Sentiment Movie Review Analysis

Aida Hallaci
220042670
Data Science (Msc)
Aida.Hallaci@city.ac.uk

1 Problem Statement and Motivation

This project uses deep-learning techniques to conduct sentiment analysis on IMDB movie reviews. Sentiment analysis allows us to automatically classify the sentiment of our dataset, such as reviews, into positive or negative. The motivation for this paper arises from the crucial role that the opinions of movie audiences play in the movie industry. Understanding the views expressed allows the film industry to make informed decisions to gauge a movie's popularity and potential success. We aim to analyze movie reviews accurately, so we propose improved systems, providing users with personalized suggestions based on their preferences. Also, sentiment analysis can be a powerful tool for marker research in the movie industry, allowing tracking the latest trends and audience preferences.

2 Research hypothesis

Based on our IMDb movie reviews, we aim to predict the sentiment of the reviews as positive or negative. We will implement and compare three deep learning methods LSTM,RNN, and pre-trained BERT. We hypothesize that the pre-trained model BERT will outperform other models such as LSTM, Bi-LSTM and RNN.This is also supported by previous research showing BERT to be highly effective in polarity classification tasks [13].

3 Related work and background

In [1], Ibrahim et al., 2019. introduce a novel approach to improving movie recommendations using neural network techniques. The authors address the importance of cleaning and normalizing text, incorporating stemming and tokenization to reduce sparsity and shrink the feature space. They emphasise the significance of tokenization, stemming, POS-tag generation and word sense disambiguation (WSD) for practical text analysis. Additionally, the authors explore the use of Recurrent Neural Networks (RNN) and Long-Short-Term Memory (LSTM); incorporating these methods in their study demonstrates the capacity to improve the accuracy and precision, recall and F1-score and reliability of movie recommendation systems.

The second paper [2]. by Nehal Mohamed Ali et al. has contributed to the sentiment analysis field for a dataset of 50K movies reviews from IBMD. This study explores three deep learning network-MLP, CNN and LSTM-along with a hybrid network, CNN_LSTM.To enrich the analysis, they utilized the Word2Vector technique for word embedding. Results showed that the CNN_LSTM network outperformed other models, achieving an accuracy of 89.20%.

Another paper by Biswarup Ray et al.[3],proposes a system for a hotel recommendation that classifies the sentiments of hotel reviews and groups them into different categories. This paper implemented a BERT model and obtained an accuracy of 92.36% while also providing detailed evaluation metrics and visualizations for the model performance.

The fourth paper studied using LSTM classifiers for sentiment analysis on IMDb movie reviews, combined with Adam optimiser [4]. The research findings reveal that the highest accuracy achieved is 89.9%

H. M. Keerthi Kumar et al. developed [5] a paper comprising four stages: preprocessing, feature extraction, feature selection and classification. The best accuracy (74.66%) was obtained with the Maximum Entropy classifier using RLPI (Relevance Likelihood Probability Index) as the future selection method.

In 2021, Ramadhan, N. G et al. conducted a study to analyse the sentiment of opinions on the IMDb website, where the focus was the star rating aspect [6]. The users struggled to analyze the movie

comments, particularly the star ratings, so this study proposed the SVM method for the classification and the technique TF-IDF for text feature extraction. The presented results indicate that SVM attained an accuracy of 79%, a precision of 75% and a recall of 87%. So, this study contributed to how users can understand and navigate the opinions of others based on the star rating.

Another paper by Jain, A et. al, presents a framework for four different datasets about sentiment classification. The main purpose of this paper [7] is to improve the performance of sentimental classification models by reducing the dimensionality of the feature space through a two-step process: feature extraction and reduction. This study used four supervised classifiers: SVM, Random Forest, Naïve Bayes (NB), and Logistic Regression. NB classifier achieved the highest accuracy for the IMDB Movie Review dataset, with a value of 78.4%.

The research conducted by Alaparthi, S. et.al, compares supervised and unsupervised learning [8]. This study mainly focuses on deep learning, modelling the pre-trained Bert with an accuracy of 92%, which outperformed other deep learning models. Meanwhile, the techniques for the unsupervised model include using Sent WordNet, which had the lowest accuracy by 63%.

