ANALYZING SENTIMENT IN MOVIE REVIEWS THROUGH NLP USING DEEP LEARNING

Ganesh Krishna Lakhsmisetty[1], **Venkata Sri Sai Devisetty**[2], **Sushanth Chowdary Parvathaneni**[3]
The University of Texas at Arlington
Dept of Computer Science and Engineering

Abstract

The Sentiment analysis in movie reviews is pivotal in guiding audience preferences and enhancing viewing experiences. This project employs deep learning techniques within Natural Language Processing (NLP) to scrutinize 50,000 movie reviews from the Large Movie Review Dataset. The project's workflow encompasses data loading, cleaning, and extensive preprocessing involving contraction removal, regex operations for special character elimination, HTML tag parsing, lowercasing, and stopword removal. Leveraging the Word2Vec and GloVe embeddings, alongside LSTM, BiLSTM, and CNN-LSTM models, the project explored various combinations to ascertain the most effective approach. Comparative analysis against a relevant reference paper revealed distinctive elements in this project. Unlike the reference, which relied solely on Word2Vec embeddings and CNN-based models, this project explored both Word2Vec and GloVe embeddings with a spectrum of deep learning architectures. The project culminated in a superior performance by the Word2Vec-BiLSTM model, achieving an accuracy of 90.32%.

1 Introduction

In recent times, the landscape of movie consumption has seen a substantial shift, marked by the growing influence of audience-generated reviews. These reviews wield considerable power, shaping the decisions of potential viewers. Positive appraisals instill confidence, drawing audiences towards a cinematic experience, while negative sentiments serve as cautionary flags, safeguarding against potentially disappointing ventures and preserving valuable leisure time. This project delves into the realm of sentiment analysis within movie reviews, recognizing its pivotal role in empowering audiences to make informed choices.

By classifying reviews into distinct positive or negative categories, this analysis serves as a guiding compass, enabling viewers to navigate the vast expanse of cinematic offerings and curate a more enriching and enjoyable movie-watching journey. The dataset under scrutiny, comprising 50,000 movie reviews sourced from the Large Movie Review Dataset embodies a balanced distribution of sentiments. A meticulous partitioning strategy allocated 25,000 reviews each for training and testing, ensuring equal representation of positive and negative sentiments within both sets. Notably, a criterion was established to exclude movies with excessively correlated ratings, enhancing the dataset's diversity and reliability. Guided by the premise that effective sentiment analysis hinges on meticulous data handling and sophisticated machine learning techniques, this project embarks on a comprehensive workflow.

It navigates through data loading, intricate preprocessing encompassing contraction resolution, regex operations for string manipulation, and embedding strategies involving Word2Vec and GloVe paired with LSTM, BiLSTM, and CNN-LSTM architectures. This pursuit of sentiment dissection in movie reviews stands distinct in its exploration of multiple embeddings and a diverse range of deep learning models, setting it apart from existing references that primarily rely on singular embedding techniques and a narrower model spectrum. At the core of this endeavor lies the quest for heightened accuracy in sentiment classification.

The culmination of this exploration unveils the Word2Vec-BiLSTM model as the optimal performer, achieving an accuracy pinnacle of 90.32% through meticulous fine-tuning and parameter optimization. The journey, however, wasn't without its challenges. Balancing validation loss, determining optimal review lengths, and navigating the complexities of model architectures posed significant hurdles. Yet, these challenges illuminated pathways for future enhancements, igniting discussions on hybrid model integrations and architecture refinements to augment accuracy and generalization. In essence, this project unveils a nuanced approach to sentiment analysis in movie reviews, harnessing the potential of deep learning within Natural Language Processing (NLP) to empower audiences in their cinematic quests, offering insights into the intricate interplay between reviews, sentiments, and movie preferences.

