitability of input/output information for passing. If the result is zero for the forget gate, the information is discarded, however if the output is near to one, the information is preserved. The ability of LSTM to address the challenges of exploding problem and disappearing gradient is due to its functioning at the forget gate. The cell state stays unchanged by the output gate.nevertheless, the date serves to differentiate between the actual information and the cell state[23] in equation 4-11.

$$F(t) = \sigma\left(W_f \cdot [H_{t-1}, X_t] + b_f\right) \quad (4)$$

$$I(t) = \sigma(W_i \cdot [H_{t-1}, X_t]) + b_i) \quad (5)$$

$$\tilde{C}(t) = tanh(W_c \cdot [H_{t-1}, X_t] + b_c) \quad (6)$$

$$C(t) = f_t * C_{t-1} + I_t * \tilde{C} \quad (7)$$

$$O(t) = \sigma(W_o \cdot [H_{t-1}, X_t] + b_o) \quad (8)$$

now, input weight is W_f, W_i , W_c, and W_c , bias is b_f, b_i, b_c and b_o, t is time state, $t-1$ is prior time state, X is input; H is output, and C is cell status.

$$H(t) = O_t * tanh(C) \tag{9}$$

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{10}$$

$$Tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{11}$$

3.9. Evaluation Metrics

There are four potential outcomes for the provided data: true negative (TN), false negative (FN), true positive and false positive. TP data is categorized as positive and labeled as positive, while FN data is categorized as negative but labeled as negative. FP data is mislabeled as negative but classed as positive, whereas TN data is correctly labeled and classified as negative[24] [25].

The Accuracy Rate refers to the capacity to accurately classify user evaluations according to their relevant polarity. It indicates the positive values that are truly positive. A Higher value shows less false positive rate (FPR). The recall is a metric that quantifies the accuracy of our model in properly detecting True Positives. F1-score analyzes the accuracy of the proposed system based on recall and precision rates.

The accuracy rate, precision, recall and F1 score is provided as

$$\begin{aligned} \text{Overall Accuracy } &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\ \text{Precision } &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Recall } &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ f_1\text{-Score } &= 2 \cdot \frac{\text{Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}}, \end{aligned} \tag{12}$$

4. RESULTS AND DISCUSSION

This study utilized distinct classification algorithms to analyze the Amazon review dataset, evaluating the effectiveness of two Feature Extraction techniques: TFIDF in isolation, combination of TF-IDF and Bag of Words (BOW) .Finally compared with deep learning technique LSTM. The analysis of revealed significant differences in the effectiveness of these methods and the accuracy is given in table 1.

Table 1: Comparison of ML algorithm with TFIDF ,TFIDF-BOW and LSTM

Model	Feature Extraction	Accuracy	Precision (Negative)	Precision (Positive)	Recall (Negative)	Recall (Positive)	F1-Score (Negative)	F1-Score (Positive)
Logistic Regression	TF-IDF	0.9416	0.69	0.95	0.18	0.99	0.28	0.97
Logistic Regression	BOW	0.9410	0.59	0.95	0.30	0.99	0.40	0.97
Logistic Regression	Hybrid	0.9413	0.59	0.95	0.30	0.99	0.40	0.97
Naive Bayes	TF-IDF	0.9358	0.83	0.94	0.01	1.00	0.02	0.97
Naive Bayes	BOW	0.9183	0.39	0.96	0.45	0.95	0.42	0.96

Naive Bayes	Hybrid	0.9158	0.38	0.96	0.46	0.95	0.41	0.95
SVM	TF-IDF	0.9404	0.79	0.94	0.11	1.00	0.19	0.97
SVM	BOW	0.9357	1.00	0.94	0.01	1.00	0.01	0.97
SVM	Hybrid	0.9358	1.00	0.94	0.01	1.00	0.02	0.97
LSTM	None	0.9358	0.51	0.95	0.31	0.98	0.38	0.97

