Journal of Theoretical and Applied Information Technology
30th November 2024. Vol.102. No. 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MACHINE AND DEEP LEARNING MODELS FOR MULTI-CLASS SENTIMENT CLASSIFICATION

LAMIAA A. GAAFAR[1*], ATEF Z. GHALWASH[2], ALIAA A. YOUSSIF[2,3], HAITHAM A. GHALWASH[4]

[1]Faculty of Computer Science and Information Technology, Ahram Canadian University, Cairo, Egypt.
[2]Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt.
[3]College of Computing and Information Technology, Smart Village,Arab Academy for Science and Technology, Cairo, Egypt.
[4]Coventry University, Egypt Branch, School of Computing, New Cairo, Egypt.

*Corresponding author(s). E-mail(s): lamia.ali@acu.edu.eg.

ABSTRACT

Nowadays, Artificial Intelligence (AI) is renowned for embedding human-like intelligence in computers, enabling them to mimic human behavior. A pivotal domain within AI is recommendation frameworks, which aid users by suggesting various choices, thereby facilitating optimal decision-making in contexts like purchasing items, selecting healthcare services, movies, etc. This paper introduces classification based on sentiment analysis, aimed at extracting opinions from user reviews. The analysis employs eight models: five machine learning models— Extreme Gradient Boosting (XGBoost), Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF); two deep learning models— Long Short-Term Memory (LSTM) and a distilled version of Bidirectional Encoder Representations from Transformers (DistilBERT) transformer model; and a proposed model integrates Convolutional Neural Networks (CNNs) and Feedforward Neural Networks (FFNNs), alongside the Mamdani Fuzzy System. Notably, the LSTM model demonstrates superior performance, especially attributed to its efficacy in processing shorter sentences, typically ranging from 15 to 20 words as in the used data set, thus slightly outperforming the DistilBERT transformer model in this context. A comparative analysis between 3-class (positive, negative, and neutral) and 4-class (strongly positive, positive, negative, and neutral) classifications reveals the LSTM predominance across all models. Notably, the Long Short-Term Memory (LSTM) model excels in the 3-class sentiment classification, achieving an accuracy of 0.99, precision of 0.99, recall of 0.99, and an F1 score of 0.99 after oversampling.

Keywords: *Natural Language Processing, Sentiment, Classification, Deep Learning, Oversampling*

1. INTRODUCTION

A recommendation system is an AI algorithm that examines user data on preferences and behavior to create customized suggestions for products, services, or other items. These systems use multiple methods, including collaborative and content-based filtering, to detect trends and forecast what a user may prefer. AI plays a crucial role in recommendation systems by utilizing machine learning algorithms and other techniques to analyze large datasets, identify patterns, and predict user preferences. This improves the accuracy and relevance of recommendations, providing more personalized suggestions based on individual interests. Additionally, AI enhances the efficiency and scalability of recommendation systems, enabling them to process large amounts of data and generate recommendations in real time. Business owners post their products on social media to advertise them and to know the clients' feedback. Analyzing the clients' reviews using sentiment analysis helps them in improving their products. Sentiment analysis has a great role in business

intelligence, it responds to BI questions about the products and services [1]. Natural language processing (NLP) is employed to preprocess the users' reviews before conducting sentiment analysis. Text preprocessing has various techniques including removing punctuation, tokenization, changing case, removing stop words, stemming, and lemmatization [2]. Following the preprocessing phase, the word embedding phase starts to convert the tokens to numerical values. Various techniques are used in this phase, for instance, word2vec [3], Glove [4], and TF-IDF (Term Frequency - Inverse Document Frequency) [5]. Various Machine Learning techniques (ML) are used in this domain to classify the reviews, for instance, Support Vector Machines (SVM) [6], K-Nearest Neighbors (KNN) [7], Random Forest (RF) [8], Logistic Regression (LR) [9], Naïve Bayes (NB) [10], XGBoost [11] or Decision Tree (DT) classifier [12]. Deep learning techniques(DL) have also shown significant performance in the sentiment analysis domain, for instance, Long Short-Term Memory (LSTM) [13], Gated Recurrent Unit (GRU) [14], Recurrent Neural Network (RNN) [15] and DistilBERT transformer model [16].

This research aims to compare different algorithms for the classification task. The eight performed classifiers are XGBoost, Naïve Bayes (NB), Logistic Regression (LR), Support vector machine (SVM), Random Forest (RF), Long short-term memory (LSTM), DistilBERT transformer model, and the integration of Convolutional Neural Networks (CNNs) and Feedforward Neural Networks (FFNNs), alongside the Mamdani Fuzzy System. Word embedding task is done with TF-IDF for ML models, Glove for the LSTM model, a pre-trained tokenizer is implemented for DistilBERT and the Word2vec method for the integrated model. The number of reviews in the data set is 15060 after oversampling performed in the preprocessing task. The classification is performed into 4 classes, namely strongly positive, positive, neutral, and negative. The strongly positive and positive classes are concatenated into a single positive class and the classification is performed again into 3 classes. The 3-class classification outperforms the 4-class classification in all models. The LSTM model achieves optimal outcomes.

In this paper, section 2 represents the literature review, section 3 outlines the machine learning models employed, Section 4 details the deep learning models utilized, and section 5 explains the methodology performed. The results of the methodology are detailed in section 6. Section 7 introduces the conclusion. Finally, Section 8 contains the future work.

2. LITERATURE REVIEW

D. D. Dsouza et. al [17] proposed a sentiment analysis model of student feedback that will be used to suggest ways to enhance the college. Machine learning techniques are used for this task, for instance, Support Vector Machine (SVM), Multinomial Naïve Bayes Classifier, and Random Forest (RF). A comparison of the three algorithms is conducted. It is discovered that the Multinomial Naïve Bayes Classifier, with an accuracy of 80%, is more accurate than the other 2 approaches.

K. Devipriya et. al [18] conducted a study of several deep-learning algorithms. The study reveals that RNN, which is built on a tree representation of the phrase, performs better and is more suitable for sentiment label training since it can better detect the similarity between words in a sentence. Convolutional neural networks fail when phrase-level labels are recognized during the training data set. However, this problem can be solved by using pre-trained word2vec vectors, which also increase performance by reducing the problem of overfitting. It is possible to apply naive with complicated decision boundaries to create recommender systems for a variety of socially relevant applications. Naive is considered suitable for sentiment analysis.

M. Mohammed et. al [19] proposed a novel sentiment classification method that improves the accuracy of the fuzzy rule-based system (FRBS) when dealing with overlapping rules by integrating it with the Crow Search Algorithm (CSA). The proposed classification model is evaluated using three large-scale datasets from Amazon, Yelp, and IMDB and compared against other commonly used machine learning techniques such as SVM, Maximum Entropy, Boosting, and SWESA. The results indicate that the proposed model outperforms the machine learning techniques based on efficiency and accuracy with an accuracy rate of 94% for the IMDB dataset surpassing SVM (77.29%), Boosting (90.27%), Maximum Entropy (91.02%) and SWESA (81.04%).