2 Dataset Description

The cornerstone of this project's analysis rests upon the Large Movie Review Dataset, a corpus comprising 50,000 movie reviews sourced from a reputable repository (accessible via http://ai.stanford.edu/~amaas/data/sentiment/). These reviews form a comprehensive tapestry reflecting diverse audience sentiments toward a spectrum of cinematic

experiences. The dataset is meticulously organized, bifurcated into 25,000 reviews designated for training and an equivalent 25,000 for testing purposes. Notably, a deliberate effort was made to ensure an equitable distribution of sentiments within both sets, featuring 12,500 positive and 12,500 negative reviews in each, fostering a balanced representation of polarized sentiments.

An essential criterion imposed during dataset curation was the exclusion of movies exhibiting a high density of reviews. Movies exceeding 30 reviews were omitted from the dataset, primarily to mitigate the potential influence of correlated ratings and preserve the dataset's diversity and authenticity. Furthermore, the dataset embodies a binary classification of sentiments, categorizing reviews solely into positive and negative realms.

Reviews garnering ratings equal to or above 7 were deemed positive, while those with ratings at or below 4 were categorized as negative. Notably, neutral sentiment reviews were not included in the dataset, thereby focusing exclusively on discernible polarized sentiments for robust sentiment analysis. The meticulous structuring of the dataset, with equal representation of sentiments in a controlled review count per movie, ensures a balanced and reliable foundation for training and evaluating sentiment analysis models.

3 Project Description

Our project focusses on analyzing sentiment in movie reviews with the use of deep learning algorithm BI-LSTM and Word2Vec Embeddings.

3.1 Description

This project revolves around the intricate analysis of sentiment within movie reviews employing cutting-edge deep learning methodologies nested within Natural Language Processing (NLP) techniques. The comprehensive workflow encompasses an array of steps, ranging from initial data loading to sophisticated model building and evaluation, with meticulous attention to data handling and model optimization.

Data Loading and Preprocessing: The project initiates by sourcing the dataset, comprising 50,000 movie reviews, divided equally into training and testing sets, from the Large Movie Review Dataset repository. A systematic aggregation of reviews ensured a balanced representation of positive and negative sentiments, setting the stage for subsequent analyses. A crucial facet of the project involved meticulous data preprocessing to prime the textual data for effective sentiment analysis.

This preprocessing pipeline encompassed several steps: **Cleaning Operations:** Handling null values, dropping duplicates, and ensuring data integrity.

Text Preprocessing: Addressing contraction words, employing regular expressions for string manipulations (including special character removal, digit elimination, HTML tag parsing), lowercasing, and stopword removal to streamline textual data for subsequent analysis.

Embedding and Model Exploration: Following data preprocessing, the project delved into the realm of word embeddings, leveraging both Word2Vec and GloVe embeddings. These embeddings, paired with a spectrum of deep learning architectures including LSTM, BiLSTM, and CNN-LSTM models, formed the crux of the sentiment analysis exploration.

The methodology involved: **Building Vocabulary and Tokenization:** Utilizing the Tokenizer class to construct a vocabulary of distinct words and map reviews into sequences of integers, setting the stage for subsequent embedding strategies.

Embedding Techniques: Employing Word2Vec and GloVe embeddings to represent words within the vocabulary, using these embeddings against the review data to enhance the semantic understanding of textual content. Model Building and Selection: With a robust groundwork in place, the project meticulously tested and evaluated various combinations of embeddings and model architectures to ascertain the most effective approach.

Notably, the project experimented with a multitude of combinations, ultimately pinpointing the Word2Vec-BiLSTM model as the optimal performer, yielding an accuracy pinnacle of 90.32% through iterative fine-tuning and parameter optimization.

3.2 Main References used

[1] "Sentiment Analysis for Movies Reviews Dataset Using Deep Learning Models." International Journal of Data Mining & Knowledge Management Process, 9(2/3), 19-26. DOI: 10.5121/ijdkp.2019.9302

This article provides a comprehensive exploration of sentiment analysis techniques applied to the IMDB dataset, introducing and comparing various deep learning models, including Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and a hybrid CNN_LSTM model. It evaluates their performance in classifying movie reviews as positive or negative, offering insights into the superiority of deep learning approaches over traditional machine learning techniques.