Singh, A. et. al, a proposed system of sentiment analysis of movie reviews using Bi-LSTM.The proposed system uses text vectorization, model building, evaluation, input sanitization and sentiment score calculation. The best accuracy was achieved by combining TF-IDF and ReLu activation functions.[15]

Nkhata, G. implemented fine-tuning BERT by adding BiLSTM and training the model on the movie reviews sentiment analysis. On film examines sentiment analysis, the BERT model achieved an accuracy of 92.28%.[16]

Akhtar, S. al; discusses the use and importance of BERT to understand the opinions and emotions of movie reviews. They provided a methodology for data preprocessing and model creation using TF-hub and TensorFlow Keras[17]. Results are displayed in every metric form, including confusion matrix,ROC curve and confusion matrix. The BERT model in this study achieved an accuracy of 92%.

4 Accomplishments

For our coursework, we used the same dataset shown in the proposal. To enrich our analysis, we added two more deep-learning models, RNN and pre-trained BERT. In our proposal, we suggested using Words of Bag (BoW), but we also used another method TF-IDF to compare which one works better for our analysis.

Task 1: This step included preprocessing the dataset and checking for the imbalanced dataset; in our case reviews. Also, we cleaned and formatted the text into a readable dataset for machine and deep learning models.

Task 2: Tokenization of the dataset; we presented it in a numerical format so that it could be helpful for the machine learning algorithms.

Task 3: We splitted the IBMD movie review data into two sets: training (80%) and testing (20%)

Task 4:We applied feature selection (Chi-Square) and feature extraction, Words of Bag (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF)

Task 5: We build one baseline model and three deep-learning algorithms. Our baseline model was logistic regression (LR). Meanwhile, the deep learning models were LSTM, KNN, and pre-trained BERT.

Task 6: The performance of deep learning models was further explored by applying features extraction methods, TF-IDF and BoW.

Task 5: The performance of RNN, LSTM and pre-trained BERT was successfully implemented and evaluated in different metrics, including the ROC curve and confusion matrix.

Task 6: The performance of all deep learning models was compared to the baseline model using different statistical measures.

5 Approach and Methodology

Our approach for sentiment analysis to successfully involves the implementation of the cleaning text data by removing stop words, HTML tags, square brackets, special characters, tokenization, and applying stemming and lemmatization technique.

After pre-processing part, we divided our dataset into training and testing,80% and 20%, respectively.

Before training the models, we pre-processed the text data using the TF-IDF technique and Bag

of Words (BoW), which converted the pre-processed text data into numerical so that the data could be fed to the neural networks. We also used chi-square feature selection to select the most essential features for the training models.

After pre-processing the text data and extracting features, we trained baseline and deep learning models. We experimented with various models.

We have selected one baseline model, Logistic Regression and three deep learning models: LSTM, RNN and pre-trained BERT. These models were chosen because our dataset possesses a sequential structure, where the organization of words plays a crucial role in interpreting the sentiment analysis. LSTM, and RNN are specifically designed to process sequential data in capturing contextual information efficiently. Also, we applied a pre-trained BERT model due to its adaptability to more than 100 languages and the capability of fine-tuning metrics [14].

We implemented several libraries, such as pandas, NLTK and BeatifulSoup, for manipulation and pre-processing. For feature extraction, we used CountVectorizer, TfidfVectorizer and Tokenizer from Keras. Also, we used specific libraries for different types of models. For LogiSTIC Regression, we used the sckit-learn library. We used LSTM and RNN from Keras for deep learning models. We also used the pre-trained model using BertForSequenceClassification, which was fine-tuned on a large corpus of text data.

5. Supervised deep learning models:

1. **RNN:** After we pre-processed the data tokenizing it and converting to text we applied RNN. We choose this deep learning model for its ability to capture the sequence of words in the review and analyse it to determine whether the sentiment is positive or negative. This model works by processing the terms of a review one at a time and updating its hidden state at every step. However, the RNN and the vanishing gradient problem can make the model difficult to train, leading to underperformance. Our RNN model consists of the embedding layer that maps the SimpleRNN layer. And a dense layer with a sigmoid activation function to output a binary sentiment classification.

