International Journal of Advanced and Applied Sciences, 11(8) 2024, Pages: 146-157

Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Sentiment analysis of movie review classifications using deep learning approaches

Sarwar Shah Khan [1, 2, *], Yasser Alharbi [3]

[1]*Department of Computer and Software Technology, University of Swat, Swat, Pakistan*
[2]*Department of Computer Science, IQRA National University, Swat, Pakistan*
[3]*College of Computer Science and Engineering, University of Hail, Hail, Saudi Arabia*

ARTICLE INFO

Article history:
Received 7 April 2024
Received in revised form
9 August 2024
Accepted 18 August 2024

Keywords:
Sentiment analysis
Deep learning models
XLNet
Rotten Tomatoes dataset
Movie reviews

ABSTRACT

Movie reviews reflect how the public feels about a movie they have watched. However, because many reviews are posted on various websites, it is practically impossible to read each one. Summarizing all movie reviews can help people make informed decisions without reading through all of them. Previous studies have used different machine learning and deep learning techniques for sentiment analysis (SA), but few have combined comprehensive hyperparameter tuning and novel datasets for better performance. This paper presents an SA approach using deep learning models with optimized hyperparameters and a novel Rotten Tomatoes (RT) dataset to help viewers make better movie choices. SA, or opinion mining, is a computational technique to extract and analyze opinions and emotions expressed in text. We explore deep learning models such as Long Short-Term Memory (LSTM), XLNet, Convolutional Neural Networks-LSTM (CNN-LSTM), and Bidirectional Encoder Representations from Transformers (BERT). These models are known for capturing complex language patterns and context from raw text data. XLNet, a pre-trained model, effectively understands context by considering all possible permutations of the input sequence, BERT excels at using bidirectional context to understand text, LSTM retains information about long-term patterns in sequential data, and CNN-LSTM combines local and global context for reliable feature extraction. The RT dataset was pre-processed with data cleaning, spelling correction, lemmatization, and handling of informal words to improve the results. Our experiments show that XLNet performed better than other models on the Rotten Tomatoes dataset. The study demonstrates that SA of movie reviews provides insights into emotions and attitudes, allowing us to estimate a movie's performance based on its overall sentiment.

1. Introduction

Movies can be a valuable part of our lives and positively impact our personal growth (Rahman and Hossen, 2019). The film industry produces an increased number of movies and sells more tickets annually to boost their earnings. From a viewer's perspective, the cost of watching a movie includes both the price of the ticket and the time spent watching it. Trailers can sometimes be misleading, leading viewers to waste their time and money on movies that are not as good as advertised (Banik and Rahman, 2018). Movie reviews can be a valuable resource for movie lovers. They can help you decide what movies to watch, learn about cinema, and connect with other movie fans. Because there are so many reviews on movie review websites, it can be hard for people who have never seen a movie before to decide which one to watch. It might be challenging to read through all of the lengthy and in-depth movie reviews in a short period of time (Danyal et al., 2023). The summary of all movie reviews into positive and negative categories can save people time reading through them all. This is achievable through sentiment analysis (SA) (Danyal et al., 2024a). SA or opinion extraction is an approach that finds and gathers emotions from source materials by using natural language processing (NLP), machine learning (ML), and text analysis techniques (Khan et al., 2018; Danyal et al., 2024b).

* Corresponding Author.
Email Address: sskhan0092@gmail.com (S. S. Khan)
https://doi.org/10.21833/ijaas.2024.08.016
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-6387-4114

Sentiment mining enables businesses to gain insights from customer feedback, enhance decision-making, and improve overall customer experience (Khan et al., 2020). SA categorizes movie reviews as positive or negative based on the occurrences of words in the review text. Words that have previously been used in a positive or negative context are used to train the SA model. The better the sentiment extraction model is trained, the better it will be able to understand the review process and the needs of viewers (Yasen and Tedmori, 2019; Danyal et al., 2024c). Deep learning (DL) is gaining popularity as an efficient machine learning (ML) technique that learns a number of layers of data representations or capabilities and generates better results for prediction. In recent years, SA has extensively used deep learning, resulting in its success in numerous other applications (Zhang et al., 2018). DL models can learn complex relationships among phrases and specific words. This enables them to properly understand all aspects of human language and make accurate predictions within the sentiment of the text.

XLNet is a language model that integrates BERT's bidirectionality with Transformer-XL's autoregressive property. This enhances its ability to recognize long-term dependencies in text (Yang et al., 2019). BERT is a language model that has been pre-trained on an enormous text dataset along with code. It can be customized for several NLP tasks, such as text classification, query responding, and summarizing (Devlin et al., 2018). LSTM is a neural network type that improves with sequential data like text. It is appropriate for various natural language processing tasks, such as language modeling, SA, and text summarizing (Sherstinsky, 2020). CNN-LSTMs improve natural-language processing applications by combining CNN's local feature extraction and LSTM's long-term learning capabilities.

This study is based on the performance of the proposed technique for models such as XLNet, BERT, and LSTM. Our main objective is to make measurable improvements over the existing state-of-the-art models, with a special focus on the higher accuracy that the suggested methodology achieves. To pre-process the dataset and carefully select the hyperparameter settings to achieve higher results. The dataset utilized is the Rotten Tomatoes Dataset, which has been through multiple pre-processing stages to improve its quality. These procedures include lemmatization, case normalization, spelling correction, and data cleaning. In data cleaning, I removed duplicate reviews, hashtags, stop words, punctuation, links, and special characters. We made the following contributions to this study:

- To improve our results, we increased our preprocessing efforts on the dataset. This includes multiple processes to improve the data before analysis. First, we fix any spelling mistakes to ensure consistency and accuracy. Next, we simplify words to their base form using lemmatization, which improves analysis by reducing word variations. We then remove common words like "and," "the," and so on, which are known as stop words because they often add little meaning to the analysis. In addition, we handle special chat language or abbreviations to ensure they are properly interpreted. Removing duplicate reviews helps to streamline the dataset and ensures that we are not analyzing redundant information. Finally, we perform general data cleaning to remove any inconsistencies, special characters, or errors, ensuring that our dataset is ready for effective analysis.
- To improve the model's performance, various hyperparameters are fine-tuned. These include parameters such as the maximum sequence length, learning rate, batch size, and epoch count. By carefully adjusting these hyperparameters, the model can improve its performance in tasks like training accuracy, generalization, and convergence speed. The continuous procedure of hyperparameter tuning is vital for increasing the model's effectiveness across different datasets and tasks.
- The performance of XLNet, BERT, LSTM, and CNN-LSTM is assessed through a comparative evaluation.
- Using DL techniques like CNN-LSTM, LSTM, XLNET, and BERT for SA on the Rotten Tomatoes English Movie reviews dataset is a novel application that contributes to the exploration of these models in the context of movie reviews.

The structure of this paper is as follows: Section 2 presents the literature review, followed by the proposed methodology in Section 3. The experimental results are described in Section 5. Section 6 discusses the limitations of the study, and the conclusion is provided in Section 7.

2. Literature review

This section provides a comprehensive review of previous efforts in SA for movie reviews. It explores various techniques used in SA, with the goal of simplifying the evaluation of movie critiques. These methods include a range of approaches designed to extract detailed sentiments from text-based movie reviews. Table 1 offers a summary of the key findings and insights from previous studies in this area.