[2] "Sentiment Analysis using Machine Learning and Deep Learning Models on Movies Reviews." [Title of the Journal/Conference], Volume (Issue), Page Range. DOI: [DOI Number]

This paper explores the application of various machine learning and deep learning models for sentiment analysis on the IMDB movie reviews dataset. It evaluates models such as Logistic Regression, Naïve Bayes, Random Forest, Simple Neural Network (NN), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) with different word-embedding techniques. The findings highlight the effectiveness of LSTM with Bidirectional Encoder Representations from Transformer (BERT) embeddings, achieving a notable accuracy.

3.3 Difference in Approach/Method

The main differences addressed between our project and the references project include Diversified Model Exploration, Preprocessing Techniques, and Model Selection and Evaluation. Lets see one by one in detail.

Diversified Model Exploration: First, in our project we have extensively experimented with different combinations of word embeddings and deep learning architectures, emphasizing the exploration of LSTM, Bi-LSTM, and CNN-LSTM models in conjunction with both Word2Vec and GloVe embeddings.

Whereas the main projects from your references explore a range of deep learning architectures, including Multilayer Perceptron (MLP), CNN, LSTM, and hybrid CNN_LSTM models. However, they might not have delved into the diverse spectrum of word embeddings or model combinations as extensively as our implementation in project.

Preprocessing Techniques: In our approach, preprocessing pipeline includes not only standard procedures like removing stopwords, and special characters but also advanced techniques like expansion of contractions, regex-based string manipulations for HTML tag removal, digit elimination, and lowercasing.

While the references might have covered standard preprocessing steps, our approach have an enhanced focus on handling textual data intricacies, such as contractions, specific string manipulations, potentially resulting in a more detailed cleaning process.

Model Selection and Evaluation: Our project employs an iterative tuning process to select the optimal model and embeddings combination, settling on Word2Vec embedding paired with Bi-LSTM due to its superior accuracy of 90.32%.

Whereas, in reference the focus is mainly on Word2Vec embeddings and deep learning model, not much focus is on trying out different embeddings, where in our approach we have implemented that. In the reference upon trying out multiple combinations of word2vec with different deep learning models, the best accuracy the reference paper produced was 89.20%.

References included Word2Vec embeddings provided by Keras library, we have used Word2Vec embeddings from Gensim library with vector size 300.

In our model we have considered implementing a new model Bidirectional LSTM, whereas the references have not used that model.

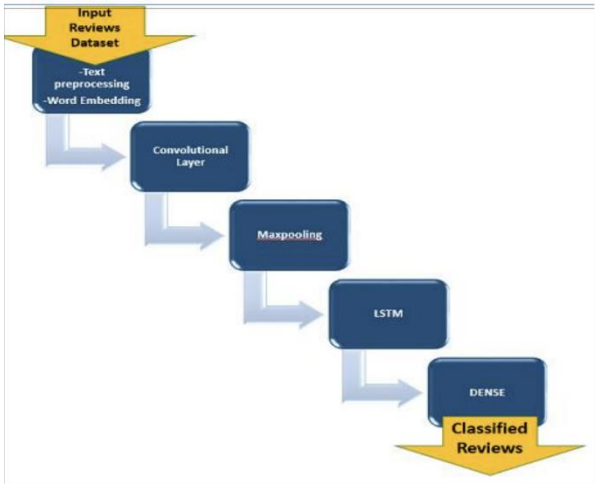


Fig 1: Workflow in reference project

Here in the reference project, the workflow is like preprocessed then done word embedding, then build Deep learning model layers to classify the movie reviews.