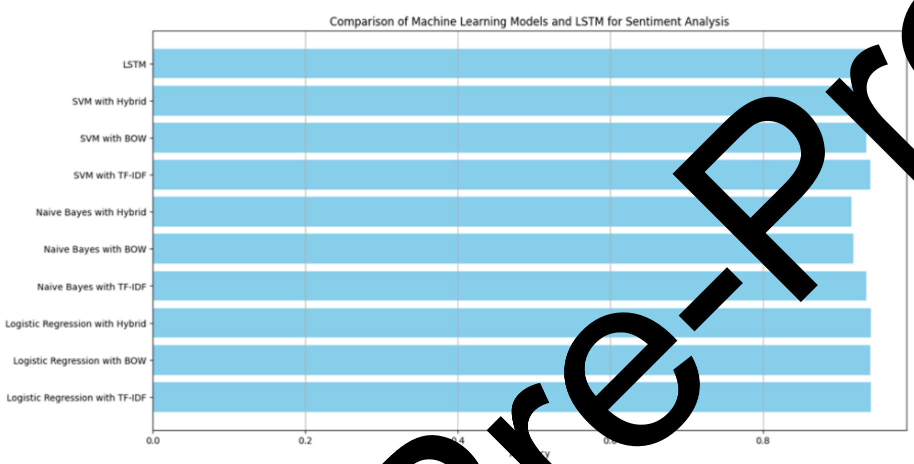
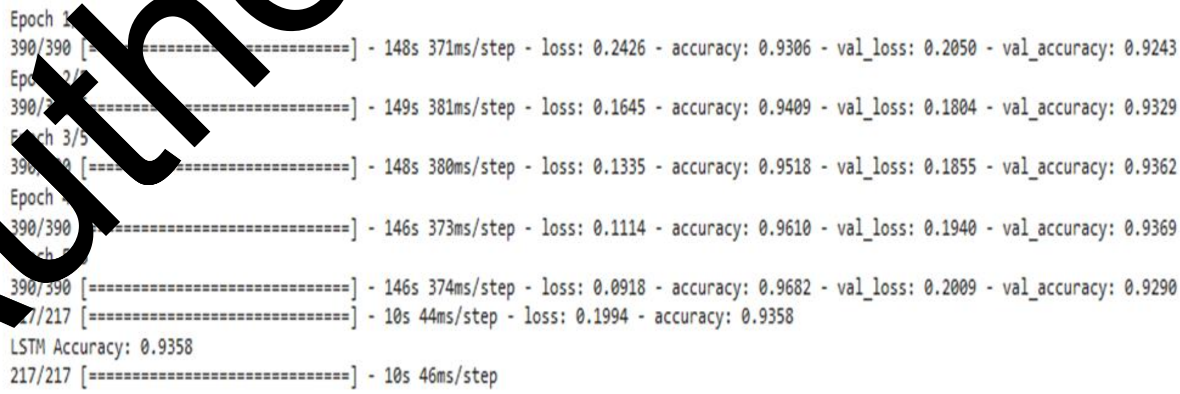


Figure 3. Performance of ML algorithm with feature extraction and LSTM Model.

In figure 3 comparison shown clearly. Logistic Regression consistently performed well across all feature extraction methods (TF-IDF, BOW, and Hybrid), with the highest accuracy observed using TF-IDF (0.9416). Naive Bayes showed a notable drop in performance with the Hybrid approach, achieving the lowest accuracy among the tested algorithms (0.9158). SVM demonstrated competitive accuracy with both TF-IDF and BOW, closely following the performance of Logistic Regression.LSTM without explicit feature extraction achieved an accuracy of 0.9358, which is competitive but slightly lower than the top-performing Logistic Regression with TF-IDF.