The research articles cited in the literature review describe various approaches to SA. These include methods for preprocessing, extraction of features, representation, and classification, as well as approaches for dealing with imbalanced datasets and pre-training large language models. These articles demonstrate the benefits of combining methods to achieve optimal performance in sentiment mining of movie reviews. The complexity and multiple emotion analysis, combined with long-distance and local semantic processing of data, presented the most significant challenges. Our method enhances the results by presenting a cleaner movie reviews dataset to XLNet, BERT, LSTM, and

CNN-LSTM. We chose the better-performing hyper-parameters and performed an in-depth comparative evaluation of all models.

Table 1: Summary of literature review

Reference	Techniques	Preprocessing techniques	Advantages and disadvantages	Evaluation measures	Datasets
Dhivyaa et al. (2023)	BERT, XLNet, transformer coder SVM	Data cleaning, tokenization	Enhances model performance and difficulty in handling subjective content	Accuracy, precision, recall, F1 score, loss	IMDb
Dashtipour et al. (2021)	LSTM, SVM, MLP, logistic regression, CNN	Text cleaning, lemmatization, tokenization, parts of speech tagging	Novel context-aware approach enhances understanding, difficult to implement	Accuracy, precision, recall, F1 score	Persian
Chakraborty et al. (2018)	K-means algorithm, Word2vec	Data cleaning	Scalability, automatic feature extraction, algorithm complexity	Time complexity	IMDb
Dholpuria et al. (2018)	Naïve bayes, SVM, logistic regression, KNN, CNN, ensemble methods	Data cleaning	Robust feature extraction, complexity, and resource-intensive	Accuracy, precision, recall, F1 score	IMDb
Dang et al. (2020)	RNN, CNN, DNN, word embedding TF-IDF	Data cleaning	Improved efficiency, applicable to various datasets	Accuracy, precision, recall, AUC, time	IMDb, Cornell, book reviews, tweets, sentiment
Lou (2023)	CNN, TF-IDF, bag of words	Data cleaning, normalization	Improved CNN and comparison of TF-IDF and count vectorizer	Accuracy, precision, recall, F1 score	IMDb
Ullah et al. (2022)	1D-CNN	Data cleaning, lemmatization, case normalization	Used state-of-the-art techniques, Requires substantial data, computationally intensive	Accuracy, precision, recall, F1 score	IMDb, binary classification
Tripathy et al. (2023)	ANN, genetic algorithm	Data cleaning	Enhanced feature representation, increased complexity	Accuracy, precision, recall, F1 score	IMDb
Abimanyu et al. (2023)	Logistic regression technique and information gain, feature selection	Lemmatization, tokenization, data cleaning, stemming,	Shows a decrease in performance as stemming and lemmatization perform better	Precision, recall, F1 score	Rotten tomatoes
Aziz et al. (2023)	Logistic regression, SVM, Naïve Bayes	Lemmatization, data cleaning, tokenization	Streamlined SA process, sensitivity to noise or irrelevant features in reviews	Accuracy, precision, recall, F1 score, training time	Rotten tomatoes
Palomo et al. (2024)	BERT, RoBERTa, XLNet, TF-IDF	Data cleaning	ML automates and expedites movie review analysis, complex transformer-based models	Accuracy, precision, recall, F1 score	IMDb
Deepa et al. (2021)	XLNet, BERT, RNN	-	Auto feature extraction and process larger data, discrepancies in pre-train-fine-tune approaches may impact model performance	Accuracy	IMDb and Coursera dataset

3. Proposed methodology

The proposed approach comprises five main stages: Data collection, data preprocessing, division into training and testing sets, optimization of model hyper-parameters, and deployment of XLNet, CNN-LSTM, LSTM, and BERT models, followed by performance evaluation. The research methodology is shown in Fig. 1. The first step is to collect data from fifty thousand reviews from the Rotten Tomatoes Movie Reviews dataset. In the second phase, the dataset is processed and cleaned in order to prepare it for modeling. In the third phase, the data is split into training and testing sets, allocating 75% for training and 25% for testing purposes. The fourth stage involves optimizing hyper-parameters to improve model performance. Finally, in the fifth phase, the models' performance is measured using a number of metrics to determine their effectiveness.

3.1. Data preprocessing

The first step of the proposed approach is preprocessing. This involves preparing the RT dataset for use by machine learning algorithms, such as label encoding, removing duplicate reviews, and case normalization. Label encoding is an approach to transforming categorical data into numerical data. This is done so that algorithms for machine learning can process the data. Case normalization is the conversion of all text in a dataset to lowercase. This ensures that the algorithms do not treat different cases of the same word as different entities. Duplicate Reviews removal is the process of removing duplicate reviews from a dataset. This ensures that the algorithms are not trained on the same data multiple times (Danyal et al., 2023).

3.2. Test train split

After preprocessing the data, the dataset is divided into two parts: the training set and the test set. The training set is used to train the algorithm, while the test set is used to evaluate its performance. In this experiment, 75% of the data is allocated for training, and 25% is reserved for testing (Fig. 2).

Dividing the reviews into training and testing sets (75% for training, 25% for testing) provides a balanced method for developing the model. The larger training set allows the model to learn patterns and features from a substantial amount of data, improving its ability to make accurate predictions.

3.3. Hyper-parameters tuning

Hyper-parameters are configuration settings that guide a machine learning algorithm's training process. Unlike model parameters, which are learned from the training data, hyper-parameters are set manually prior to training. They act as higher-level instructions, shaping the algorithm's behavior and performance (Sherstinsky, 2020). The specific hyper-parameters tuned for this experiment are detailed in Tables 2, 3, 4, and 5.

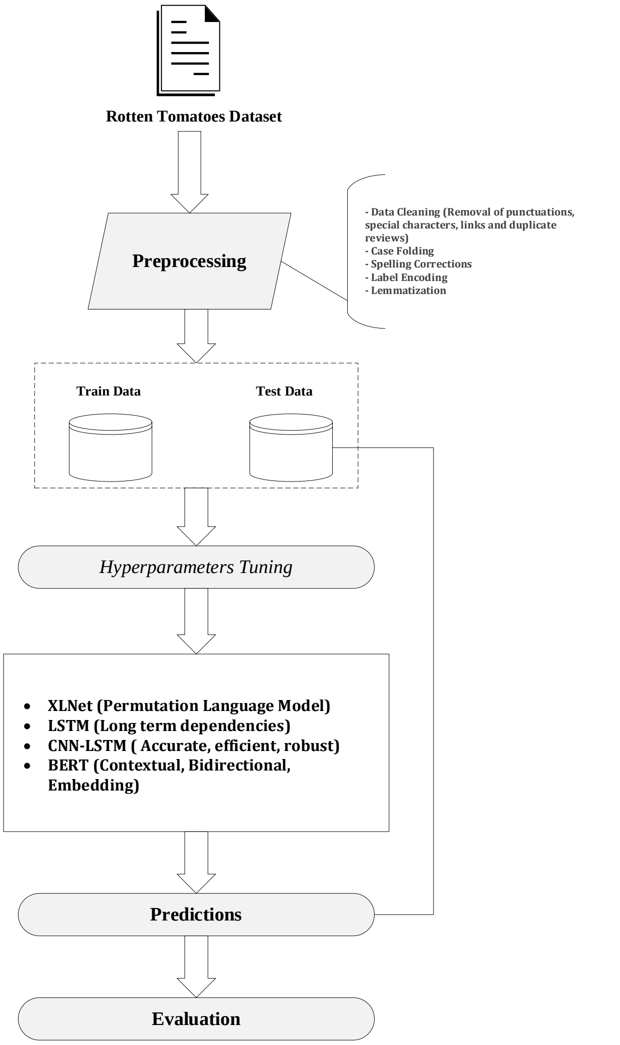

Fig. 1: Proposed methodology

Fig. 2: Train test split

Table 2: XLNet parameters

Parameters	Value
Number of epoch	3
Maximum sequence length	128
Batch size	128
Learning rate	0.0001