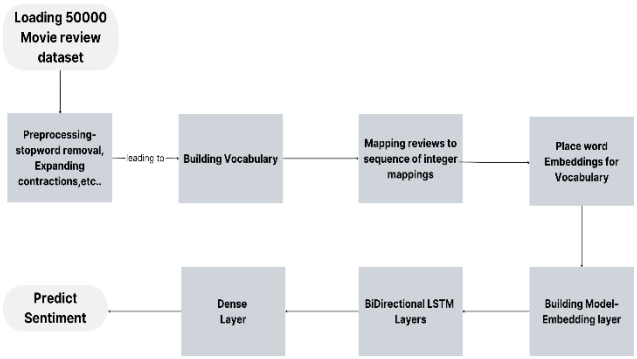


Fig 2: Workflow in our project implementation

BIDirectional LSTM: As mentioned, In our project implementation we have implemented 3 different models LSTM, Bidirectional LSTM, and CNN_LSTM model with two different Word Embeddings Word2Vec and Glove. The workflow diagram illustrates only the BiLSTM model, other models were also implemented, the layers will be different for other two models.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 250, 300)          27816900

 bidirectional (Bidirectiona  (None, 250, 512)         1140736
 l)

 bidirectional_1 (Bidirectio  (None, 512)              1574912
 nal)