Figure 4 Shows the LSTM Network being trained using training dataset. A total of 5 epochs are executed to train Lstm network



```
Epoch 1/5
390/390 [==============================] - 148s 371ms/step - loss: 0.2426 - accuracy: 0.9306 - val_loss: 0.2050 - val_accuracy: 0.9243
Epoch 2/5
390/390 [==============================] - 149s 381ms/step - loss: 0.1645 - accuracy: 0.9409 - val_loss: 0.1804 - val_accuracy: 0.9329
Epoch 3/5
390/390 [==============================] - 148s 380ms/step - loss: 0.1335 - accuracy: 0.9518 - val_loss: 0.1855 - val_accuracy: 0.9362
Epoch 4/5
390/390 [==============================] - 146s 373ms/step - loss: 0.1114 - accuracy: 0.9610 - val_loss: 0.1940 - val_accuracy: 0.9369
Epoch 5/5
390/390 [==============================] - 146s 374ms/step - loss: 0.0918 - accuracy: 0.9682 - val_loss: 0.2009 - val_accuracy: 0.9290
217/217 [==============================] - 10s 44ms/step - loss: 0.1994 - accuracy: 0.9358
LSTM Accuracy: 0.9358
217/217 [==============================] - 10s 46ms/step
```

Figure 4. Validation accuracy and loss of LSTM Model

The training process was carried out for a total of 5 epochs. The following information provides a breakdown of each epoch:

During Epoch 1, the LSTM model attained a training accuracy of 93.06% and a training loss of 0.2426. The validation accuracy was 92.43%, and the validation loss was 0.2050. During Epoch 2, the training accuracy rose to 94.09%, while the training loss fell to 0.1645. The validation accuracy improved slightly to 93.29%, accompanied by a validation loss of 0.1804. During Epoch 3, there was a notable enhancement in training accuracy, reaching 95.18%, and a reduction in training loss to 0.1335. The validation accuracy increased to 93.62%, with a validation loss of 0.1855. Throughout Epoch 4, the model achieved a training accuracy of 96.10% and a training loss of 0.1114. The validation accuracy was 93.69%, and the validation loss was 0.1940. Epoch 5 concluded with a training accuracy of 96.82% and a training loss of 0.0918. The validation accuracy was 92.90%, and the validation loss was 0.2009. Finally, the model was evaluated on the test set, achieving an accuracy of 93.58% and a loss of 0.1994.

The figure depicts the graphical representation of accuracy and validation accuracy, whereas figures 5 exhibit the graphical representation of accuracy and validation loss.

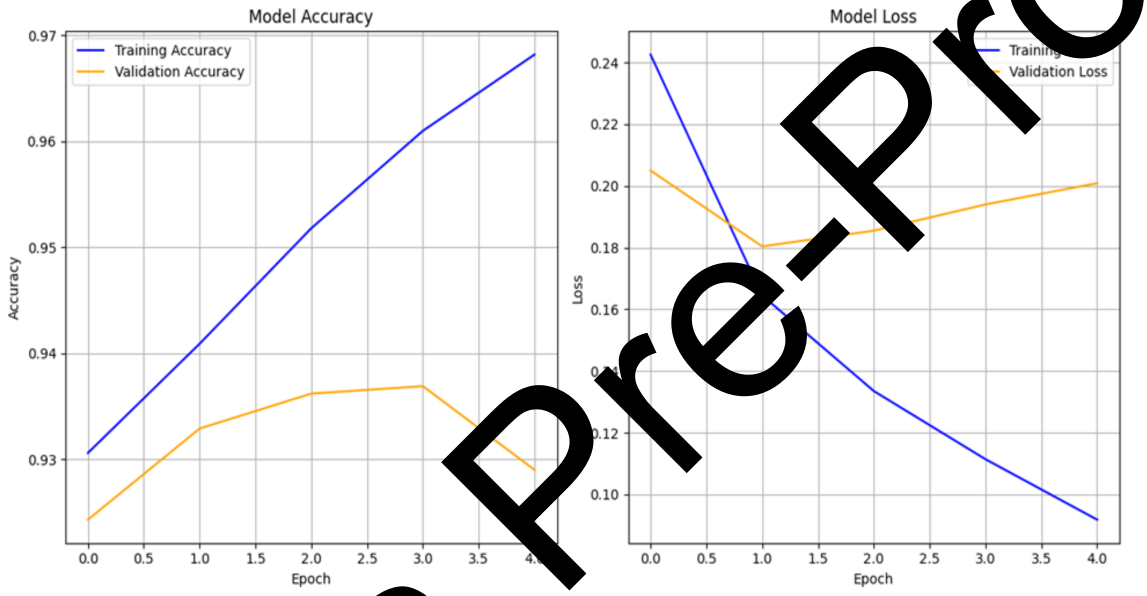


Figure 5. The model accuracy and loss of the LSTM model.