A. Mounika et. al [20] suggested a four-level procedure to recommend books to users, which includes semantic network grouping of related sentences, sentiment analysis, reviewer clustering, and a recommendation system. The SA phase involves training and testing using deep

learning methodologies such as convolutional neural networks (CNN) using the n-gram approach. The results are used in the clustering phase to group reviewers by K-nearest neighbor (KNN). The final level uses the collaborative filtering algorithm to recommend the top-n most interesting books to users. The goal is to achieve high accuracy in a short amount of time.

During the sentiment analysis step, word embedding and feature extraction are carried out. The text documents are converted to vectors during the training phase by using the document-to-vector (doc2vec) model, tokenizer, and padding code sequence. To integrate the vectors quickly and effectively with multi-channeling, a CNN classifier and n-gram are used. A dense activation layer is created by merging and max pooling the channels which are retained as a hierarchical model to be compared to other data sets. In the testing step, pre-processed test data is used to execute the same feature extraction process, and the outcomes are compared and categorized as positive, negative, or neutral. The model's performance and accuracy are computed and recorded in the database [20].

B. T. Hung [21] proposed a recommender system based on sentiment analysis. It is decided to use deep learning techniques for the sentiment analysis task. They built a hybrid approach of long short-term memory (LSTM) and convolutional neural networks (CNN). A comparative study is done between the hybrid model and (LSTM and CNN) models separately. The results yield an accuracy of 0.8254 for the LSTM model, 0.8276 for the CNN model, and 0.8345 for the hybrid LSTM and CNN model so that the hybrid model gets the best accuracy results.

Y. Wang et. al [22] proposed a recommendation framework for movies using recommendation and sentiment analysis. Vector space representation is applied to represent the users' reviews as vectors. The sentiment analysis task is done using the lexicon so that the reviews are classified into positive and negative classes based on the sentiment lexicon. If the reviews of a movie contain words such as "amazing"," good" and "perfect", it is classified as a positive review and the movie is recommended to the users. A comparative study is carried out between (the collaborative filtering and content-based method) with and without sentiment analysis. The model with sentiment analysis gets higher results with a precision of 0.782 while the precision is 0.531 without performing the sentiment analysis.

S. L. Lo et. al [23] discussed the multilingual sentiment analysis in this review paper. Informal and scarce resource languages with formal languages are considered because the comments of users on social media contain lots of informal languages. Many approaches are discussed like lexicon-based approaches, support vector machine (SVM), Naïve Bayes (NB), etc.

U. Srinivasarao et. al [24] proposed a model to extract opinions from the reviews and assign actual ratings to opinions, a lexicon-based opinion mining or sentiment analysis method, such as the AFINN lexicon [25], is used in this research. Deep neural network technique is used like Neural Collaborative Filtering (NCF) [26]. NCF consists of two techniques: GMF (Generalized Matrix Factorization) and MLP (Multi-Layer Perceptron). Whereas MLP utilizes a non-linear kernel to learn the interaction function, GMF models interactions using a linear kernel. The NCF framework lets GMF and MLP train independently, and the results are fed into the final hidden layer. The experimental findings using two datasets demonstrate that the suggested model performs well at making recommendations.

G. Khanvilkar et. al [27] proposed a product recommendation model based on sentiment analysis using machine learning techniques such as Support Vector Machine (SVM) and Random Forest (RF). While random forest is more effective for accuracy as well as robustness, SVM is well known for its accuracy in prediction and classification. SVM and Random Forest machine learning algorithms in this system will help to enhance sentiment analysis for product recommendation using multiclass classification.

T. Karnan et. al [28] proposed a product recommendation framework based on sentiment analysis. A comparison is carried out between Collaborative filtering and stochastic learning [29]. Collaborative filtering has led to an average precision of 75% and an average recall of 30% while stochastic learning has led to an average precision of 80% and an average recall of 32%. Stochastic learning has proved its efficiency compared to collaborative filtering.

P. S. Bhargav et. al [30] proposed a sentiment analysis model that is performed on hotel reviews to rate the hotel as positive, negative, or neutral. Machine learning algorithms are used such as Naïve Bayes. When scaling a dataset and applying a linear equation to features and predictors, Naive Bayes performs well. It has retrieved some incorrect values for neutral outcomes.

S. Devi et. al [31] describe a methodology for performing sentiment analysis on product reviews, intending to classify them accurately as positive or negative. In this approach, natural language processing (NLP) is utilized to preprocess the text, while a Naïve Bayes Classifier is used for rating classification. The paper proposes augmenting the model's accuracy by identifying and excluding fake reviews and spam. The Support Vector Machine (SVM) algorithm is used to accomplish this. The model achieves a notable success rate, with accuracy reaching 80%.

M. U. Salur et. al [32] proposed a model for analyzing the huge amounts of users' feedback on social media applications. Natural language processing and deep learning are very important in performing sentiment analysis on these huge amounts of user' reviews. The proposed model demonstrates that various deep learning approaches successfully combine with various text representation approaches. Hence, embedding methods like character level and fastText are used in conjunction with CNN and LSTM algorithms. There are two branches used in the feature extraction procedure. The first branch employs character-level embedding with the CNN algorithm, whereas the second employs fastText embedding with the LSTM algorithm. All the features from both branches are combined as input for the softmax layer to accomplish the classification. Two separate experiments are done to confirm the proposed model performance. In the first experiment, the various deep learning models (BiLSTM, LSTM, GRU, CNN) are used with three word-embedding methods (FastText, Word2Vec, character-level embedding) so there are twelve deep learning models. In the experiment, the embedding methods word2vec and fastText have achieved better results when used with recurrent neural networks (RNN). When combining BiLSTM and FastText, the accuracy of the classification is 80.44%. Using CNN and character-level representation, the classification accuracy is 75.67%. The classification accuracy of the proposed hybrid model is 82.14%. From the first experiment, we notice that the accuracy of the proposed model is better than the other models. In the second experiment, a comparison is made between the proposed model and a previous model. The comparison is done on the same data set. While the proposed model's accuracy is 82.14%, the accuracy of the other model is 69.25%.

Suresh Kumar et. al [33] introduced a technique for broadening the sentiment word dictionary by utilizing Stanford's GloVE tool alongside sentiment analysis. It improved the weighting of sentiment words using a suitable model, which enhanced the accuracy of the results. To add further value, an emotional sentiment factor was included, along with additional features to increase accuracy. Information retrieval techniques were applied using the Vector Space Model, leading to notable improvements in accuracy. The proposed method achieved an overall accuracy of 92.78%, with positive sentiment accuracy at 91.33% and negative sentiment accuracy at 97.32%.