Table 3: BERT parameters

Parameters	Value
Number of epoch	3
Maximum sequence length	128
Batch size	128
Learning rate	0.0003

Table 4: CNN-LSTM parameters

Parameters	Value
Number of epoch	3
Embedding layer (output dimensions)	128
Dropout	0.5
Batch size	128
MaxPooling1D (pool_size)	2

Table 5: LSTM parameters

Parameters	Value
Maximum sequence length	128
Batch size	128
Embedding dimensions	128
LSTM units	128
LSTM dropout	0.2
LSTM recurrent_dropout	0.2
Number of epochs	3

Choosing the right hyper-parameters is important for building an effective machine-learning model. Their impact on the model's capacity to generalize and generate accurate predictions is significant. Finding the optimal settings is essential, a process often involving careful experimentation with different combinations of values to maximize model performance (Agrawal, 2021).

3.4. Techniques

In this study, deep learning models like XLNet, LSTM, and CNN-LSTM were employed to capture the sentiment expressed in movie reviews. Deep learning models offer numerous advantages, such as automatic feature extraction, the ability to handle complex data, enhanced performance, and the capability to process non-linear, structured, or unstructured data (Yasen and Tedmori 2019). XLNet and BERT were chosen for their advanced learning capabilities, outperforming other techniques in accuracy and efficiency. LSTMs excel in tasks involving sequences, as they are adept at learning long-term dependencies and can effectively handle inputs of varying lengths. CNN-LSTMs leverage the strengths of both CNNs, which excel in local feature extraction, and LSTMs, which are proficient in long-term learning, to enhance the performance of natural language processing applications. These models make significant contributions to a variety of fields thanks to their powerful capabilities and architectural designs. More details are provided below.

3.4.1. XLNet

XLNet is a modified autoregressive (AR) pre-training approach that combines the benefits of AR and Autoencoder approaches with the goal of permutation language modeling. XLNet's neural architecture is designed to function in tandem with the AR goal, which involves the incorporation of Transformer-XL and carefully designing the two-stream attention system. It was proposed by Li et al. (2019), and it outperforms other models in various NLP tasks, including question answering, natural language inference (NLI), and opinion extraction (Yang et al., 2019). XLNet classifies movie reviews using a permutation-based training technique that randomly orders tokens in a phrase. Pre-training and fine-tuning comprise this model. XLNet pretrains to predict the next word in a phrase permutation by viewing all the words but without knowing which one comes next. It handles long movie reviews well, thanks to Transformer-XL's design. XLNet improves the pre-trained model for sentiment categorization during fine-tuning. The model predicts “positive” or “negative” from a complete review using contextual representations of the text. Thus, XLNet can categorize movie review sentiment using permutation-based training, two-step training, and a fine-tuning approach. The XLNet architecture for movie reviews in Fig. 3 tokenizes input text into sub-words, adds positional embeddings for sequence order, and uses random permutation to eliminate left-to-right context dependency. Transformer layers with self-attention and feed-forward sub-layers handle the permuted sequence. An inverse permutation layer restores the original order, while the transformer layers' hidden states predict the next token, similar to traditional language models.

3.4.2. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a bidirectional model, which means it can understand the meaning of words in a sentence based on both the words that come before and after them. This makes it stronger than previous language models, which could only understand the meaning of words based on the words that came before them (Devlin et al., 2018). BERT is trained on an immense amount of text and code and is appropriate for various NLP tasks such as question answering, sentiment extraction, and NLI. BERT, a transformer-based machine learning model, processes text data for tasks like movie review classification. It starts by tokenizing the review into pieces and embedding these into a high-dimensional space. These embeddings are then processed through stacked transformer layers, capturing bidirectional context for each token. The final representation is fed into a classification layer trained to predict sentiments, providing the final output. The BERT architecture is shown in Fig. 4.

3.4.3. LSTM

Long-term memory (LSTM) is a recurrent neural network (RNN) version capable of learning long-term dependencies. As a result, it is well-suited to tasks like machine translation, speech recognition, and text generation. LSTMs use a gating mechanism to regulate the flow of data into and out of the cell state. This enables them to retain information for extended periods of time, even if it is unrelated to the current task (Van Houdt et al., 2020).

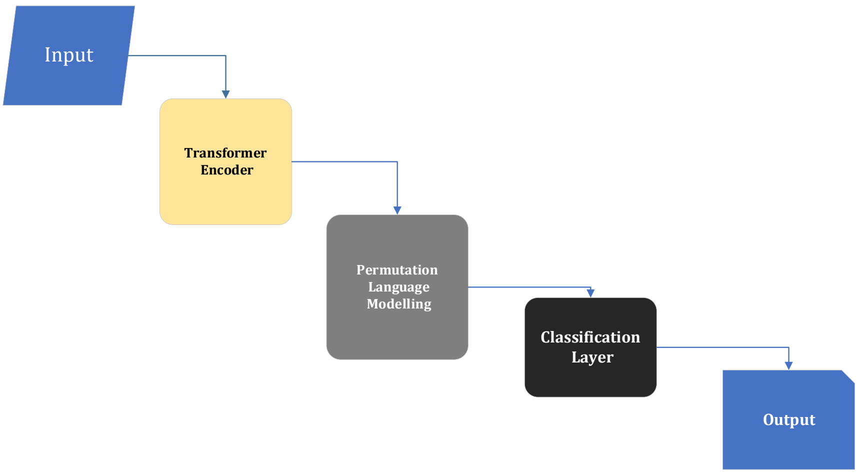

Fig. 3: XLNet architecture

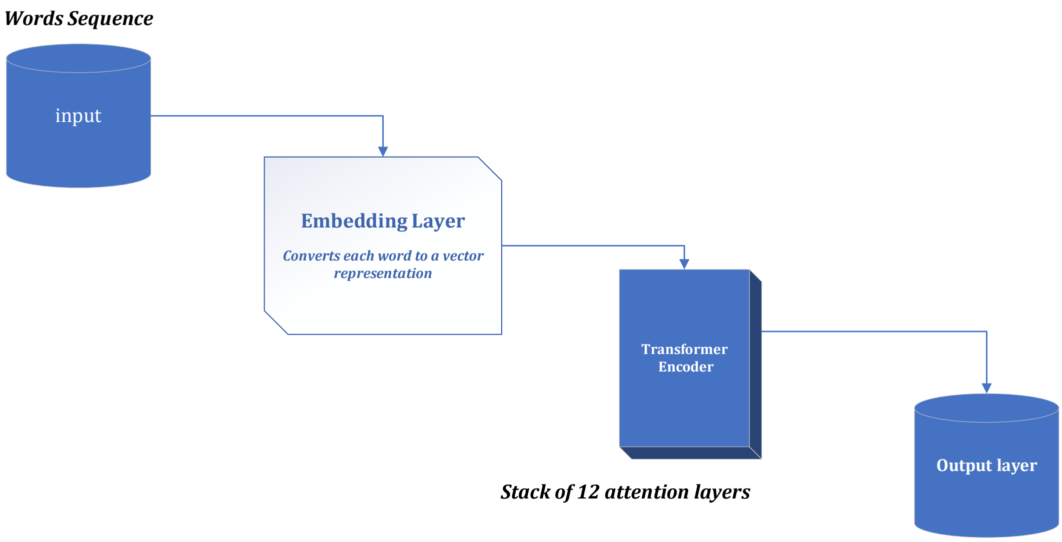

Fig. 4: BERT architecture

The model in Fig. 5 implements six layers, beginning with an embedding layer for converting input sequences into dense vectors. A dropout layer is then applied to prevent overfitting. Two LSTM layers process the sequential data, with the second one producing the final output. A spatial dropout layer helps regularize the LSTM output. The model concludes with a dense layer for classification using a sigmoid activation function. Together, these layers form a comprehensive deep-learning model capable of learning and making predictions.