The findings indicate Logistic Regression consistently performed well across all feature extraction methods (TF-IDF, BOW, and Hybrid), with the highest accuracy observed using TF-IDF (0.9416). Naive Bayes showed a notable drop in performance with the Hybrid approach, achieving the lowest accuracy among the tested algorithms (0.9158).SVM demonstrated competitive accuracy with both TF-IDF and BOW, closely following the performance of Logistic Regression.LSTM without explicit feature extraction achieved an accuracy of 0.9358, which is competitive but slightly lower than the top-performing Logistic Regression with TF-IDF.

5. CONCLUSION

In this study, we utilized distinct classification algorithms to analyze the Amazon review dataset with a specific emphasis on two feature extraction techniques: TF-IDF and BOW. Traditional machine learning algorithms like Logistic Regression and SVM with feature extraction methods (especially TF-IDF) outperform the LSTM model for this particular sentiment analysis task. The Naive Bayes classifier, while effective with TF-IDF, shows a significant decline with the Hybrid approach, indicating a possible overfitting or inefficiency in combining features. The deep learning model (LSTM) still provides a strong performance without the need for explicit feature extraction, demonstrating its potential for handling raw text data directly. However, it slightly lags behind the best-performing traditional machine learning approaches in terms of accuracy.

While this study focused on deep learning without explicit feature extraction, there is significant potential for improving accuracy by integrating feature extraction techniques with deep learning models. Feature extraction methods such as TF-IDF and BOW could provide richer input representations for deep learning architectures, potentially enhancing their performance beyond what was achieved in this study. Exploring the combination of feature extraction techniques with deep learning models could yield even better results.

Future research should focus on integrating feature extraction methods with deep learning models to leverage the strengths of both approaches. Implementing advanced architectures such as transformers, combined with feature extraction techniques like TF-IDF and BOW, could lead to higher accuracy and better performance. Additionally, fine-tuning hyperparameters and incorporating domain-specific knowledge could further enhance model performance.By combining the rich feature representations from traditional methods with the powerful learning capabilities of deep learning, future studies have the potential to significantly advance the state-of-the-art in sentiment analysis.

REFERENCES

[1] N. Raveendhran and N. Krishnan, “A novel hybrid SMOTE oversampling approach for balancing class distribution on social media text,” Bulletin of Electrical Engineering and Informatics, vol. 14, no. 1, pp. 638–646, Feb. 2025, doi: 10.11591/eei.v14i1.8380.

[2] N. R, N. K, S. R, S. Banu S, S. P, and B. P, “Graph-Based Rumor Detection on Social Media Using Posts and Reactions,” International Journal of Computing and Digital Systems, vol. 15, no. 1, pp. 173–182, Jul. 2024, doi: 10.12785/ijcds/160114.

[3] M. A. Fauzi, “Random forest approach fo sentiment analysis in Indonesian language,” Indones. J. Electr. Eng. Comput. Sci., vol. 12, no. 1, pp. 46–50, 2018, doi: 10.11591/ijeecs.v12.i1.pp46-50.

[4] Nishit Shrestha and Fatma Nasoz,” Deep Learning Sentiment Analysis Of Amazon.Com Reviews And RatingS”, International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.8, No.1, February 2019

[5] N.C. Dang, M.N. Moreno-García, F. De la Prieta, Sentiment analysis based on deep learning: a comparative study, Electronics 9 (3) (2020) 483.