Dash, D. P. et. al [34], in this paper, reviewed various algorithms for efficient seizure detection using intracranial and surface EEG signals, focusing on advances in machine learning and deep learning over the past decade. The paper evaluated methods such as support vector machines (SVM), artificial neural networks (ANN), convolutional neural networks (CNN), and long-short term memory (LSTM) networks, discussing the effectiveness of different feature domains and the incorporation of additional physiological signals like electrocardiograms. Notably, the LSTM approach achieved a 96.5% accuracy in distinguishing seizure from non-seizure EEG signals, and the paper suggested that sentiment analysis might also be a viable method for seizure detection.

Mozumder et. al [35] introduced a new framework that utilized advanced machine learning and deep learning techniques, particularly BERT, to categorize customer feedback into essential drivers of customer satisfaction (CSAT). By combining TF-IDF methods with BERT embeddings, the framework considerably enhanced prediction accuracy, attaining an F1 score of 0.84 based on a dataset of 5,943 responses from 39 companies. This approach surpassed traditional methods such as SVM and MLP networks. The findings underscored

the potency of deep learning in improving CSAT modeling and emphasized the necessity of implementing advanced ML and DL models for effective strategic decision-making.

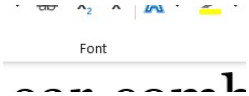
The existing literature on multi-class sentiment classification utilizing machine and deep learning techniques highlights several criticisms. A primary concern is the significant dependence on large labeled datasets, which are difficult to acquire and raise issues regarding the generalizability of the models. Furthermore, many studies overlook the interpretability of complex models, making it difficult for practitioners to grasp the decision-making processes involved. Additionally, traditional evaluation metrics frequently fall short in capturing the intricacies of sentiment in multi-class situations. Many methods also fail to account for linguistic context, such as sarcasm and irony, which can impact classification accuracy.

Deep learning and transformers play a significant role in sentiment classification, demonstrating superior performance compared to traditional machine learning techniques.

3. MACHINE LEARNING MODELS

In this paper, a variety of machine learning models are utilized, including Logistic Regression, Multinomial Naive Bayes, Support Vector Classifier (SVC), and Random Forests. Each model operates based on distinct principles and methodologies, as explained below:

1. Logistic Regression: Logistic Regression is a linear model for binary classification tasks. The probability p of a sample belonging to a particular class is modeled as a logistic function of a linear combination of the input features x:



Where w represents the weight vector and b is the bias [9].

2. Multinomial Naive Bayes: This model is used for multi-class classification problems. It applies Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable y and a dependent feature vector x_1 through x_n, the probability $P(y \mid x_1, \ldots, x_n)$ is given by [10]:

3. Support Vector Classifier (SVC): SVC is a powerful, non-probabilistic binary classifier. It constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification. The objective is to find the hyperplane that has the maximum margin, i.e., the maximum distance between data points of both classes. Mathematically, it solves the following optimization problem:

Subject to $y_i(w^T\phi(x_i)+b) >= 1$ for all i, where ϕ is the feature mapping function [6].

4. Random Forests: This is an ensemble learning method for classification (and regression) that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the mode of the classes output by individual trees. The general equation for a decision tree, which forms the basis of Random Forests, is:

$$y = f(\mathbf{x}, \Theta_k),$$

where Θk are the parameters of the k-th tree in the ensemble, and y is the output class [8].

5. XGBoost: XGBoost is a scalable and accurate implementation of gradient boosting machines (GBMs). It optimizes both computational speed and model performance. The model involves constructing a series of additive trees in a sequential manner, where each tree corrects the errors of the previous ones. The objective function that XGBoost minimizes is given by:

Where l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i, Ω represents the regularization term which penalizes the complexity of the model, and fk are the functions corresponding to the trees in the model.

XGBoost also incorporates techniques like sparsity-aware learning for handling missing data, and tree pruning based on the depth to prevent overfitting [11].

4. DEEP LEARNING MODELS

In the task of sentiment classification, two distinct neural network architectures are employed: Long Short-Term Memory (LSTM) and DistilBERT. The LSTM, a type of recurrent neural network (RNN), is particularly adept at processing sequences of data (such as text) by maintaining a 'memory' of past inputs through its internal state. This allows it to capture long-range dependencies within the text, crucial for understanding context and sentiment. The architecture of the LSTM Model is shown in Figure 1. Its mathematical formulation can be represented as:

ıt neural networ
: text) by mainta
: to capture long
and sentiment. '

Where σ is the sigmoid function, · denotes matrix multiplication, ∗ is the Hadamard product, W and b are the weights and biases, respectively, ft, it, and ot are the forget, input, and output gates, Ct is the cell state, and ht is the hidden state [13]. On the other hand, DistilBERT is a streamlined version of BERT (Bidirectional Encoder Representations from Transformers) as shown in Figure 2, designed for faster performance and lower memory consumption while retaining most of BERT's key features. It utilizes the transformer architecture, which leverages self-attention mechanisms to weigh the importance of different words in a sentence relative to each other for encoding the meaning of the text. The core idea of transformers can be encapsulated as follows:

where σ is the sigr

Here, Q, K, and V represent the query, key, and value matrices, respectively, derived from the input embeddings, and dk is the dimension of the key vector [16].

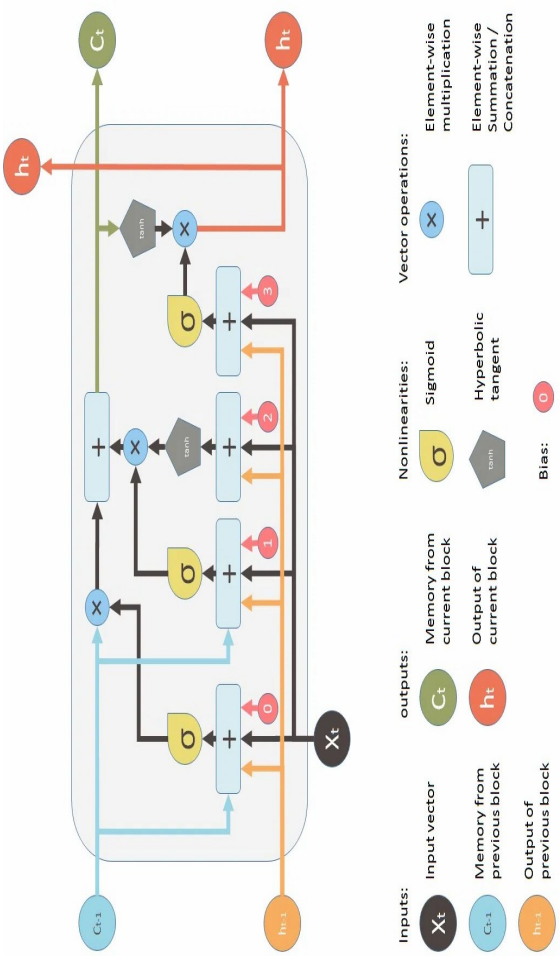


Figure 1: The Architecture of the LSTM Model [36]