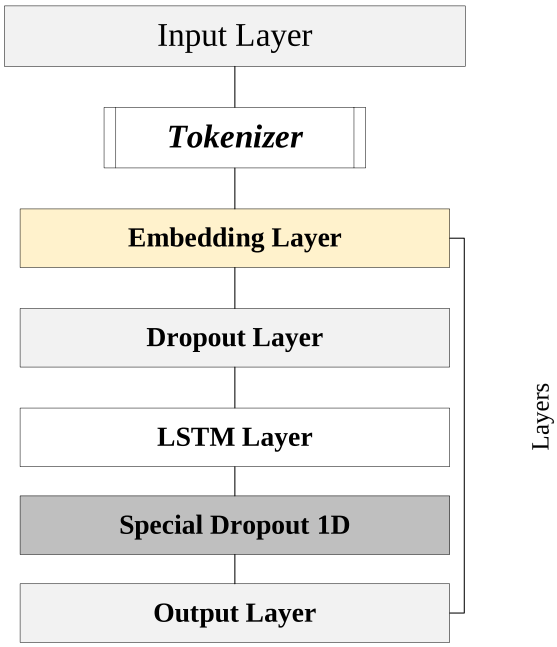

Fig. 5: LSTM architecture

3.4.4. CNN-LSTM

CNN-LSTM represents a sophisticated deep learning architecture that merges the strengths of convolutional neural networks (CNNs) with long short-term memory networks (LSTMs). While CNNs excel at capturing spatial information from data, LSTMs are particularly skilled at recognizing and understanding long-term dependencies. This fusion of CNNs and LSTMs in the CNN-LSTM model proves highly effective across a range of tasks, including natural language processing (NLP), speech recognition, and image classification (Mutegeki and Han, 2020).

In Fig. 6, we observe a visualization of the CNN-LSTM hybrid model architecture, beginning with an input layer that defines the input data's structure. Subsequently, an embedding layer is utilized to transform the input integer sequences into dense vector representations. Following this, the convolutional layer applies filters to capture local patterns within the data, while the max pooling layer serves to reduce dimensionality. To mitigate overfitting, a dropout layer is incorporated, and the LSTM layer is introduced to discern sequential dependencies within the data. Ultimately, the output layer employs a sigmoid activation function to generate the predicted sentiment of the input sequence. This model effectively amalgamates the feature extraction capabilities inherent in CNNs with the sequential understanding prowess of LSTMs, rendering it well-suited for SA tasks.

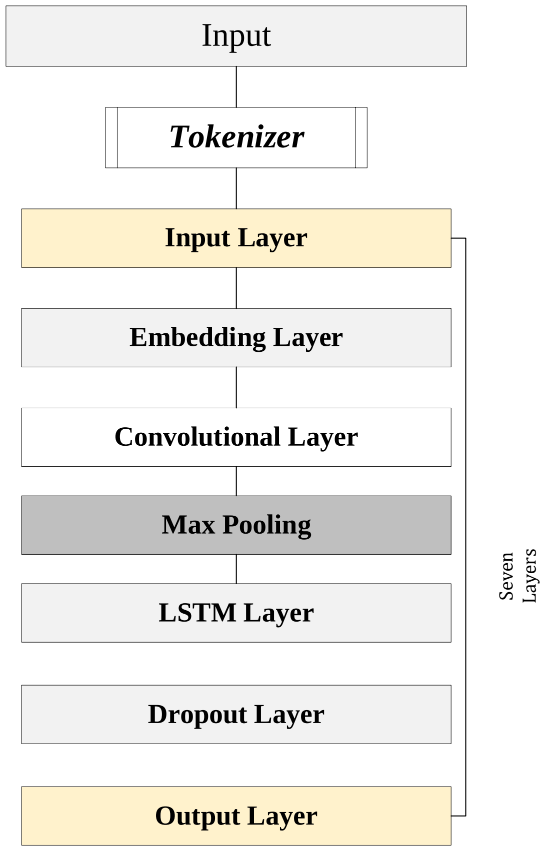

Fig. 6: CNN-LSTM architecture

3.5. Performance metrics

In this research study, we assess performance using metrics such as accuracy, F1- score, recall, precision, and evaluation time (seconds). The efficiency of the proposed model is determined through a confusion matrix, where TP represents true positives (correctly predicted positives), FN stands for false negatives (incorrectly predicted negatives), FP denotes false positives (incorrectly predicted positives), and TN represents true negatives (correctly predicted negatives). Here are further details.

The accuracy of the predictions is calculated by dividing the predicted number of reviews by the total number of reviews (Khan et al., 2023).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision is calculated by dividing the number of accurately predicted positive reviews by the total number of expected positive reviews.

$$Precision=\frac{TP}{TP+FP} \quad (2)$$

Recall is calculated by dividing the number of accurately predicted positive reviews by the total number of positive reviews.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The F1-measure, commonly known as the f-score or f-measure, measures an algorithm's performance by taking precision and recall into account. The F1-measure equation can be seen below.

$$F1\text{-}Score = \frac{2*Precision*recall}{Precision+recall} \quad (4)$$

The training time refers to the total duration of all steps performed in each epoch across the model.

4. Experimentation setup

The research experiment was carried out on a system equipped with an 11th-generation Intel Core i5 processor, 16GB of RAM, and the Jupyter Lab environment via Anaconda, which uses the Python programming language.

4.1. Dataset description

Movie review datasets can be found from various sources, including open-source platforms such as Kaggle, Rotten Tomatoes, the UCI Machine Learning Repository, and the IMDb website. For this study, movie reviews were extracted from the Rotten Tomatoes (RT) Dataset. A total of 50,000 reviews were randomly selected for experimentation (Leone, 2020). Fig. 7 provides a visual representation of the RT dataset.

	review	sentiment
0	Just as director Antoine Fuqua start to clos ...	rotten
1	... A film that could have been a bold feminis...	rotten
2	It's comforting, really, to see the movies s...	fresh
3	Wild is too artistically tame to wild abo...	fresh
4	By casting its villain is suspiciously ISIS...	rotten
...	...	...
49995	Arevalo's first feature film is a tense story...	fresh
49996	Sure, Budapest essential pastiche of the b...	fresh
49997	Make no mistake, this Toy story trilogy, and ...	fresh
49998	The disjunction between what's being said and...	rotten
49999	The film characters are stick figure artic...	rotten

50000 rows x 2 column

Fig. 7: Rotten tomatoes dataset

The review distribution in the Rotten Tomatoes dataset is shown in Fig. 8. Positive reviews in the RT dataset are referred to as fresh, while negative reviews are referred to as rotten, as shown in Fig. 8.