2. **LSTM:** due to its architecture, LSTM has an advantage consists of a memory cell responsible for remembering information at each time step. Unlike RNN, LSTM avoids the vanishing problem, so LSTM can capture long-term dependencies, so in thus, the data can be utilized from previous tome steps for a more extended period. Since our data contains 50 K, the speed and memory were very challenging on our limited hardware resources.

3. **Pre-trained BERT*:*** This model is already trained on a large corpus of data, which can be adapted for various tasks, such as the movie review dataset. When fine-tuned on our dataset, it achieved the highest accuracy because of its pre-existing knowledge. Because of its architecture, having multiple layers of Transformer encoders, it allows the model to complement the complex language in the movie reviews. On the other hand, the size of the BERT model can be millions of parameters; it takes a significant time to train and consumes much memory, especially during training time.

6 Dataset

The IMDb movie review dataset is from the Kaggle website and will be used to develop an effective model for classifying sentiments as positive or negative. This dataset contains reviews from users who have watched and rated the movies. Every review has a label based on their opinion about the positive and negative film. It includes 50,000 movie reviews; the sentiment distribution is 25,000 positive and 25,000 negative. Also, this dataset contains 101895 unique words (Figure. 2), where the most common words are “the”,”and”,”of” and “this”. The dataset was already balanced so we didn’t have to balance it, as it is shown in Figure 1.

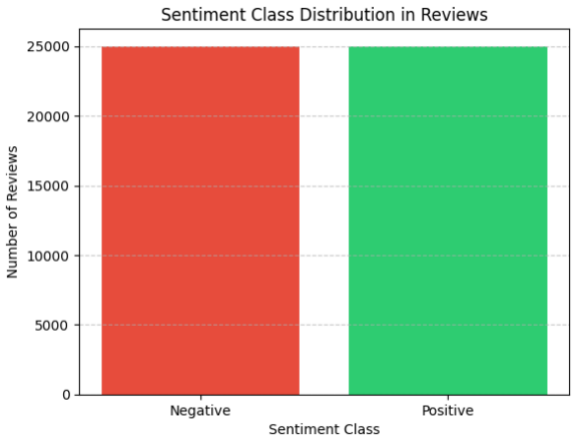


Figure. 1

Below, we are showing an example of positive ang negative review from our dataset:

- Positive review: "I loved this movie! The plot was engaging, and the acting was superb. I highly recommend it to anyone who enjoys a good thriller."
- Negative review: "Unfortunately, this film was a huge disappointment. The storyline was predictable, and the acting felt forced. I wouldn't recommend it."

There were some properties of the data that made our task challenging, such as:

- Informal language: "This movie was lit! The acting was on point, and the plot kept me hooked. Def a must-watch!"

  From this review, we can see that the reviewer uses words like: "lit," "on point," and "def," which can be challenging for a model to understand the sentiment.
- Ambiguity and irony: "Oh great, another superhero movie. Just what we needed."

  In this review, the commentator uses a touch of sarcasm to express the feelings related to the movie. But for the sentiment analysis model, it may interpret as a positive review; meanwhile, it expresses their disapproval.

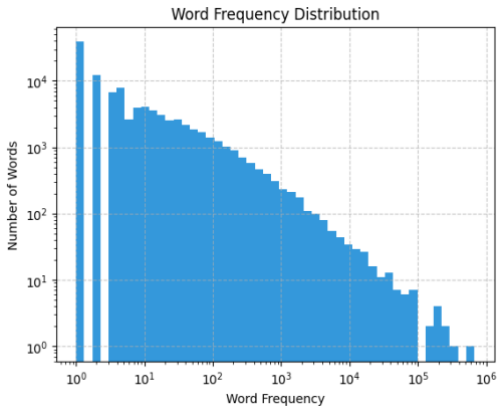


Figure. 2

6.1 Dataset preprocessing

Part-of-Speech (POS) Tokenization: In this step, each IMDB movie review is tokenized to gain valuable information about the grammatical structure of the sentences in the review column. We utilized the pos_tokenizer () function to create a new representation of the movie review text and assigned their grammatical roles.

Stop words removal: in this pre-processing part, we first printed the stop words that were part of the dataset. Then we eliminated all of the stop words.