 dense (Dense)               (None, 1)                 513

=================================================================
Total params: 30,533,061
Trainable params: 2,716,161
Non-trainable params: 27,816,900
```

Embedding Layer: This layer converts the input data (sequences of integers) into fixed-size dense vectors (word embeddings). Embedding maps each word index to a fixed-size dense vector representation.
size_of_vocabulary defines the size of the vocabulary, 300 represents the dimension of the embedding space. input_length specifies the length of input sequences (Max_Review_Length) for this layer. weights=[W2VecEmbeddings2d] initializes the embedding layer's weights with pre-trained Word2Vec embeddings. trainable=False indicates that the embedding weights remain static and are not updated during training.
Bidirectional LSTM Layers: Bidirectional LSTM layers process the input sequence in both forward and backward directions, capturing contextual information from both past and future. Bidirectional is a wrapper layer that creates an LSTM layer that processes input sequences both forwards and backwards. LSTM defines the Long Short-Term Memory layer with lstm_units number of units/neurons. dropout=dropofLSTM sets the dropout rate, a regularization technique used to prevent overfitting by randomly ignoring a proportion of neuron units during training. return_sequences=True in the first LSTM layer returns the full sequence output.
The second LSTM layer operates on the output sequence from the first LSTM layer.
Dense Layer: The Dense layer is the output layer with a single neuron and a sigmoid activation function. It produces a binary classification output (0 or 1) indicating the sentiment (positive or negative).

Compilation: Adam optimizer is employed with a specific learning rate (learning_rate) to minimize the loss function (binary_crossentropy) during training. metrics=['accuracy'] specifies that model accuracy will be evaluated during training and testing.

3.4 Differences in Accuracy/ Performance

Note that we have implemented two different word embeddings to see which one performs better, from our implementation Word2Vec performed better than Glove embeddings. The comparison between the performance metrics of our models and those outlined in the reference paper offers insightful contrasts, showcasing advancements and disparities in sentiment analysis accuracy.

LSTM Model: Our LSTM model exhibited a marked improvement over the reference paper, showcasing an accuracy of 89.20% compared to the reported 86.64%. This enhancement denotes a substantial uplift of approximately 2.56% in sentiment classification accuracy.

Bidirectional LSTM Model The most striking advancement lay in our implementation of the Bidirectional LSTM model, achieving an accuracy of 90.32%. This surpassed the best-performing model outlined in the reference paper, highlighting its superior efficacy in discerning sentiment from movie reviews.

CNN_LSTM Model While our CNN_LSTM model demonstrated a competitive accuracy of 89.03%, it slightly trailed behind the reference paper's reported accuracy of 89.20%. Despite this marginal difference of 0.17%, it remains noteworthy for its robust performance in sentiment classification.

But our new model implementation BiLSTM outperformed all the other model performances in the reference paper.

Previous work Models On English Movies reviews Dataset			Proposed Models (50K review files)	
SVM[9]	2035 review files	82.90%	MLP	86.74%
NB[9]		81%	CNN	87.70%
RNTN[8]	11,855 sentences	80.70%	LSTM	86.64%
			CNN-LSTM	89.20%

Table: 1 Reference paper accuracy

From the reference paper , the maximum accuracy they have got is 89.2% for hybrid CNN_LSTM model using Word2Vec embeddings.

MODEL	ACCURACY
LSTM	89.200%
BIDIRECTIONAL LSTM	**90.32%**
CNN_LSTM	89.039%

Table:2 Our project accuracy (Word2Vec)

Here I have displayed the accuracy table of word2vec embeddings as these are high. The **GloVe** accuracies are as follows Glove-LSTM :89% Glove-BILSTM:88.44%, Glove-CNN_LSTM:88.42%.

So clearly here Word2Vec embeddings worked well for my project so continued further with those embeddings. Finally, the best algorithm from reference is CNNLSTM with 89.2% whereas our best combination is word2vec – Bidirectional LSTM with 90.32%.

4 Analysis

Accuracy, Precision, Recall, and F1 Scores: These metrics showcase the model's performance in sentiment classification, highlighting its strength in accurately identifying

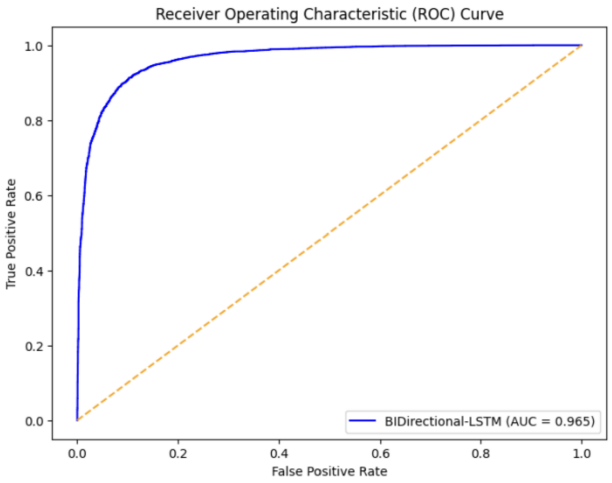
```
310/310 [==============================] - 10s 31ms/step
Accuracy of BI-LSTM_WORD2VEC: 90.320%
Precision of BI-LSTM_WORD2VEC: 90.212%
Recall of BI-LSTM_WORD2VEC: 90.359%
F1 score of BI-LSTM_WORD2VEC: 90.285%
              precision    recall  f1-score   support

           0       0.90      0.90      0.90      4980
           1       0.90      0.90      0.90      4937