5. Results and discussion

The experiment results are discussed in this section. We tested XLNet, BERT, CNN-LSTM, and LSTM models over three epochs. XLNet has the highest accuracy of 87.68% in the third epoch. BERT also performed well, with an accuracy of 82.24% in the third epoch, which was better than LSTM, which had an accuracy of 75.16% in the third epoch. It showed progress over each epoch. CNN-LSTM also performed better than LSTM, with an accuracy of 76.98%. The bold results indicate that the models achieve the highest performance in their respective metrics. The performance of XLNet on the RT Dataset is shown in Table 6 and visualized in the bar chart in Fig. 9.

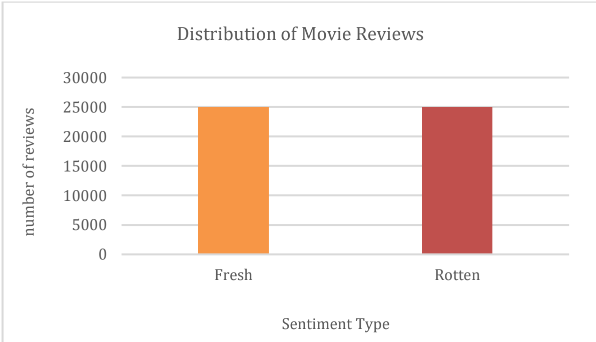

Fig. 8: Reviews distribution in rotten tomatoes dataset

Table 6: Performance of XLNet on rotten tomatoes dataset

Model	XLNet		
Epochs	1	2	3
Precision	83.71%	85.29%	86.18%
Recall	85.29%	86.01%	85.72%
F1-Score	84.49%	85.65%	85.95%
Accuracy	86.05%	87.16%	87.68%
Training time	918s	924s	949s

XLNet has the highest accuracy because it is trained on a larger dataset, and XLNet uses a technique called" permutation language modeling," which allows it to learn long-range dependencies between words. This makes it better at understanding the context of a sentence, which is important for tasks like SA and natural language inference. BERT also performed well, with an accuracy of 82.24%, as it is a well-established model that has been shown to be effective for various tasks. The rest of the results are presented in Table 7. The performance of BERT on the RT dataset is visualized in Fig. 10.

Table 7: Performance of BERT on RT dataset

Model	BERT		
Epochs	1	2	3
Precision	81.94%	82.50%	82.69%
Recall	80.91%	82.32%	82.14%
F1-Score	80.95%	82.35%	82.19%
Accuracy	80.91%	82.19%	82.24%
Training time	1312s	1301s	1288s

BERT uses a "masked language modeling" approach to learn the meaning of words by predicting them from their context. This makes it good at understanding the meaning of words in movie reviews. Based on the transformer model, BERT's architecture also contributes to its effectiveness. The transformer model implements self-attention mechanisms, allowing the model to attend to different words in a sentence while building representations. This attention mechanism enables BERT to capture dependencies within a text, facilitating a better understanding of the relationships between words. Table 8 shows the performance of LSTM on the RT dataset.

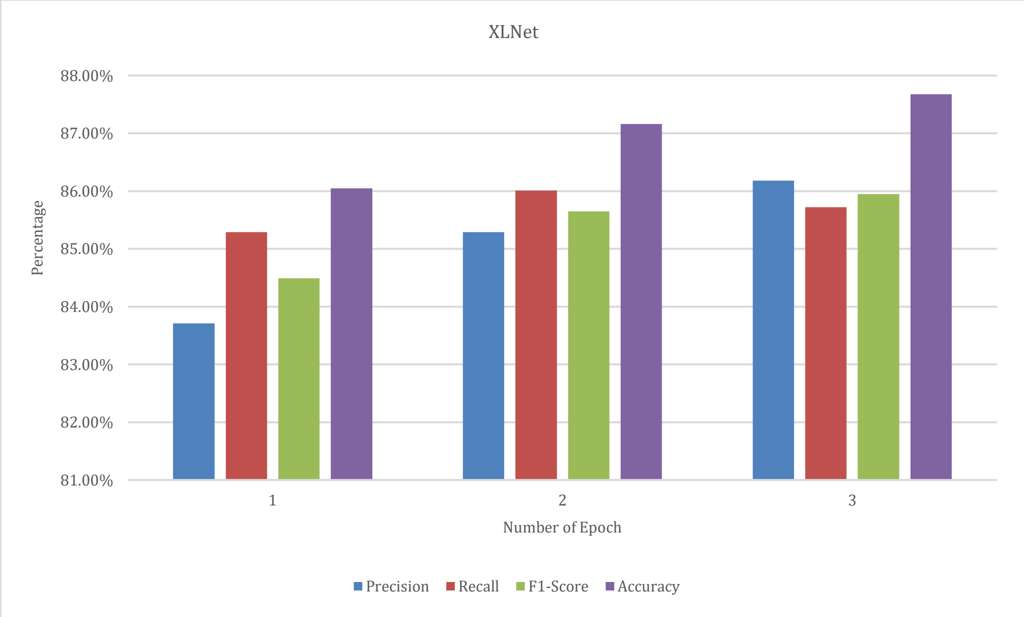

Fig. 9: Performance of XLNet on RT dataset

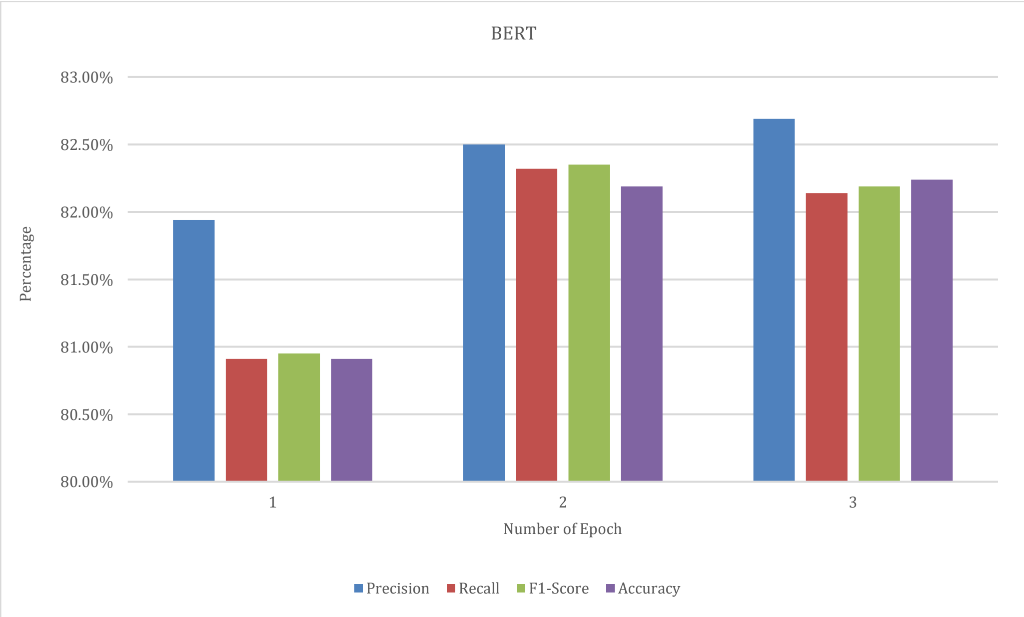

Fig. 10: Performance of BERT on RT dataset

Table 8: Performance of LSTM on RT dataset

Model	LSTM		
Epochs	1	2	3
Precision	67.41%	72.87%	70.88%
Recall	75.68%	66.70%	75.09%
F1-Score	71.31%	69.75%	72.92%
Accuracy	72.87%	74.10%	75.16%
Training time	313s	307s	264s

The performance of LSTM is visualized in Fig. 11 using bar chart. The relatively lower performance of the LSTM model compared to XLNet and BERT on the RT dataset can be attributed to factors such as the simpler architecture of LSTM compared to transformer-based models, limited ability to capture long-range dependencies in movie reviews, the dataset size possibly not providing enough data for effective training, and the absence of pre-training and transfer learning, which gives XLNet and BERT an advantage in understanding textual data and achieving higher performance in SA tasks. Table 9 illustrates the performance of the CNN-LSTM model on the RT Dataset.