HTML Tags Removal: we used the 'BeautifulSoup' library to remove any HTML tags that may be present.

Cleaning text: we created this function to remove punctuation, digits, square brackets, memorable characters, and other noisy text elements that may be present.

Text Stemming: we applied the **simple_stemmer ()** function to stem the review text data, which reduces the words to their root forms. For example, the stemming function converted the terms "happily," "happiness," and "happy" to their root form, "happy". Doing this allows us to gather the exact group words with similar meanings.

Text Lemmatization: We used the simple_lemmatizer() function to transform the words into their base forms to lemmatize the text data. The terms "worse" and "worst" would be lemmatized to the "bad" in the IMDb movie review dataset.

Tokenization:

Feature extraction: we used two feature extraction techniques.

- **TF-IDF:** we utilized this technique to extract features from the pre-processed IMDB movie review dataset. This technique calculates the importance of words in the dataset. The max features parameter was set to 5000; to limit the dimensionality of the feature space, the training model can be more efficient.
- **Bag of words:** this approach counts the occurrence of each word in the review column; it provides a numerical representation of text data about the reviews.

8. Baselines

We applied Logistic Regression as our baseline model. The logistic regression algorithm uses a logistic function to convert the input features into a probability distribution over the output labels, which can be used to predict the sentiment label of a given text data as positive or negative. The reason we chose LR, it's for the fact that it is an algorithm widely used for binary classification. One limitation that this algorithm could have is the ability to handle complex nuances in the language used in the movie reviews.

9. Results, error analysis

We trained one machine learning model and three deep learning methods to evaluate or model. We also applied two feature extraction methods, TF-IDF and BoW, to the LSTM and RNN models.

- ***Baseline model, Logistic Regression***

Model	Accuracy
LR (TF-IDF)	89%
LR (BoW)	89%

Table. 1

In our experiment, two models for baseline models were trained Logistic Regression with different extraction techniques, BoW and TF-IDF.The accuracy of both models is shown in the Table. 1 was found to be the same, 89%. This suggests that both feature extraction works well on the LR model. Furthermore, the feature selection improved the performance of the models.

We plotted the confusion matrix to analyze further which feature extraction works best on the baseline model. Looking at the confusion matrices, we can tell that the BoW model has more false positives than the TF-IDF; knowing that we tend to avoid false positives (avoiding predicting negative reviews as positive), the LR model with TF-IDF model may be more effective, because it will give less weight to the familiar words.