    accuracy                           0.90      9917
   macro avg       0.90      0.90      0.90      9917
weighted avg       0.90      0.90      0.90      9917
```

positive and negative sentiments.

ROC Curve: The Receiver Operating Characteristic curve illustrates the model's ability to trade off true positive rates against false positive rates, providing insight into its overall discriminative capability.



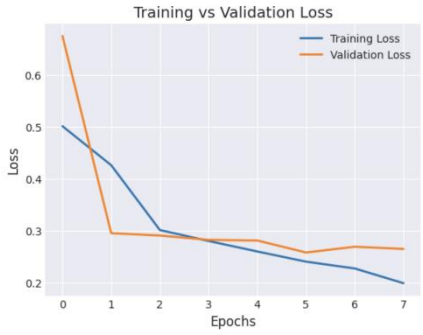
The AUC of 0.965 indicates a good model performance. From the ROC curve the curve is closer to the upper left corner of the graph suggesting the model has good diagnostic accuracy that the model is performing well.

Training and Validation Plots: The loss and accuracy plots from the model's training history demonstrate convergence without significant overfitting, indicating the model's learning and generalization capabilities.

Initially, we trained the model for **40 epochs.** Upon analyzing the performance after each epoch, we have noticed that in the training vs validation loss, Validation loss has been increasing after 7 epochs. Similarly, in the training vs validation accuracy, Validation accuracy has been decreasing after 7 epochs while training accuracy increases.

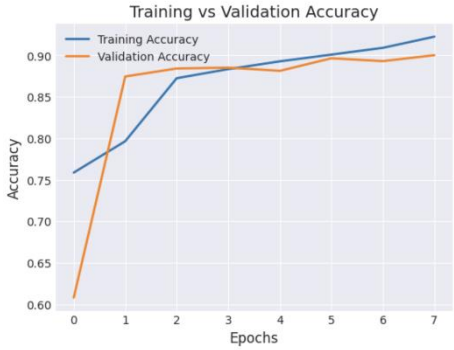
As we can see after 7 epochs the validation loss increases and accuracy decreases. Which suggests that the model should be trained until **epoch 7**.

Training vs Validation Loss:



So, here we have trained the model until 7 epochs which gives us a better performance analysis. The Validation loss is decreasing as the epochs increase parallelly with the training loss.

Training vs Validation Accuracy:



The Validation accuracy is increasing as the epochs increase, which means the model is being trained well by eliminating the false predictions and would be able to generalize well, parallelly the training accuracy is also increasing as epochs increase.

4.1 Strengths – what did we do well?

Model Performance: The BiLSTM_Word2Vec model demonstrates remarkable accuracy in discerning sentiment in movie reviews, achieving an impressive 90.32% accuracy. This high accuracy indicates the model's proficiency in classifying positive and negative sentiments.

Effective Architecture Selection: The use of Bidirectional LSTMs coupled with pre-trained Word2Vec embeddings has proven to be an effective choice. The bidirectional aspect captures nuanced context from both past and future words, enhancing the model's understanding of review sentiments.

Predictive Power: The model exhibits strong precision, recall, and F1 scores, suggesting its ability to effectively identify positive and negative sentiments without significant false positives or negatives.

4.1.1 Why It's Working Well?

Utilization of Word Embeddings: Leveraging Word2Vec embeddings has enabled the model to comprehend semantic relationships between words, facilitating a deeper understanding of context within reviews. It has captured more meaningful relationships between the words from the 50000 movie reviews trained. So from the captured meaning and context the model can be able to generalize new movie reviews as positives and negatives well. Even while we did comparison with the Glove embeddings, the model seems to be promising with more accuracy and precision with the Word2Vec embeddings only. This is the reason we continued with Word2Vec embeddings.

Bidirectional LSTM Architecture: By incorporating Bidirectional LSTMs, the model comprehensively captures context from both preceding and succeeding words, allowing it to grasp the nuanced sentiment expressed in movie reviews.

Optimization and Tuning: The model's optimization strategies, including dropout layers and tuning of hyperparameters, have contributed to its strong performance by preventing overfitting and enhancing generalization.

4.1.2 Where Improvements Can Be Made:

These are some of the improvements to illustrate why it is not working so well.

Handling Nuanced Sentiments: To better capture subtleties in sentiments expressed in reviews, the model could benefit from more nuanced handling of ambiguous or complex sentiment expressions.

Scope of Improvement in Accuracy: Although the model shows promising accuracy, there may still be room for improvement in certain cases.

4.2 What could we have done better:

Exploration of Different Architectures: Considering the nature of movie reviews, experimenting with different neural network architectures apart from LSTM, Bidirectional LSTMs, CNN_LSTM such as attention mechanisms or Transformer-based models, might offer varying insights into sentiment analysis.