Table 9: CNN-LSTM Performance on RT dataset

Model	CNN-LSTM		
Epochs	1	2	3
Precision	70.79%	71.74%	72.33%
Recall	81.89%	79.48%	75.04%
F1-Score	75.86%	75.39%	73.79%
Accuracy	76.19%	76.89%	76.98%
Training time	27s	24s	22s

CNN-LSTM performed better than LSTM because it effectively uses convolutional neural network (CNN) features to capture spatial information while complementing the temporal understanding capabilities of long short-term memory (LSTM) networks, resulting in improved overall performance. The performance of CNN-LSTM is visualized in Fig. 12 using bar chart.

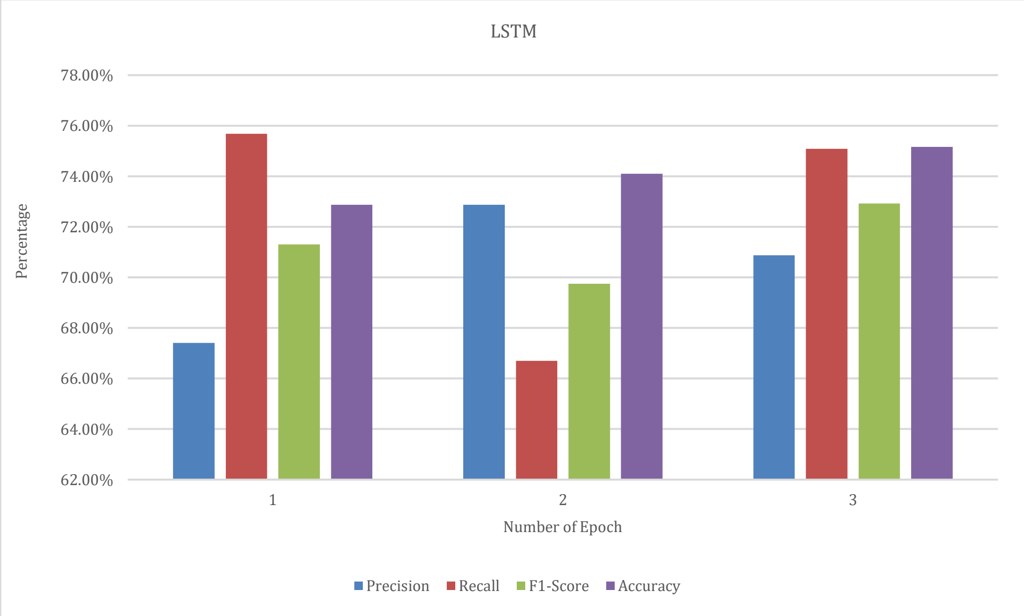

Fig. 11: Performance of LSTM on RT dataset

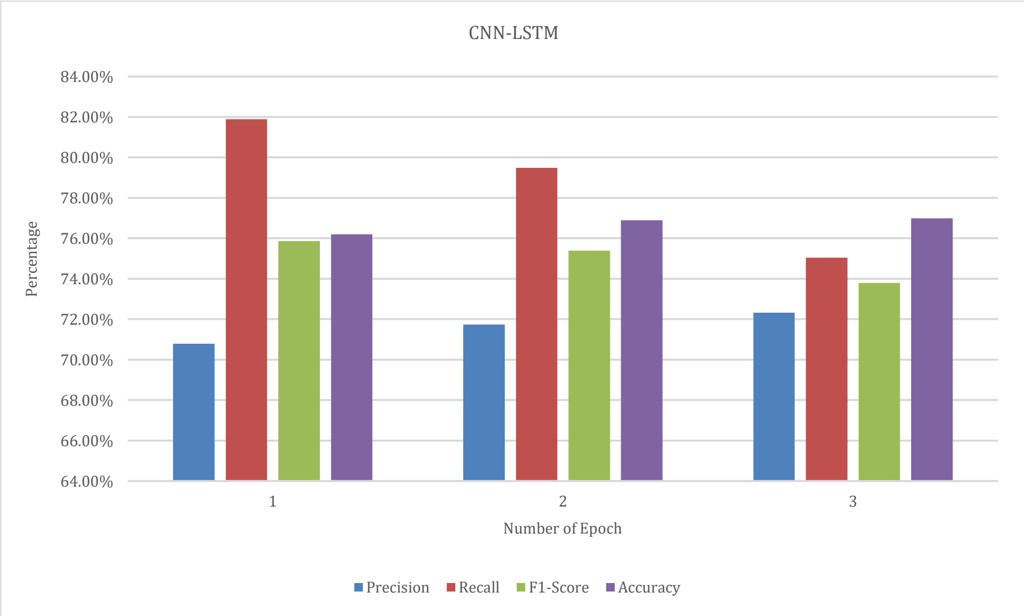

Fig. 12: Presenting the CNN-LSTM model's performance on the RT dataset

The evaluation time is measured in seconds (s) and varies for each model and the number of epochs. It provides insights into the computational efficiency of the models and can be useful in comparing their speed of inference. BERT takes a longer evaluation time due to its large model size, complex attention mechanisms, tokenization process, and resource-intensive computations. CNN-LSTM has less evaluation time due to its parallel processing capabilities, making it more efficient compared to traditional sequential models. The accuracy comparison of models is shown in Fig. 13.

After the third epoch, increasing the epochs did not significantly improve the model's performance, so we limited the epochs to three. The reason behind this phenomenon lies in the fact that the model had already grasped most of the pertinent patterns and insights from the training data during the initial three epochs. Further training beyond this point did not yield substantial advantages. Stopping training in such cases saves computational costs and time.

5.1. Discussion

This study proposes a model for SA of movie reviews using advanced models such as XLNet, BERT, LSTM, and CNN-LSTM. Before evaluation, the dataset is cleaned through several preprocessing steps, and the hyperparameters of all models are fine-tuned. Our experimental results show that XLNet outperforms the other models with an accuracy of 87.68%, which is 5% higher than BERT. BERT also performs better than LSTM and CNN-LSTM, achieving an accuracy of 82.24%, which is 5% higher than CNN-LSTM. XLNet's superior performance is due to its advanced permutation language modeling and extensive pre-training, which improve its ability to understand long-term dependencies and context. BERT also performs well because of its masked language modeling and self-attention mechanisms, which help it capture complex word relationships. CNN-LSTM benefits from combining CNN's ability to extract spatial features with LSTM's strength in learning temporal sequences. However, LSTM falls behind because of its simpler architecture and lack of pre-training, making it less effective for complex SA tasks. Table 10 compares the performance of our proposed method with other existing models.