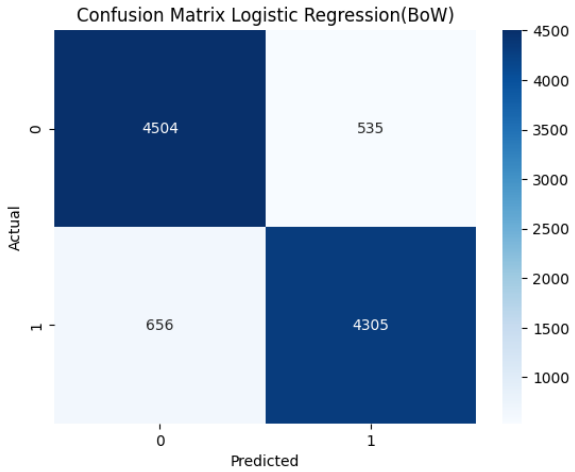


Figure. 1

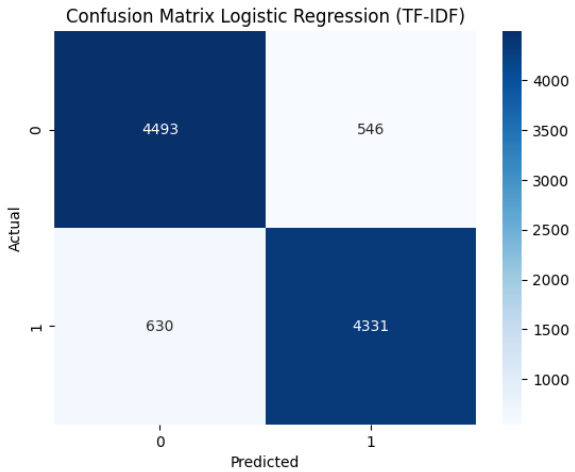


Figure. 2

B. Deep learning models

- ***LSTM***

Model	Accuracy
LSTM (TF-IDF)	87.57 %
LSTM (BoW)	87.47 %
LSTM	78.09%

Table. 2

We used two different feature extraction for the LSTM model: BoW and TF-IDF. Based on our findings, both techniques perform well with the LSTM model. This highlights the importance and effectiveness of these methods in extracting the relevant features from the text data, particularly in the movie industry data. The movie industry needs to have significant findings because, based on the findings, they gain a better understanding of audience preferences and improve the profitability

of the movie industry. Based on the confusion matrix, we can see that the LSTM model using TF-IDF feature extraction outperformed the other model using BoW. The TF-IDF model correctly predicted 4446 instances of negative sentiment and 4324 instances of positive sentiment while misclassified only 539 negatives and 637 positives. Meanwhile, the BoW model predicted 4207 and 499 negative and positive cases, respectively. Also it misclassified 832 negatives and 512 positives. So we can conclude with the result that LSTM with TF-IDF performs better in predicting sentiment in IMDb movie reviews.

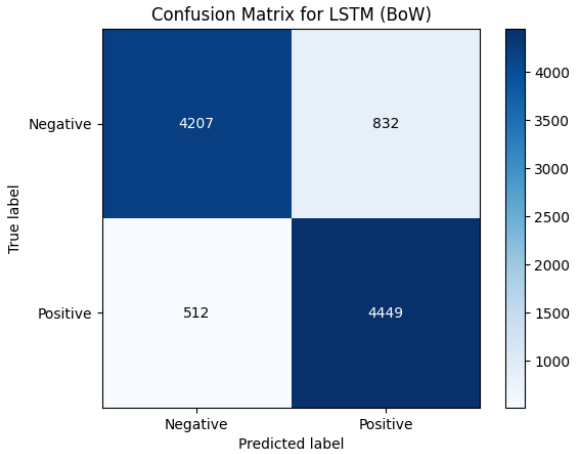


Figure.3

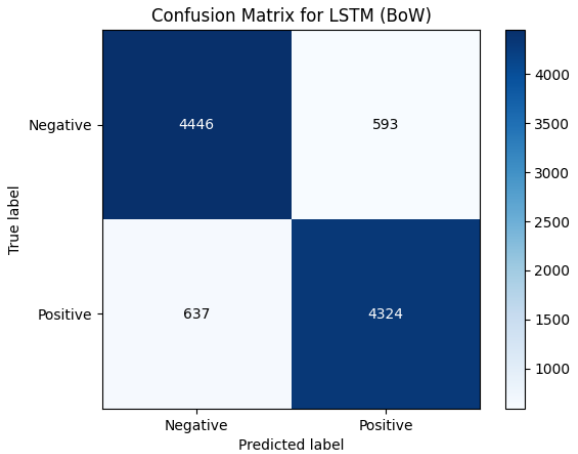


Figure. 4

We also trained LSTM on pre-processed data to compare its performance with the feature extraction model. The accuracy of this model was 78.09% lower than other models. This suggests that applying the feature extraction technique improved the accuracy of the sentiment analysis model. For the movie industry movie, it is imperative to highlight this because it improves the decision-making process based on sentimental analysis.

- ***RNN***

Model	Accuracy
RNN (TF-IDF)	87.85%
RNN (BoW)	87.70 %
RNN	54.44 %

Table. 3

We will experiment with RNN the same way we did with LSTM. Based on the result of the RNN model with feature extraction, both performed well, but with a very slight difference, like LSTM, RNN performs better with TF-IDF. Our basic RNN model performed with an accuracy of 54.44%, suggesting that for the RNN model, the feature extraction was very effective in capturing the crucial reviews.