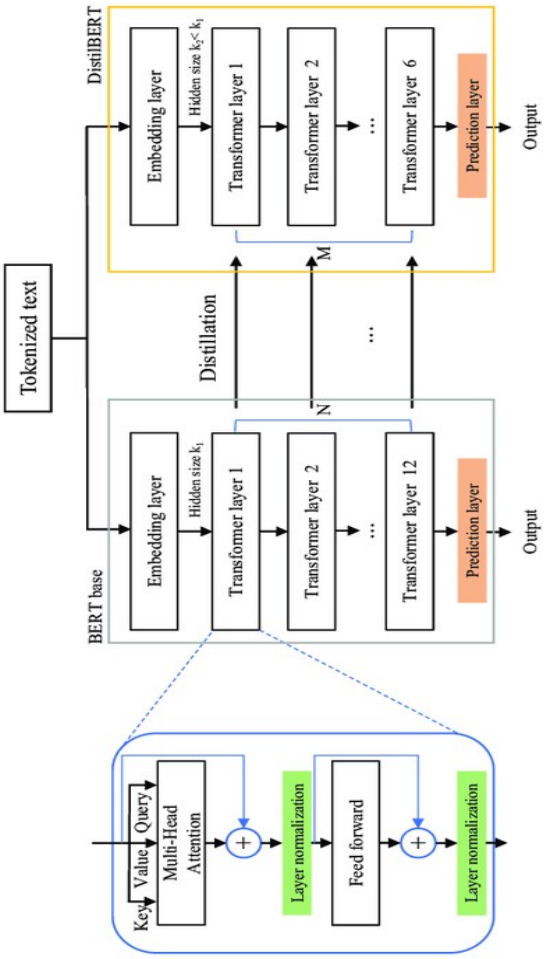


Figure 2: The Architecture of the DistilBERT Model [37]

5. METHODOLOGY

5.1 Data Collection Phase

This paper utilizes the Participants Data dataset, which contains 6364 reviews of products. The reviews are classified into four categories ranging from negative to strongly positive (0 = Negative, 1 = Neutral, 2 = Positive, and 3 = strongly positive). The dataset includes four attributes, namely Text ID, Product Description, Product Type, and Sentiment. Text ID and Product Type are excluded from the learning process as having no impact. The dataset is split into training and testing data using the Sklearn train test split function in Python, with the testing data constituting 20% of the total dataset.

5.2 Data Preprocessing Phase

The text preprocessing methods used to clean the data are punctuation and special characters removal from reviews as having no impact on the sentiment, Lower-casing to decrease the number of dimensions for words, stop words removal using NLTK stop words list in Python as having no impact on sentiment, tokenization to split sentence to tokens. The preprocessed data is the input for the next phase.

Text processing constitutes only a part of the overall methodology. As illustrated in Figure 3, there is a notable imbalance in the class labels, which impedes the ability of models to accurately discriminate between classes. Data imbalance refers to a scenario where certain classes are represented by many samples, while others have significantly fewer samples.

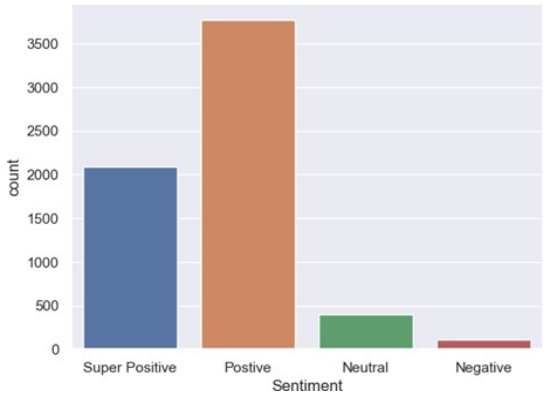


Figure 3: Illustration of Class Label Imbalance

Figure 3 illustrates the sentiment distribution within the analyzed dataset. The data exhibits a significant imbalance, with 'Positive' sentiments being the most prevalent at over 3500 instances. 'Super Positive' sentiments follow with approximately 2000 instances. 'Negative' and 'Neutral' sentiments are considerably less common, each recorded at just under 500 occurrences. This disproportionate distribution suggests a potential bias towards positive sentiment within the dataset.

To address the imbalance depicted in the dataset, The RandomOverSampler technique from the imbalanced-learn library is employed. RandomOverSampler works by randomly duplicating instances in the minority classes to balance the dataset. The process would involve fitting the sampler to the dataset and generating new samples until all sentiment classes have approximately the same number of instances. This approach can help mitigate the bias introduced by the imbalanced dataset and improve the performance of machine learning models trained on this data. Table 1 illustrates the Number of Reviews per Sentiment for each experiment before and after oversampling.

Table 1 Number of Reviews per Sentiment.

Label	Samples	Oversampling	Concatenation of S+ and +	3 Classes Oversampled
S+(3)	3765	3765		
P (2)	2089	3765	7530	7530
Neg. (1)	399	3765	3765	7530
Neu. (0)	111	3765	3765	7530
Total	6364	15,060	15,060	22,590

5.3 Word Embedding Phase

Diverse methodologies for word embedding were employed to enhance classical machine learning models. For models like Random Forest, XGBoost, and Support Vector Machines (SVM), a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer was utilized, supplemented with an N-Gram range of 1 to 3. In the context of deep learning algorithms, specifically for the DistilBERT model, a pre-trained tokenizer was implemented. Furthermore, for Long Short-Term Memory (LSTM) networks, the Global Vectors for Word Representation (GloVe) was applied, and the Word2vec method for the integrated model. The Word2vec method combined the Continuous Bag-of-Words model and the Skip-Gram model.

5.3.1 Techniques

1. TF-IDF Vectorizer: This method transforms text into a meaningful representation of numbers which is used to fit machine algorithms for prediction. It reflects how important a word is to a document in a collection or corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others [5].
2. N-Gram: An N-Gram model uses the probability of a word occurring based on the occurrence of its N - 1 previous words. This technique is crucial in statistical natural language processing and computational linguistics for predicting the next item in a sequence [38].
3. DistilBERT with Pre-trained Tokenizer: DistilBERT is a smaller, faster, cheaper, and lighter version of BERT. It uses a pre-trained tokenizer that converts input text into tokens that can be processed by the model. This approach significantly reduces the complexity and computational cost without substantial loss in performance, particularly in large-scale text data [16].
4. GloVe (Global Vectors for Word Representation): GloVe is an unsupervised learning algorithm for obtaining vector representations for words. It efficiently leverages the aggregate global word-word co-occurrence statistics from a corpus to generate word embeddings. This approach is advantageous as it allows the model to discern the probability of finding a word in a particular context, capturing the semantic relationships between words [4].