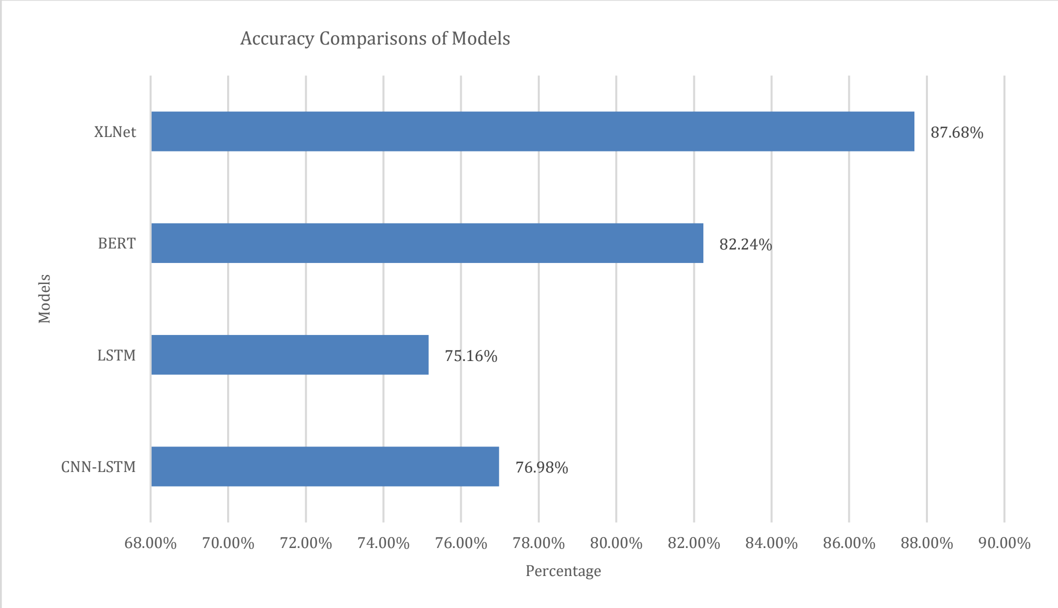

Fig. 13: Accuracy comparison of Models on the RT dataset

Table 10: Performance comparison of proposed model with existing techniques

No.	1	2	3	4	5	6
Models	Nath and Roy (2023)	Abimanyu et al. (2021)	Başarslan and Kayaalp (2023)	Liachoudis (2020)	Putrada et al. (2023)	Proposed
Techniques	CNN	Logistic Regression	Voting Model + TF-IDF	SVM + TF-IDF	BERT	XLNet
Accuracy	80%	76.88%	87.20%	78%	75.30%	87.68%

SA models can be biased by their training data, such as choosing specific populations or opinions, which leads to unfair and inaccurate results. For example, a model trained in Western movie reviews may not accurately reflect non-Western audiences' sentiments. Ethically, privacy concerns occur because personal opinions and reviews are analyzed, so data anonymity and consent are important. Furthermore, biased models can influence important decisions, such as hiring and lending, potentially leading to unfair treatment. As a result, transparency in model development and the implementation of fairness measures are critical to ensuring ethical and unbiased SA.

6. Limitations

The Rotten Tomatoes English movie reviews dataset produced promising results for the model; however, several challenges must be considered. The main limitation is that the dataset focuses only on English movie reviews, which raises concerns about language and cultural biases. This may affect the model's performance when applied to reviews in other languages or from different cultural backgrounds. Additionally, the focus on movie reviews may not translate well to SA in other contexts, such as social media posts or product reviews. Further research could explore domain-specific adjustments to improve the model's performance in different areas. Although we tuned the hyperparameters, not all configurations were explored due to computational constraints. While XLNet performed well in this study, its effectiveness may vary across other datasets. In conclusion, while our study provides useful insights, the limitations highlight the need for further research and improvements.

7. Conclusions and future work

Movie reviews are a form of textual data that provide valuable insights, opinions, and recommendations, helping people make informed decisions, discover new films, and appreciate the art of cinema. This paper proposed a SA model using advanced deep learning and transformer models, including CNN-LSTM, XLNet, LSTM, and BERT. XLNet achieved the highest accuracy on the Rotten Tomatoes dataset, demonstrating the ability of deep learning models to extract complex sentiments from textual opinions. This study highlights the importance of SA as a key tool for understanding emotions expressed in movie reviews. However, there are some limitations. The dataset used focuses only on English reviews, which may not be applicable to other languages or cultures. While the model works well with movie reviews, it may not perform as effectively on other types of text, such as news articles or social media posts. Due to time and computational limitations, we were unable to test all possible hyperparameter settings. Although XLNet performed well, its effectiveness may vary with different datasets, requiring further research. Future work will address these limitations by incorporating Aspect-Based SA and expanding the analysis to include multilingual datasets. These improvements are expected to enhance both research outcomes and the overall effectiveness of SA across diverse contexts of movie reviews.

List of symbols

AI	Artificial intelligence
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional neural networks
CV	Cross-validation
DL	Deep learning
IMDB	Internet movie database
KNN	K-nearest neighbors
LR	Logistic regression
LSTM	Long short-term memory
SVM	Support vector machines
ML	Machine learning
NB	Naïve Bayes
NLP	Natural language processing
NLTK	Natural language tool-kit
OP	Opinion mining
RT	Rotten tomatoes
SA	Sentiment analysis

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abimanyu A, Pranowo WS, Faizal I, Afandi NK, and Purba NP (2021). Reconstruction of oil spill trajectory in the Java Sea, Indonesia using SAR imagery. Geography, Environment, Sustainability, 14(1): 177-184. https://doi.org/10.24057/2071-9388-2020-21

Abimanyu AJ, Dwifebri M, and Astuti W (2023). Sentiment analysis on movie review from rotten tomatoes using logistic regression and information gain feature selection. Building of Informatics, Technology and Science, 5(1): 162-170. https://doi.org/10.47065/bits.v5i1.3595

Agrawal T (2021). Introduction to hyperparameters. In: Agrawal T (Ed.), Hyperparameter optimization in machine learning: Make your machine learning and deep learning models more efficient: 4-5. APRESS, New York, USA. https://doi.org/10.1007/978-1-4842-6579-6

Aziz MM, Purbalaksono MD, and Adiwijaya A (2023). Method comparison of Naïve Bayes, logistic regression, and SVM for analyzing movie reviews. Building of Informatics, Technology and Science, 4(4): 1714-1720. https://doi.org/10.47065/bits.v4i4.2644

Banik N and Rahman MHH (2018). Evaluation of Naïve Bayes and support vector machines on Bangla textual movie reviews. In the International Conference on Bangla Speech and Language Processing, IEEE, Sylhet, Bangladesh: 1-6. https://doi.org/10.1109/ICBSLP.2018.8554497

Başarslan MS and Kayaalp F (2023). Sentiment analysis with ensemble and machine learning methods in multi-domain datasets. Turkish Journal of Engineering, 7(2): 141-148. https://doi.org/10.31127/tuje.1079698