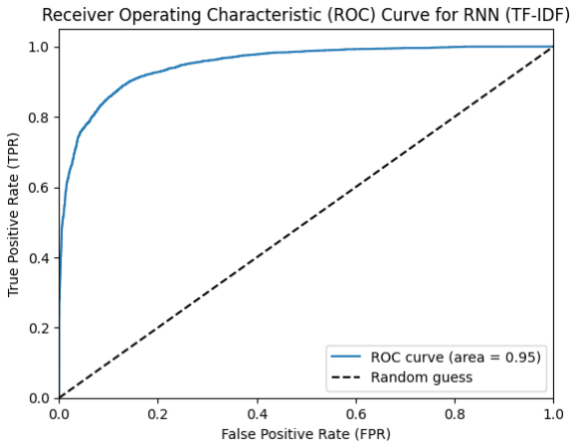


Figure. 4

Adding also the fact that the ROC curve for the TF-IDF resulted better, than the BoW, with a difference of 1%.

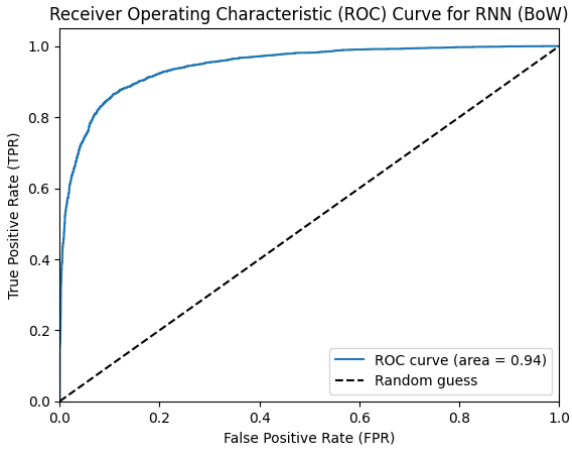


Figure. 5

As we can see from the Table. 3, the RNN model without any feature extraction performed better,

with an accuracy of 54.44%. We can tell the same thing as we mentioned above about why feature extraction did not give a good accuracy. But unlike the LSTM model that achieved a higher accuracy, the RNN has almost the same accuracy as the other model with feature extraction.

- ***Pre-trained BERT model***

Our fine-tuned BERT model outperformed all the models we trained in our dataset, with an accuracy of 92% . As it is shown in Figure. 6, the confusion matrix represents that the model correctly predicted 4715 instances of negative sentiment and 4532 instances of positive sentiment, meanwhile misclassified 324 negatives and 429 positives.

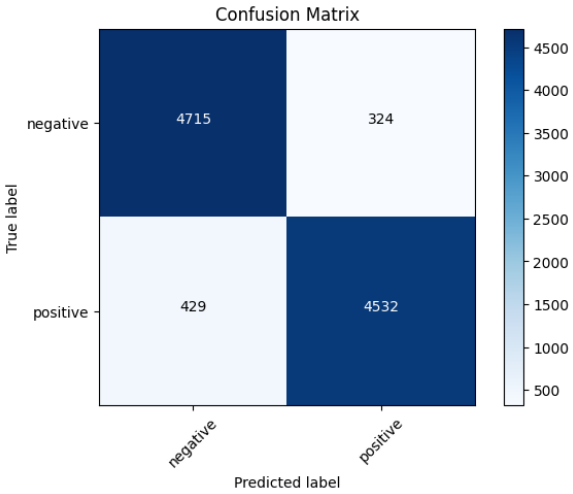
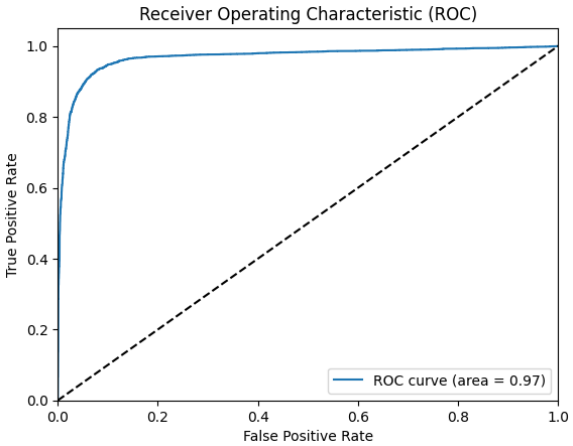


Figure.6

Furthermore, we plotted the ROC curve, which has a value of 97%, suggesting that our model can proficiently differentiate between reviews that comment positive reviews about a film and those that convey negative perspectives.



