Text Classification on IMDb Reviews:A Comprehensive Comparison of BERT, LSTM, and Traditional Models

Asad Uzzaman
sdzzaman@memphis.edu
U00930541

Kamrun Naher
knaher@memphis.edu
U00769436

Abstract

Text analysis, especially on a large scale, is both time-consuming and complex when performed manually. With advancements in technology, there is an increasing need for efficient methods to analyze customer feedback, enabling improvement and cost savings. Natural Language Processing (NLP) has emerged as a crucial approach for addressing such challenges. In this project, we focus on the critical task of text classification, specifically sentiment analysis. Utilizing the IMDb dataset for training for all models and testing on LSTM, we evaluated various models, including deep learning approaches like BERT and LSTM, alongside traditional machine learning models such as Logistic Regression, Naïve Bayes, Linear SVM, and XGBoost. Our comparative study highlights that deep learning models, particularly LSTM, demonstrated superior performance in accuracy and handling sequential data, while BERT provided strong contextual understanding. Traditional models, though computationally efficient, were outperformed in terms of accuracy. Our study found better performance with deep learning approaches compared to other traditional models in terms of accuracy and effectiveness in the sentiment classification task.

Introduction

This project focuses on sentiment classification of movie reviews collected from IMDb, a binary classification task aimed at determining whether a given review expresses a positive or negative sentiment. Sentiment analysis, also known as opinion classification, plays a crucial role in understanding public perception, influencing consumer behavior, and shaping marketing strategies. To address this task, we evaluate various models, including deep learning techniques such as LSTM and BERT, as well as traditional machine learning models like Logistic Regression, Naïve Bayes, Linear SVM, and XGBoost.

LSTM (Long Short-Term Memory) models are particularly well-suited for sequential data tasks like sentiment analysis due to their ability to capture long-term dependencies and maintain context across sequences. This characteristic enables LSTM to deliver superior performance in understanding linguistic structures, making it an ideal candidate for this study.

On the other hand, BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art language model known for its context-encoding accuracy. BERT's bidirectional nature allows it to grasp complex relationships within the text, making it highly effective for tasks like sentiment classification. While BERT offers strong contextual understanding, it requires significant computational resources, which we also evaluate in this study.

Additionally, traditional machine learning models such as Logistic Regression, Naïve Bayes, Linear SVM, and XGBoost were included to provide a comparative baseline. These models, while computationally efficient, are often less accurate compared to deep learning approaches, but they remain valuable for applications requiring rapid predictions and minimal resource consumption.

This project explores contemporary deep learning techniques like LSTM and BERT to address sentiment analysis while benchmarking them against traditional methods. By examining the strengths, limitations, and trade-offs of each approach, the study provides insights into their practical implications for sentiment analysis tasks, highlighting the balance between performance, efficiency, and resource requirements.

Proposed Approach

The methodology focuses on evaluating and comparing the performance of various models, starting LSTM networks. LSTM was chosen because of its strong ability to process sequential data and learn long-term dependencies, making it well-suited for sentiment classification tasks (Behera et al., 2021). Its architecture allows it to effectively analyze the structure of IMDb movie reviews, making it a reliable choice for handling text data.

Alongside LSTM, the study also incorporates BERT, which is widely recognized for its ability to capture deep contextual relationships in text. BERT’s fine-tuning process enables it to adapt to specific datasets, which can improve sentiment classification accuracy for IMDb reviews (Arora et al., 2023). Despite its high computational demands, BERT

remains a powerful tool for analyzing complex linguistic data.

For comparison, the study also examines simpler models such as Logistic Regression, XGBoost, Linear SVM, and Multinomial Naïve Bayes. These models are less computationally intensive and faster, making them useful benchmarks for understanding trade-offs between speed, resource efficiency, and accuracy. By including both deep learning models like LSTM and BERT, as well as traditional machine learning approaches, the study provides a well-rounded analysis of performance, efficiency, and scalability across different tools.

The BERT model

It uses a large-scale corpus to train and obtain text containing rich semantic information, which is semantic representation of the text, and then fine-tune that text for a specific NLP task and then finally apply it to that specific NLP task (Wu et al.,2024). Attention Mechanism is another important key. In essence, Attention is to select a small amount of important information from a large amount of information, and focus on this important information, ignoring most of the unimportant information.

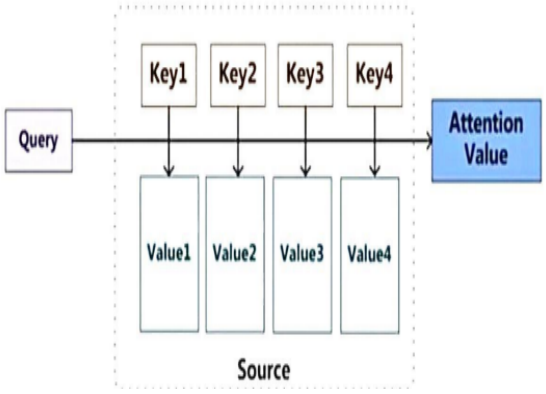

Figure 1: Bert model Attention mechanism architecture diagram [source: (Wu et al. 2024)]

The greater the weight, the more it focuses on its corresponding Value, that is, the weight represents the importance of information, and Value is its corresponding information (Wu et al. 2024).