Each of these techniques contributes uniquely to the overall performance and accuracy of the respective machine learning and deep learning models.

5.4 The Integrated Model

This model, as shown in Figure 4, integrates a Convolutional Neural Network (CNN) and a Feedforward Neural Network (FFNN) to extract features from the collected reviews. Additionally, it employs the Mamdani Fuzzy System to classify the output generated by the FFNN model.

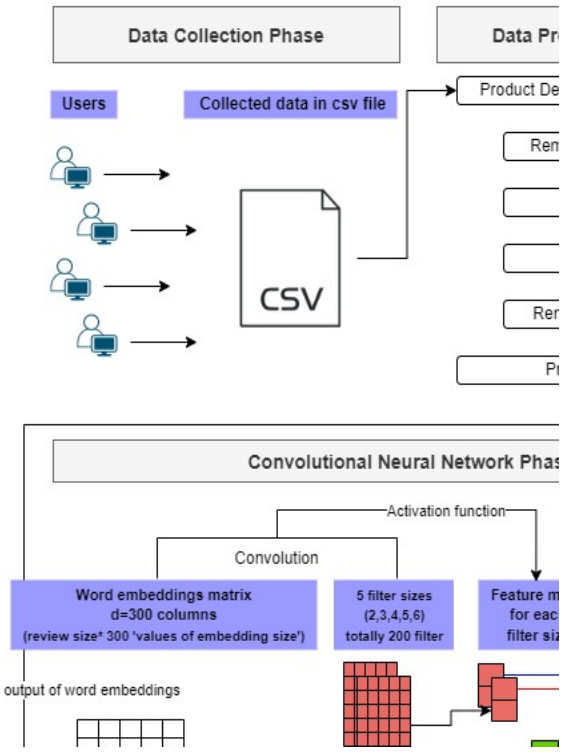


Figure 4: Illustration of the Integrated Model

5.4.1 Convolutional Neural Network (CNN) Phase

The CNN phase for feature extraction consists of 4 layers. The first layer is the embedding matrix which is the output of the previous phase. The second layer is the convolutional layer that produces a feature map by applying a filter matrix on the embedding matrix using the ReLU activation function. Many convolutional operations are conducted with many filters and many filter sizes (2,3,4,5,6) to produce many feature maps. The filter matrix width is the same as the embedding matrix so the created feature map is a one-dimensional vector. The third layer is a max pooling operation that reduces the size of each feature map to a single value by selecting the maximum value. In the Fourth layer, all feature maps are concatenated into one vector which is the output of the CNN phase and input for the FFNN phase.

5.4.2 Feedforward Neural Network (FFNN) Phase

The FFNN phase is responsible for producing positive sentiment value (PV) and negative sentiment value (NV). The output vector of the CNN phase is converted to a fully connected layer which is the first layer in the FFNN phase. The fully connected layer neurons represent the extracted features. The following 2 layers are sigmoid hidden layers. The third layer is the softmax output layer which produces PV and NV that range between 0 and 1. A dropout layer is used to overcome the overfitting problem.

5.4.3 Mamdani Fuzzy System Phase

The Mamdani fuzzy system phase comprises the classification phase, which accepts two input values from the previous phase, namely PV and NV. Although the CNN and FFNN methods are efficient in extracting features from a dataset, they encounter difficulties in handling ambiguous and vague data. Fuzzy logic is employed to enhance accuracy and efficiency. The Mamdani fuzzy system is composed of 4 steps:

- The definition of input and output linguistic variables, as well as the linguistic terms associated with each variable. The input variables are PV and NV, while the output variable is C = class. The linguistic terms for the input variables are very very low (0 to 1/6), very low (1/6 to 1/3), low (1/6 to 1/2), moderate (1/3 to 2/3), high (1/2 to 5/6), very high (2/3 to 1), very very high (5/6 to 1), and for the output variable are negative (0 to 0.4), neutral (0.4 to 0.6), positive (0.6 to 0.8), strongly positive (0.8 to 1).
- The process of fuzzification involves using a triangular membership function to calculate the membership degree of input linguistic variables to all input linguistic terms.
- The fuzzy inference system is composed of three distinct phases. Firstly, the application process involves the utilization of 49 IF-Then fuzzy rules that have been generated. Subsequently, the implication process computes the membership degree of the output class to each linguistic term based on the aforementioned application process. Finally, the aggregation process is responsible for aggregating each class label and determining the maximum membership degree for each class label.
- In the process of defuzzification, the output of the aggregation process is transformed into a real number through the use of a weighted average technique. The defuzzified value is then subjected to defuzzification rules to determine the final output class label.

LSTM and DistilBERT Performance

Both LSTM and DistilBERT models exhibited superior performance, especially in the 3-class experiment. Their high accuracy, precision, recall, and F1 scores indicate an excellent ability to model complex patterns in data. The 3-class experiment, post-oversampling, further solidified their robustness, showing minimal variance in performance metrics, a testament to their generalization capabilities.

Learning Curves Analysis

1. LSTM Learning Curve (Figure 5): The LSTM model displayed a steady improvement in learning, evidenced by the convergence of training and validation accuracy over 20 epochs. This suggests a well-tuned model with effective learning and generalization to unseen data. The absence of overfitting or underfitting is indicative of LSTM's capability to handle

temporal dependencies in data efficiently.

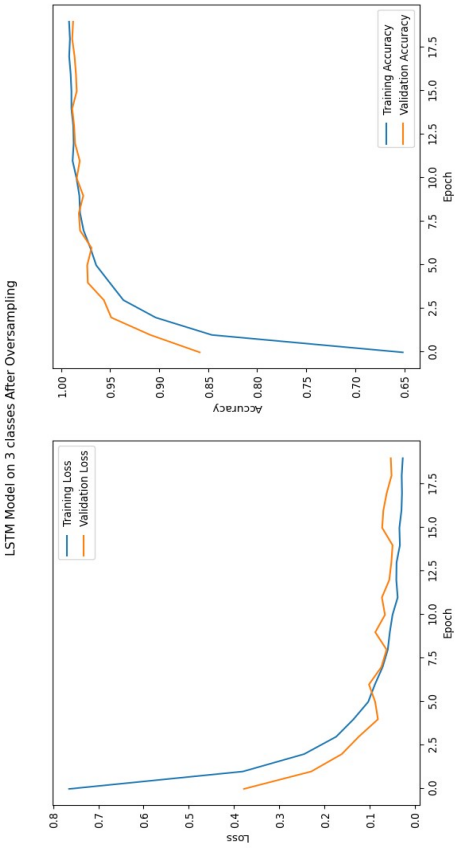


Figure 5: LSTM Model Learning Curve

2. DistilBERT Learning Curve (Figure 6): DistilBERT, known for its efficient adaptation of BERT architecture, showed rapid attainment of high accuracy in initial epochs, plateauing thereafter. This quick convergence demonstrates DistilBERT's efficiency in text representation and classification tasks. Its performance in the 4-class experiment, though slightly lower than in the 3-class, still outperforms other models, highlighting its robustness across different classification challenges with 50 epochs of training.