Lessons learned and conclusions

This project taught several lessons about how to adapt feature extraction to sentiment analysis. Both feature extraction methods significantly impacted optimizing the accuracy of Logistic Regression, RNN and LSTM.It is essential to mention that the baseline model, which involved simple feature extraction performed very well, achieving an accuracy of 89 %. Another important lesson learned is that a pre-trained model can provide, in our case BERT model, provide significant advantages. This model had itself the tokenizer part and embedding layers, which means we didn't had apply feature extraction, and still achieved to have the highest accuracy.Pre-trained model achieved the highest accuracy with 92%,as we proposed at our hypothese.

The TF-IDF technique feature extraction method, was slightly better than BoW, in all the models we applied it. Even though it was an small difference-IDF resulted to perform better in movie review sentiment analysis.

References

[1]Ibrahim, M., Bajwa, I.S., Ul-Amin, R., & Kasi, B. (2019). A Neural Network-Inspired Approach for Improved and True Movie Recommendations. Computational Intelligence and Neuroscience, 2019,6701398.

[2]Ali, N. M., Abd El Hamid, M. M., & Youssif, A. (2019). Sentiment Analysis for Movies Reviews Dataset Using Deep Learning Models. International Journal of Data Mining & Knowledge Management Process (IJDKP), 9(2/3), 19-38. DOI: 10.5121/ijdkp.2019.

[3]Saraswathy, R., & Kavitha, K. (2018). Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method. *ResearchGate*. Retrieved https://www.researchgate.net/publication/330014159_Sentiment_Analysis_on_IMDb_Movie_Reviews_Using_Hybrid_Feature_Extraction_Method

[5] Kumar, H. M. K., Harish, B. S., & Darshan, H. K. (2018). Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method. Procedia Computer Science, 132, 104-113.

[6] Ramadhan, N. G., & Ramadhan, T. I. (2021). Analisis Sentimen Opini Beberapa Komentar Pengguna Situs Web IMDB Menggunakan Metode

Support Vector Machine (SVM). Jurnal Sinkron, 6(1), 208-219. Retrieved from https://jurnal.polgan.ac.id/index.php/sinkron/article/view/11204

[7] Jain, A., & Jain, V. (2021). Efficient Framework for

Feature Reduction. EAI Endorsed Transactions on Scalable Information Systems, 8(29). https://doi.org/10.4108/eai.16-2-2021.168715

[8] Alaparthi, S., & Mishra, M. (2021). Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey. Decision Support Systems, 146, 113511. doi: 10.1016/j.dss.2021.113511

[9] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[10] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5-6), 602-610.

[11] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

[12]Zhang, Y., Zong, C., & Lv, Y. (2019). Using BERT for Checking the Polarity of Movie Reviews. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 190-195).

[13] Akhtar, S., & Khan, S. (2019). Using BERT for Checking the Polarity of Movie Reviews. International Journal of Advanced Computer Science and Applications, 10(12), 386-392. doi: 10.14569/IJACSA.2019.0101243

[14] Nascimento, R., de Souza, J. M., & Batista, G. E. A. P. A. (2022). A comparative analysis of pre-trained language models for sentiment analysis in the movie review domain. Electronic Commerce Research, 1-24. https://doi.org/10.1007/s10660-022-09560-w

[15] Singh, A., Thapliyal, R., Vanave, R., Shedge, R., & Mumbaikar, S. (2022). Analysis of hyperparameters in Sentiment Analysis of Movie Reviews using Bi-LSTM. In S. Panda, M. Pattnaik, & R. Das (Eds.), *2022 International Conference on Advanced Computing and Communications (ICACC)* (pp. 1-8). ITM Web of Conferences. https://doi.org/10.1051/itmconf/20224803012

[16]Nkhata, G. (2022). Movie Reviews Sentiment Analysis Using BERT. (Graduate Theses and Dissertations). University of Arkansas, Fayetteville. Retrieved from ScholarWorks@UARK.

[17] Nehal Mohamed Ali, Marwa Mostafa Abd El Hamid, & Aliaa Youssif. (2021). Sentiment analysis for movies reviews dataset using deep learning models. Faculty of Computer Science, Arab