LSTM MODEL

The Long Short-Term Memory (LSTM) network, a type of recurrent neural network (RNN), is particularly well-suited for sequential data, such as text. LSTMs work on long term dependency, which is designed to ignore the long-term dependency problem. Remembering information for a long time is something so usual for them, they do not struggle with this. All recurrent neural networks have the form of a chain of repeating modules of the neural networks (Murthy et al.,2020). A classical LSTM cell and an LSTM cells chain have been shown in figure 2. The network learns by itself what to keep and what to skip. LSTM learns long-range dependencies based on this mechanism (Murthy et al, 2020).

Traditional MODEL

Machine models explored in this analysis were Logistic Regression, XGBoost, Linear SVM, and Multinomial Naïve Bayes. Multinomial Naive Bayes is a variation of Naive Bayes. It is designed for tasks like word count etc. It works under the assumption of conditional independence. Logistic regression is a method for classification problems. Logistic regression predicts probability of a categorical outcome. mostly binary like positive, negative, true, false. XGBoost is a powerful and efficient implementation of the **gradient boosting** algorithm designed for supervised learning tasks such as classification and regression. It is widely known for its speed, performance, and scalability. Linear support vector machine is a supervised algorithm for classification.

For classification linear SVM aims to maximize the margin between two classes, where the margin is defined as the distance between the hyperplane and the nearest data points from each class (called **support vectors**).

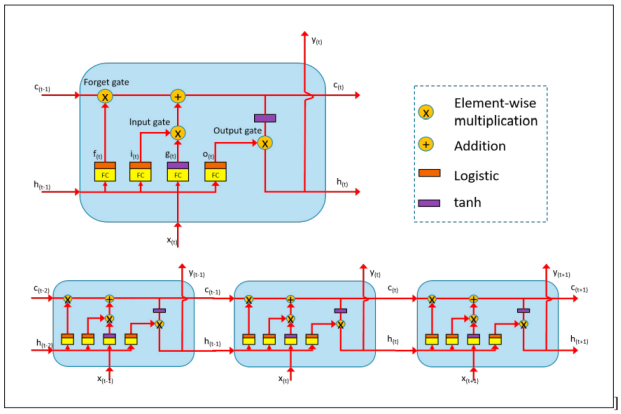

Figure 2: The architecture of LSTM Model. [Source:(Liu et al.2019)]

The evaluation metrics for these models encompass training time, feature count, and accuracy. **Accuracy** measures the proportion of correctly classified reviews relative to the total number of reviews, expressed as a percentage. **Training time** represents the duration required to train the model. **Feature count**, applicable only to traditional models, refers to the number of features employed by these models.

Experiments and Results

Dataset

We use IMDb dataset which contains movie reviews and polarity labels for our task. The reviews in the dataset are rated by users which helps to derive the polarity from the content directly (Haque et al., 2019). The diversity and varying review lengths make the dataset a good choice to train and test the models on complicated phrasing to extract sentiment. The dataset consists of 50,000 entries, with each row containing a "review" and a corresponding "sentiment" label. A snapshot from the dataset is given in figure 3.

For our task, we have converted the target variable to numeric labels, where sentiment is 1 if it is positive and 0 if it is negative. Initially, we performed data cleaning and preprocessing, which resulted in a refined dataset ready for model training. We then split the dataset into two parts: 80% for training and 20% for testing. Using the training data, we applied various machine learning and deep learning techniques to build a model. Afterward, we evaluated the model on the test data.

	review	sentiment
0	one reviewer mentioned watching oz episode hoo...	1
1	wonderful little production filming technique ...	1
2	thought wonderful way spend time hot summer we...	1
3	basically family little boy jake think zombie ...	0
4	petter matteis love time money visually stunni...	1

Figure 3: Overview of using Numerical labels on IMDB dataset.

For data analysis, we generated word clouds for both positive and negative sentiments, which aided in visualizing the distribution of words. Additionally, we created graphs to represent the word count distribution across reviews. Lastly, we calculated and compared the accuracy of the different machine learning techniques used.

The preprocessing of the dataset involved several steps. Initially, the text was converted to lowercase, and stop words were removed to eliminate irrelevant terms. For deep learning models, tokenization and padding were applied, while feature conversion was performed for traditional machine learning models.

From the data analysis, it was observed that positive reviews tend to be longer than negative ones. After preprocessing, the vocabulary size was reduced to approximately 12,000 unique words from an initial size of 126,176. The resulting training feature matrix has a shape of (39,665, 12,176), and the test feature matrix has a shape of (9,917, 12,176).

Results

To perform sentiment classification for analysis purposes, several models were built, and their performances were compared.