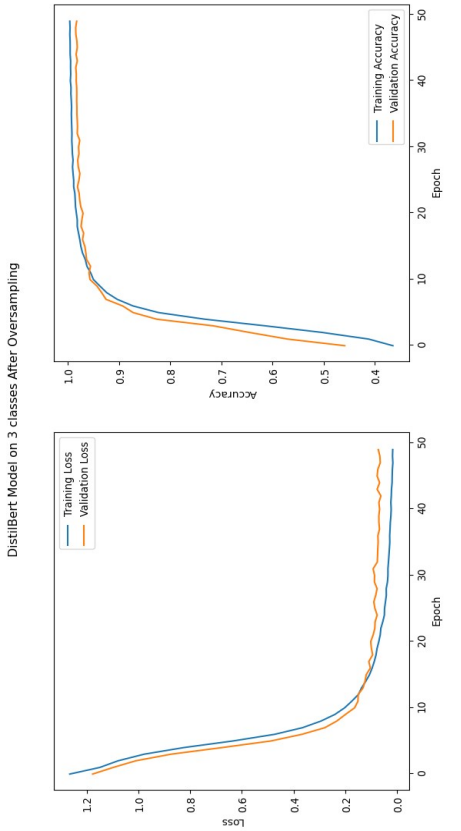
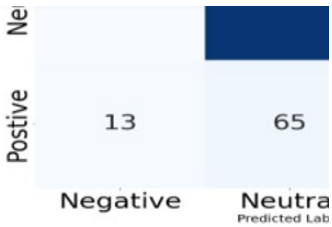


Figure 6: DistilBERT Model Learning Curve

6 RESULTS

The study compared the performance of the integrated model (CNN, FFNN, and fuzzy system), various machine learning and deep learning models (SVM, XGBoost, Random Forest, Logistic Regression, Multinomial Naive Bayes, LSTM, and DistilBERT) across two different experiments: one involving three classes and the other involving four classes. The K-fold cross-validation is applied with n=5 and the mean score is computed. The evaluation metrics included test accuracy, precision, recall, and the F1 score, defined as follows [39]:



Journal of Theoretical and Applied Information Technology
30th November 2024. Vol.102. No. 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The LSTM Model of both the 3-class and 4-class classifications Confusion matrices after oversampling are shown in Figure 7.

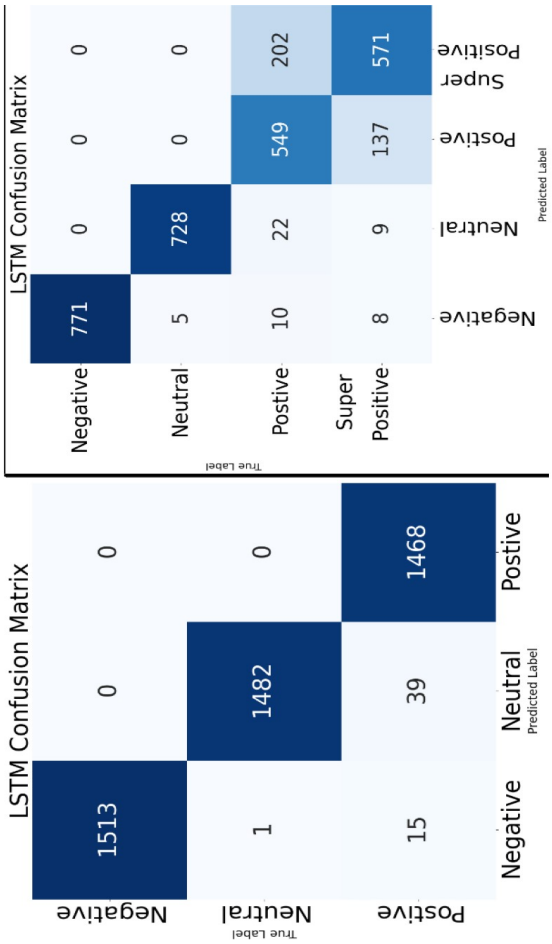


Figure 7: LSTM Model Confusion Matrices

The DistilBERT Model of both the 3-class and 4-class classifications Confusion matrices after oversampling are shown in Figure 8.

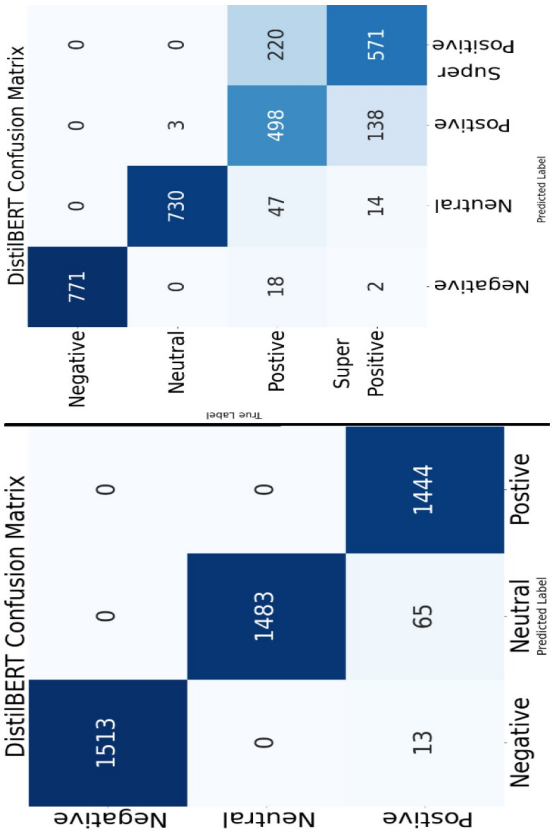


Figure 8: DistilBERT Model Confusion Matrices

The comparative performances of the models in both the 3-class and 4-class experiments are presented in Tables 2, 3, and 4 :

Table 2: Model Performance Comparison for 4-class Experiment after Oversampling.