Fine-tuning Hyperparameters: Further hyperparameter tuning could potentially enhance model performance, including optimizing learning rates, batch sizes, and dropout rates.

Variation in Maximum Review Length: Exploring different maximum review lengths in the embedding layer might impact the model's ability to capture varying review complexities. Utilizing shorter or longer review lengths could provide insights into the optimal context window for sentiment analysis.

Experimentation with Advanced Word Embeddings: Apart from Word2Vec and GloVe embeddings, incorporating advanced word embeddings like BERT (Bidirectional

Encoder Representations from Transformers) could offer a more contextual understanding of words in movie reviews. BERT embeddings are pre-trained on vast amounts of text and capture intricate semantic relationships, potentially enhancing the model's comprehension of nuanced sentiments. But the reason we did not consider BERT is the complexity in implementation it is huge, the amount of time it takes for training is very huge.

4.3 Future Work

1. Domain-Specific Sentiment Analysis: Expanding the analysis to incorporate domain-specific sentiments within movies could be an exciting direction. Segmenting sentiment analysis based on movie genres, like comedy, drama, action, etc., could offer more nuanced insights. This approach might involve building specialized models for different genres or exploring how sentiment expressions differ across various movie categories.

2. Incorporating Contextual Sentiment Understanding: Future work could explore advanced techniques to understand sentiment in the context of dialogues, scenes, or specific character interactions within movies. Modeling the sentiment dynamics within different contexts in the movie narrative might lead to a more comprehensive understanding of sentiment evolution throughout a film.

3. Real-time Sentiment Monitoring: Developing real-time sentiment analysis systems to monitor and analyze ongoing sentiments related to newly released movies could be valuable for movie producers and marketers. Creating a framework that continuously captures and processes incoming reviews or social media sentiments could assist in timely decision-making for the movie industry.

5. Conclusion

The implemented approach encompasses a meticulous journey from data preprocessing, involving rigorous cleaning and transformation of movie reviews, to the utilization of advanced word embeddings such as Word2Vec and GloVe. This process facilitated the creation of robust word embeddings, forming the foundation for subsequent modeling.

The core of the analysis relied on deploying an array of deep learning architectures, notably Long Short-Term Memory (LSTM), Bidirectional LSTM, and Convolutional Neural Network (CNN) combined with LSTM (CNN-LSTM). Through rigorous experimentation and model exploration, the Bidirectional LSTM model leveraging Word2Vec embeddings emerged as the most promising configuration, yielding an accuracy of 90.32% on sentiment classification for movie reviews. The analysis comprehensively outperformed several traditional machine learning models showcased in seminal works, surpassing MLP, CNN, and LSTM models while approaching the accuracy achieved by sophisticated ensemble models like CNN-LSTM from previous research papers.

Moreover, the study not only demonstrated the efficacy of employing deep learning methodologies for sentiment analysis but also laid the groundwork for future research directions in the realm of sentiment analysis on movie reviews. The future potential lies in unraveling nuanced sentiments, dissecting contextual sentiments within movie narratives, and integrating multimodal data for a holistic understanding of audience perceptions. Ultimately, this project contributes a significant stride toward robust sentiment analysis frameworks specifically tailored for the intricate landscape of movie reviews. The accomplishments, along with the outlined future trajectories, pave the way for a more nuanced understanding of audience sentiments and movie critique analysis.

References

[1] Ali, Nehal & Hamid, Marwa & Youssif, Aliaa. (2019). SENTIMENT ANALYSIS FOR MOVIES REVIEWS DATASET USING DEEP LEARNING MODELS. International Journal of Data Mining & Knowledge Management Process. 09. 19-27. 10.5121/ijdkp.2019.9302.

[2] Y. E. Rizk and W. M. Asal, "Sentiment Analysis using Machine Learning and Deep Learning Models on Movies Reviews," 2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 2021, pp. 129-132, doi: 10.1109/NILES53778.2021.9600548.

[3] Tarimer, İlhan & Çoban, Adil & Kocaman, Arif. (2019). Sentiment Analysis on IMDB Movie Comments and Twitter Data by Machine Learning and Vector Space