Chakraborty K, Bhattacharyya S, Bag R, and Hassanien AE (2018). Comparative sentiment analysis on a set of movie reviews using deep learning approach. In: Hassanien A, Tolba M, Elhoseny M, and Mostafa M (Eds.), The international conference on advanced machine learning technologies and applications: Advances in intelligent systems and computing: 311-318, Volume 723. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-74690-6_31

Dang NC, Moreno-García MN, and De la Prieta F (2020). Sentiment analysis based on deep learning: A comparative study. Electronics, 9(3): 483. https://doi.org/10.3390/electronics9030483

Danyal MM, Haseeb M, Khan SS, Khan B, and Ullah S (2024a). Opinion mining on movie reviews based on deep learning models. Journal of Artificial Intelligence, 6: 23-42. https://doi.org/10.32604/jai.2023.045617

Danyal MM, Khan SS, Khan M, Ghaffar MB, Khan B, and Arshad M (2023). Sentiment analysis based on performance of linear support vector machine and multinomial Naïve Bayes using movie reviews with baseline techniques. Journal on Big Data, 5: 1-18. https://doi.org/10.32604/jbd.2023.041319

Danyal MM, Khan SS, Khan M, Ullah S, Ghaffar MB, and Khan W (2024b). Sentiment analysis of movie reviews based on NB approaches using TF–IDF and count vectorizer. Social Network Analysis and Mining, 14: 87. https://doi.org/10.1007/s13278-024-01250-9

Danyal MM, Khan SS, Khan M, Ullah S, Mehmood F, and Ali I (2024c). Proposing sentiment analysis model based on BERT and XLNet for movie reviews. Multimedia Tools and Applications, 83: 64315–64339. https://doi.org/10.1007/s11042-024-18156-5

Dashtipour K, Gogate M, Adeel A, Larijani H, and Hussain A (2021). Sentiment analysis of Persian movie reviews using deep learning. Entropy, 23(5): 596. https://doi.org/10.3390/e23050596
PMid:34066133 PMCid:PMC8151596

Deepa D, Nafais AS, Kumar BM, Prasath JR, Suba T, and Jenopaul P (2021). Analyzing the performance of bidirectional transformer and generalized autoregressive permutation pre-trained language models for sentiment classification task. Annals of the Romanian Society for Cell Biology, 25(6): 7598-7604.

Devlin J, Chang MW, Lee K, and Toutanova K (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Arxiv Preprint Arxiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Dhivyaa CR, Nithya K, Sendooran G, Sudhakar R, Kumar KS, and Kumar SS (2023). XLNet transfer learning model for sentimental analysis. In the International Conference on Sustainable Computing and Smart Systems, IEEE, Coimbatore, India: 76-84. https://doi.org/10.1109/ICSCSS57650.2023.10169445

Dholpuria T, Rana YK, and Agrawal C (2018). A sentiment analysis approach through deep learning for a movie review. In the 8th International Conference on Communication Systems and Network Technologies, IEEE, Bhopal, India: 173-181. https://doi.org/10.1109/CSNT.2018.8820260

Khan B, Arshad M, and Khan SS (2023). Comparative analysis of machine learning models for PDF malware detection: Evaluating different training and testing criteria. Journal of Cybersecurity, 5: 1-11. https://doi.org/10.32604/jcs.2023.042501

Khan M, Khan MS, and Alharbi Y (2020). Text mining challenges and applications: A comprehensive review. International Journal of Computer Science and Network Security, 20(12): 138-148.

Khan SS, Khan M, Ran Q, and Naseem R (2018). Challenges in opinion mining, comprehensive review. A Science and Technology Journal, 33(11): 123-135.

Leone S (2020). Rotten Tomatoes movies and critic reviews dataset. Kaggle, San Francisco, USA.

Li H, Zhang X, Liu Y, Zhang Y, Wang Q, Zhou X, Liu J, Wu H, and Wang H (2019). D-NET: A pre-training and fine-tuning framework for improving the generalization of machine reading comprehension. In the Proceedings of the 2nd Workshop on Machine Reading for Question Answering: 212–219, Hong Kong, China. https://doi.org/10.18653/v1/D19-5828

Liachoudis G (2020). Sentiment analysis of movie reviews by merging comments from two well-known platforms. Ph.D. Dissertation, Tilburg University, Tilburg, Netherlands.

Lou Y (2023). Deep learning-based sentiment analysis of movie reviews. In the 3rd International Conference on Machine Learning and Computer Application, SPIE, Shenyang, China: 12636: 177-184.

Mutegeki R and Han DS (2020). A CNN-LSTM approach to human activity recognition. In the International Conference on Artificial Intelligence in Information and Communication, IEEE, Fukuoka, Japan: 362-366. https://doi.org/10.1109/ICAIIC48513.2020.9065078

Nath D and Roy J (2023). Forecast of movie sentiment based on multi label text classification on rotten tomatoes using multiple machine and deep learning technique. In: Mercier-Laurent E, Fernando X, and Chandrabose A (Eds.), Computer, communication, and signal processing: AI, knowledge engineering and IoT for smart systems: 128-142. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-031-39811-7_11

Palomo BA, Velarde FH, Cantu-Ortiz FJ, and Ceballos Cancino HG (2024). Sentiment analysis of IMDB movie reviews using deep learning techniques. In: Yang XS, Sherratt RS, Dey N, and Joshi A (Eds.), Proceedings of eighth international congress on information and communication technology. ICICT 2023. Lecture Notes in Networks and Systems, Volume 696. Springer, Singapore, Singapore. https://doi.org/10.1007/978-981-99-3236-8_33

Putrada AG, Alamsyah N, and Fauzan MN (2023). BERT for sentiment analysis on rotten tomatoes reviews. In the International Conference on Data Science and Its Applications (ICoDSA), IEEE, Bandung, Indonesia: 111-116. https://doi.org/10.1109/ICoDSA58501.2023.10276800

Rahman A and Hossen MS (2019). Sentiment analysis on movie review data using machine learning approach. In the International Conference on Bangla Speech and Language Processing, IEEE, Sylhet, Bangladesh: 1-4. https://doi.org/10.1109/ICBSLP47725.2019.201470

Sherstinsky A (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404: 132306. https://doi.org/10.1016/j.physd.2019.132306

Tripathy A, Anand A, and Kadyan V (2023). Sentiment classification of movie reviews using GA and NeuroGA. Multimedia Tools and Applications, 82(6): 7991-8011. https://doi.org/10.1007/s11042-022-13047-z

Ullah K, Rashad A, Khan M, Ghadi Y, Aljuaid H, and Nawaz Z (2022). A deep neural network-based approach for sentiment analysis of movie reviews. Complexity, 2022: 5217491. https://doi.org/10.1155/2022/5217491

Van Houdt G, Mosquera C, and Nápoles G (2020). A review on the long short-term memory model. Artificial Intelligence Review, 53(8): 5929-5955. https://doi.org/10.1007/s10462-020-09838-1

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, and Le QV (2019). XLNET: Generalized autoregressive pretraining for language understanding. In the 33rd Conference on Neural Information Processing Systems, Vancouver, Canada: 1-18.

Yasen M and Tedmori S (2019). Movies reviews sentiment analysis and classification. In the IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology, IEEE, Amman, Jordan: 860-865. https://doi.org/10.1109/JEEIT.2019.8717422

Zhang L, Wang S, and Liu B (2018). Deep learning for sentiment analysis: A survey. WIREs: Data Mining and Knowledge Discovery, 8(4): e1253. https://doi.org/10.1002/widm.1253