Model (4-class)	Acc.	Prec.	Rec.	F1 Score
SVM	0.81	0.80	0.81	0.80
XGBoost	0.80	0.79	0.80	0.79
Random Forest	0.85	0.85	0.85	0.85
Logistic Regression	0.55	0.55	0.55	0.54
Multinomial Naive Bayes	0.49	0.49	0.49	0.48
LSTM	**0.87**	**0.87**	**0.87**	**0.87**
DistilBERT	0.85	0.85	0.85	0.85
Fuzzy Sys.	0.85	0.85	0.85	0.85

Table 3 Model Performance Comparison for 3-class Experiment.

Model (3-class)	Acc.	Prec.	Rec.	F1 Score
SVM	0.92	0.92	0.92	0.92
XGBoost	0.93	0.92	0.93	0.92
Random Forest	0.95	0.95	0.95	0.95
Logistic Regression	0.66	0.64	0.61	0.62
Multinomial Naive Bayes	0.60	0.52	0.53	0.60
LSTM	**0.98**	**0.98**	**0.99**	**0.98**
DistilBERT	0.96	0.97	0.96	0.96
Fuzzy Sys.	0.95	0.95	0.95	0.95

Table 4: Model Performance Comparison for 3-class Experiment after Oversampling.

Model (3-class, after oversampling)	Acc.	Prec.	Rec.	F1 Score
SVM	0.93	0.93	0.93	0.93
XGBoost	0.92	0.92	0.92	0.92
Random Forest	0.96	0.96	0.96	0.96
Logistic Regression	0.67	0.67	0.67	0.67
Multinomial Naive Bayes	0.59	0.59	0.59	0.58
LSTM	**0.99**	**0.99**	**0.99**	**0.99**
DistilBERT	0.98	0.98	0.98	0.98
Fuzzy Sys.	0.96	0.96	0.96	0.96

Both LSTM and DistilBERT models provide complementary strengths in sentiment analysis, with the LSTM capturing long-term dependencies and DistilBERT efficiently handling large-scale language modeling tasks.

Analyzing these results reveals several key insights, into the performance of different machine learning and deep learning models. The study's focus on both three-class and four-class experiments, with metrics such as accuracy, precision, recall, and F1 score, provides a comprehensive view of model efficacy.

Comparison with Other Models

- The integrated model and traditional machine learning models like SVM, XGBoost, and Random Forest showed commendable performance but were outperformed by LSTM and DistilBERT, especially in handling the more complex 4-class classification.
- Logistic Regression and Multinomial Naive Bayes struggled in both experiments, possibly due to their linear nature and the inability to capture complex data relationships as effectively as deep learning models.

This paper, along with the existing literature, demonstrates that deep learning models and transformers outperform traditional machine learning methods. In this study, the LSTM model achieved an accuracy of 99%, which is higher than the 96.5% accuracy reported for LSTM in previous sentiment analysis tasks. This improvement can be attributed to the techniques used in the preprocessing phase, the size, and the type of the dataset used. We employed oversampling techniques to address the imbalanced dataset, which significantly enhanced the results.

7 CONCLUSION

In conclusion, this study highlighted the importance of recommendation frameworks in artificial intelligence for aiding user decision-making in areas like purchasing, healthcare, and entertainment. It focused on sentiment analysis using eight classification models, including five machine learning models (XGBoost, Naïve Bayes, Logistic Regression, SVM, and Random Forest) and two deep learning models (LSTM and DistilBERT). A proposed model integrated CNNs with FFNNs and a Mamdani Fuzzy System. The results showed that the LSTM model excelled, especially with shorter sentences, slightly outperforming the DistilBERT model. A comparison between 3-class and 4-class classifications revealed LSTM's superiority, achieving an accuracy of 0.99, along with precision, recall, and F1 scores of 0.99, demonstrating its effectiveness in sentiment classification tasks.

The presented models illustrate the robust capabilities of AI in sentiment analysis. The LSTM model's superior performance in handling short sentences enhances the precision of sentiment classification, which is critical for accurate recommendations. These results affirm the viability of employing LSTM for concise user-generated content.

The analysis concluded that, particularly after oversampling, the LSTM model displayed commendable performance in the 3-class sentiment classification scenario. The findings underscore the model's robustness and adaptability to sentiment analysis, especially when the dataset is preprocessed to mitigate class imbalance. Also, the

findings proved that using conventional machine learning algorithms for sentiment analysis is inefficient due to their design, which involves a static mapping of the input vector to output vectors. So they can't adapt to the dynamic behaviors. This insight provides a significant contribution to the domain of AI-driven sentiment analysis, particularly in enhancing the predictive accuracy of sentiment classification models.

8 FUTURE WORK

Future research will explore the integration of more complex deep learning models and ensemble methods to further improve accuracy and generalization across diverse datasets. Additionally, expanding the classification to more nuanced sentiment categories could offer deeper insights into user sentiment. Investigation into the adaptability of these models in real-time applications will also be a focus.

STATEMENTS AND DECLARATIONS

I declare that the work in this paper was carried out following the Regulations of the Journal of Theoretical and Applied Information Technology. The work is original except where indicated by particular references in the text. The paper has not been submitted to any other journal.

Competing Interests. All the authors declare that they do not have any conflict of interest.

REFERENCES:

[1] Zhang Z, Guo J, Zhang H, Zhou L, Wang M. Product selection based on sentiment analysis of online reviews: An intuitionistic fuzzy TODIM method. Complex Intell Syst. 2022;8(4):3349–3362.

[2] Srikanth J, Damodaram A, Teekaraman Y, Kuppusamy R, Thelkar AR. Sentiment analysis on COVID-19 Twitter data streams using deep belief neural networks. Comput Intell Neurosci. 2022;2022.

[3] Maodah F, Utami E, Sudarmawan S. Optimizing sentiment analysis of product reviews on marketplace using a combination of preprocessing techniques, Word2Vec, and convolutional neural network. Jurnal Teknik Informatika (Jutif). 2023;4(1):101–107.

[4] Giri S, Das S, Das SB, Banerjee S, et al. SMS spam classification–simple deep learning models with higher accuracy using BUNOW and GloVe word embedding. J Appl Sci Eng. 2023;26(10):1501–1511.

[5] Das M, Alphonse P, et al. A comparative study on TF-IDF feature weighting method and its analysis using unstructured dataset. arXiv preprint arXiv:2308.04037. 2023.

[6] Gopi AP, Jyothi RNS, Narayana VL, Sandeep KS. Classification of tweets data based on polarity using improved RBF kernel of SVM. Int J Inf Technol. 2023;15(2):965–980.

[7] Paul B, Guchhait S, Dey T, Das Adhikary D, Bera S. A comparative study on sentiment analysis influencing word embedding using SVM and KNN. In: Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021. Springer; 2022. p. 199–211.