[6] RaviderAhuja,akarsha chugh,shruthi kohli,shaurya gupta,pratyush” The Impact of Features Extraction on the Sentiment Analysis”,International Conference on Pervasive Computing Advances and Applications – PerCAA 2019

[7] Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. "Sentiment analysis of twitter data." In Proceedings of the workshop on languages in social media, pp. 30-38. Association for Computational Linguistics, 2011

[8] AH Quraishi” Performance analysis of machine learning algorithms for Movie Review”, International Journal of Computer Applications, 2020

[9] S. A. Aljuhani and N. S. Alghamdi, “A comparison of sentiment analysis methods on Amazon reviews of mobile phones,” International Journal of Advanced Computer Science and Applications, vol. 10, pp. 608–617, 2019.

[10] Soni J, Mathur K (2022) Sentiment analysis based on aspect and context fusion using attention encoder with LSTM. Int J Inf Technol

[11] A. S. a. A. A. a. D. P. Rathor, “Comparative study of machine learning approaches for Amazon reviews,” Procedia computer science, vol. 132, pp. 1552--1561, 2018

[12] S. Hochreiter, “The vanishing gradient problem during learning recurrent neural nets and problem solutions,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 6, no. 02, pp. 107–116, 1998.

[13] Neogi, A.S.; Garg, K.A.; Mishra, R.K.; Dwivedi, Y.K. Sentiment analysis and classification of Indian farmers’ protest using twitter data. Int. J. Inf. Manag. Data Insights 2021, 1, 100019. [CrossRef]

[14] Das M, Kamalanathan S, Alphonse P (2021) A comparative study on TF-IDF feature weighting

method and its analysis using unstructured dataset. CEUR Workshop Proc 2870:98–107

[15] S. Xiao, H. Wang, Z. Ling, L. Wang, and Z. Tang, “Sentiment analysis for product reviews based on deep learning,” in Proc. 2020 the Second International Conf. on Artificial Intelligence Technologies and Application (ICAITA), Dalian, 2020, 012103

[16] Gaur P, Vashistha S, Jha P. Twitter Sentiment Analysis Using Naive Bayes-Based Machine Learning Technique. In: Shakya S., Du KL., Ntalianis K. (eds) Sentiment Analysis and Deep Learning. Advances in Intelligent Systems and Computing, Springer, Singapore. 2023;1432. https://doi.org/10.1007/978- 981-19-5443-6_27

[17] sang woonkim,Joon min Gil,” Research paper classification systems based on TF-IDF and LDA schemes”,Human centric computing and information service. https://doi.org/10.1186/s13673-019-0192-7,26 aug 2019

[18] Alam, S., and Yao, N. (2018). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis, Computational and Mathematical Organization Theory. doi:10.1007/s10588-018-9266-8.

[19] Alsmadi, I. and Hoon, GK., (2018). Term weighting schemes for short-text classification: Twitter corpuses. Neural Computing and Applications. doi:10.1007/s00521-017-3298-8.

[20] Bao, Y., Quan, C., Wang, L., and Ren, F. (2014). The Role of Pre-processing in Twitter Sentiment Analysis, Lecture Notes in Computer Science, 615–624. doi:10.1007/978-3-319-09339-0_62.

[21] Asudani, D.S., Nagwani, N.K. & Singh, P. Impact of word embedding models on text analytics in deep learning environment: a review. Artif Intell Rev 56, 10345–10425 (2023). https://doi.org/10.1007/s10462-023-10419-1

[22] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830

[23]. Putra Fissabil Muhammad, Retno Kusumaningrum, Adi Wibowo,Sentiment Analysis Using Word2vec And Long Short-Term Memory (LSTM) For Indonesian Hotel Reviews,Procedia Computer Science,Volume 179,2021,

[24] Durairaj, S., Umar, M.M. and Natarajan, B., Evaluation of Bio-Inspired Algorithm-based Machine Learning and Deep Learning Models. In Bio-inspired Algorithms in Machine Learning and Deep Learning for Disease Detection (pp. 48-69). CRC Press

[25] Durairaj, S., S, S. and S, A.A.B., 2025. Hybrid key management WSN protocol to enhance network performance using ML techniques for IoT application in cloud environment. Peer-to-Peer Networking and Applications, 18(4), p.163.