Table 1. Model Performance Comparison

Model	Training Accuracy (%)	Training Time (sec)	Test Accuracy (%)
Log. Reg.	89.11	21.20	-
Naïve Bayes	86.8	0.10	-
XGBoost	84.38	51.73	-
Linear SVM	89.57	7.92	-
BERT	92.1	360.78	-
LSTM	95.8	186.79	88.4

From table 1. It can be said that the LSTM model achieved the **best training accuracy of 95.8%**, showcasing its ability to learn long-term dependencies, which is essential for modeling sequential data like text. To evaluate its generalization ability, the test accuracy was measured at **88.42%**, highlighting a slight drop compared to its training performance. Its training time of **186.79 seconds** indicates higher resource requirements compared to simpler models but demonstrates its strength in pattern recognition.

The fine-tuned BERT model demonstrated strong accuracy of **92.1%**, though at a significantly higher computational cost with a training time of **360.78 seconds**. This result reflects BERT's powerful architecture, which effectively captures contextual relationships in text data, making it highly suitable for complex text classification tasks.

Logistic Regression and Naïve Bayes showed comparatively lower accuracies (**89.11%** and **86.8%**, respectively). However, they excelled in terms of efficiency, with training times of **21.20 seconds** and **0.10 seconds**, respectively. These models are ideal for scenarios where computational resources are limited, and rapid predictions are prioritized.

Similarly, XGBoost and Linear SVM offered a balance between accuracy and computational cost. XGBoost achieved **84.38% accuracy** with a training time of **51.73 seconds**, while Linear SVM reached **89.57% accuracy** in just **7.92 seconds**. These models remain competitive for text classification tasks due to their efficiency and generalization capabilities.

Overall, this comparative analysis highlights the trade-offs between accuracy and training efficiency, helping to guide model selection based on application-specific needs, such as accuracy requirements, computational constraints, or resource availability. The LSTM model's high training accuracy and consistent test performance make it a strong

contender for text classification tasks, while BERT provides less training accuracy at the cost of higher computational resources.

Discussion

The experimental findings emphasize the strengths and trade-offs of each model, as seen in the scalability graph from figure 5 LSTM emerged as the most accurate model, showcasing its capability to capture long-term dependencies and intricate linguistic patterns effectively. However, its sequential nature limits its scalability for larger datasets, as its performance declines with increasing dataset size.

In contrast, BERT demonstrated excellent scalability, maintaining its performance as datasets grew larger. This scalability advantage stems from its parallel processing ability, though it comes at the expense of higher computational demands and longer training times.

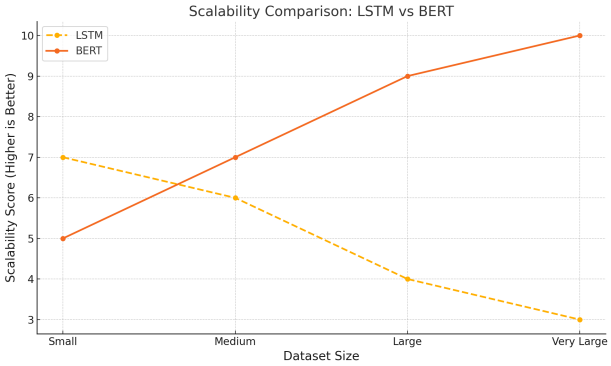

Figure 5: Scalability comparison of LSTM and BERT.

Simpler models like XGBoost and Linear SVM provide a middle ground, delivering competitive accuracy and fast training times. They are well-suited for tasks involving medium to large datasets where efficiency is important. Logistic Regression and Naïve Bayes, while excelling in speed and low computational requirements, are better suited for straightforward tasks with smaller datasets due to their comparatively lower accuracy. The scalability graph highlights BERT as the ideal choice for large-scale applications requiring contextual depth, while LSTM remains the preferred option for medium to large datasets where achieving peak accuracy is crucial.

References

Haque, M. R., Lima, S. A., & Mishu, S. Z. (2019). Performance analysis of different neural networks for sentiment analysis on IMDb movie reviews. In 2019 3rd International conference on electrical, computer & telecommunication engineering (ICECTE) (pp.161-164). IEEE.

Arora, K., Gupta, N., & Pathak, S. (2023). Sentimental analysis on imdb movies review

using bert. In 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 866-871). IEEE.

Behera, R. K., Jena, M., Rath, S. K., & Misra, S. (2021). Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. Information Processing &

Management, 58(1), 102435.

Wu, Y., Jin, Z., Shi, C., Liang, P. and Zhan, T., 2024. Research on the Application of Deep Learning-based BERT Model in Sentiment Analysis. *arXiv preprint arXiv:2403.08217*.

Murthy, G.S.N., Allu, S.R., Andhavarapu, B., Bagadi, M. and Belusonti, M., 2020. Text based sentiment analysis using LSTM. *Int. J. Eng. Res. Tech. Res*, *9*(05).

Liu, X., Gherbi, A., Li, W. and Cheriet, M., 2019. Multi features and multi-time steps LSTM based methodology for bike sharing availability prediction. *Procedia Computer Science*, *155*, pp.394-401.