[8] Sharma S, Swarup M, Mahajan T, Patel ZD. Detecting anomalies, contradictions, and contextual analysis through NLP in text. In: 2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT). 2022. p. 1–5. https://doi.org/10.1109/ICICT55121.2022.10064560.

[9] Hidayat THJ, Ruldeviyani Y, Aditama AR, Madya GR, Nugraha AW, Adisaputra MW. Sentiment analysis of Twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier. Procedia Comput Sci. 2022;197:660–667.

[10] Suasnawa IW, Caturbawa IGN, Widharma IGS, Sapteka AANG, Indrayana INE, Sunaya IGAM. Twitter sentiment analysis on the implementation of online learning during the pandemic using Naive Bayes and support vector machine. 2023.

[11] Wang T, Bian Y, Zhang Y, Hou X. Classification of earthquakes, explosions and mining-induced earthquakes based on XGBoost algorithm. Comput Geosci. 2023;170:105242.

[12] Patel A, Oza P, Agrawal S. Sentiment analysis of customer feedback and reviews for airline services using language representation model. Procedia Comput Sci. 2023;218:2459–2467.

[13] Mostafa R, Mehedi MHK, Alam MM, Rasel AA. Bidirectional LSTM and NLP based sentiment analysis of tweets. In:

Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022). Springer; 2023. p. 647–655.

[14] Zulqarnain M, Alsaedi AKZ, Sheikh R, Javid I, Ahmad M, Ullah U. An improved deep neural network based on combination of GRU and auto encoder for sentiment analysis. 2023.

[15] Robertson S. NLP from scratch: Classifying names with a character-level RNN. PyTorch Tutorials. 2023;1(0).

[16] Robinson S. Classification of toxic comments based on textual data using deep learning algorithms. Available at SSRN 4609428.

[17] Dsouza DD, Deepika, Nayak DP, Machado EJ, D AN. Sentimental analysis of student feedback using machine learning techniques. Int J Recent Technol Eng (IJRTE). 2019;8:3384–3386.

[18] Devipriya K, Prabha D, Pirya V, Sudhakar S. Deep learning sentiment analysis for recommendations in social applications. 2020;9(4).

[19] Mohammed M, Yu L, Aldhubri A, Qaid GR. Study on sentiment classification strategies based on the fuzzy logic with crow search algorithm. 2022.

[20] Mounika A, Saraswathi S. Design of book recommendation system using sentiment analysis. In: Evolutionary Computing and Mobile Sustainable Networks: Proceedings of ICECMSN 2020. Springer; 2021. p. 95–101.

[21] Hung BT. Integrating sentiment analysis in recommender systems. Reliability and Statistical Computing: Modeling, Methods and Applications. 2020. p. 127–137.

[22] Wang Y, Wang M, Xu W, et al. A sentiment-enhanced hybrid recommender system for movie recommendation: a big data analytics framework. Wireless Commun Mobile Comput. 2018.

[23] Lo SL, Cambria E, Chiong R, Cornforth D. Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artif Intell Rev. 2017;48:499–527.

[24] Li J, Yang Y. Recommender systems based on opinion mining and deep neural networks. In: MATEC Web of Conferences. EDP Sciences; 2018. p. 173:03016.

[25] Srinivasarao U, Sharaff A. Machine intelligence based hybrid classifier for spam detection and sentiment analysis of SMS messages. Multimedia Tools Appl. 2023;1–31.

[26] Do PMT, Nguyen TTS. Semantic-enhanced neural collaborative filtering models in recommender systems. Knowl-Based Syst. 2022;257:109934.

[27] Khanvilkar G, Vora D. Sentiment analysis for product recommendation using random forest. Int J Eng Technol. 2018;7(3):87–89.

[28] Karnan T, Seenuvasan G. Sentiment analysis in e-commerce using recommendation system. In: Proc. International Journal of Computer Science and Mobile Computing. 2017;6:55–62.

[29] Antonio VD, Efendi S, Mawengkang H. Sentiment analysis for COVID-19 in Indonesia on Twitter with TF-IDF featured extraction and stochastic gradient descent. Int J Nonlinear Anal Appl. 2022;13(1):1367–1373.

[30] Bhargav PS, Reddy GN, Chand RR, Pujitha K, Mathur A. Sentiment analysis for hotel rating using machine learning algorithms. Int J Innovative Technol Explor Eng (IJITEE). 2019;8(6):1225–1228.

[31] Devi S, Kumar SB, Keerthivel S, Ranjith S. E-commerce product reviews using sentimental analysis. 2020.

[32] Salur MU, Aydin I. A novel hybrid deep learning model for sentiment classification. IEEE Access. 2020;8:58080–58093.

[33] Suresh Kumar, K., Radha Mani, A. S., Ananth Kumar, T., Jalili, A., Gheisari, M., Malik, Y., … Jahangir Moshayedi, A. (2024). Sentiment Analysis of Short Texts Using SVMs and VSMs-Based Multiclass Semantic Classification. Applied Artificial Intelligence, 38(1). https://doi.org/10.1080/08839514.2024.2321555.

[34] Dash, D. P., Kolekar, M., Chakraborty, C., & Khosravi, M. R. (2024). Review of machine and deep learning techniques in epileptic seizure detection using physiological signals and sentiment analysis. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), 1-29.

[35] Mozumder, M. A. S., Nguyen, T. N., Devi, S., Arif, M., Ahmed, M. P., Ahmed, E., ... & Uddin, A. (2024). Enhancing Customer Satisfaction Analysis Using Advanced Machine Learning Techniques in Fintech Industry. Journal of Computer Science and Technology Studies, 6(3), 35-41.

[36] Aljedaani W, Abuhaimed I, Rustam F, et al. Automatically detecting and understanding the perception of COVID-19 vaccination: a Middle East case study. Soc Netw Anal Min. 2022;12(128). https://doi.org/10.1007/s13278-022-00946-0.

[37] Adel H, Dahou A, Mabrouk A, Abd Elaziz M, Kayed M, El-Henawy IM, Alshathri S, Amin Ali A. Improving crisis events detection using DistilBERT with hunger games search algorithm. Mathematics. 2022;10(3):447.

[38] Purbaya ME, Rakhmadani DP, Arum MP, Nasifah LZ, et al. Implementation of n-gram methodology to analyze sentiment reviews for Indonesian chips purchases in Shopee e-marketplace. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi). 2023;7(3):609–617.

[39] Salauddin Khan M, Nath TD, Murad Hossain M, Mukherjee A, Bin Hasnath H, Manhaz Meem T, Khan U. Comparison of multiclass classification techniques using dry bean dataset. Int J Cognitive Comput Eng. 2023;4:6–20. https://doi.org/10.1016/j.ijcce.2023.01.002.