Aspect-based sentiment analysis: A study of the IMDB review database

R. Mills SN:7168755, rm718@uowmail.edu.au
K. Goel SN:7836685, kg956@uowmail.edu.au
B. Sensha Shrestha SN:8447196, bss541@uowmail.edu.au
M. Faruk SN:7056849, mzf395@uowmail.edu.au
School of Computing and Information Technology
University of Wollongong
CSCI933,

May 31, 2024

Abstract

This report presents a detailed exploration of aspect-based sentiment analysis applied to the IMDB review database. Utilizing advanced deep learning techniques, specifically Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM), we systematically analyzed sentiments within movie reviews. The CNN model effectively parsed general sentiments, achieving an accuracy of 89%, whereas the LSTM model, adept at capturing sequential data dependencies, exhibited performance with a 87% accuracy rate. Furthermore, we delved into aspect-specific sentiment analysis using Latent Dirichlet Allocation (LDA) to classify sentiments related to distinct movie elements—actors, plot, music, and director. This nuanced approach to dissecting thematic content allowed for an accurate classification of sentiments at a granular level with a commendable combined accuracy of 78%.

1 Introduction

The surge in access to media through streaming and content platforms has enabled consumers to share and access opinions on a range of media. An opinion on a particular media can take on many different forms and often display some form of positive or negative sentiment. Sentiment analysis is a subfield of natural language processing (NLP) that aims to identify the sentiment expressed within a textual format.

Our study maintains a focus on movie reviews as it is an interesting domain that contains many subtle nuances and challenges. Analysing sentiment in movie reviews is considered more challenging than in other types of reviews. For example, in movie reviews, an evil character or a tragic storyline does not make a movie bad (Thet, Na, & Khoo, 2010).

Aspect based sentiment analysis (ABSA) can address these challenges by focusing on specific aspects of the movie the reviewer is commenting on and also whether they like them or not (Farhadloo & Rolland, 2016). ABSA is powerful in examples such as 'Food is decent but service is so bad' which contains positive sentiment towards aspect food but strong negative sentiment towards aspect service. Classifying the overall sentiment as negative would neglect the fact that the food was actually good (Wang, Xu, & Zu, 2021). ABSA consists of an aspect extraction stage and sentiment classification stage. Aspect extraction gathers terms that are often related to a given aspect. Sentiment classification attributes a positive or negative sentiment based on the text (Onalaja, Romero, & Yun, 2021).

2 Literature review

This section reviews existing methods and implementations of aspect based sentiment analysis, outlining their strenghts and weaknesses. Phan and Ogunbona (2020) method combines part-of-speech embeddings, dependency-based embeddings, and contextualised embeddings. This approach improves accuracy by effectively understanding syntactical information such as sentence structure and relationships between words. However, the reliance on complex syntactical features may introduce additional computational overhead and require more extensive training data.

The methods described in the study by Onalaja et al. (2021) analyses movie review by focusing on specific aspects of the content, allowing for a more nuanced understanding of sentiment compared to traditional sentiment analysis methods. The complexity of accurately identifying and categorising different aspects within the text, can be challenging and time-consuming.

Wang et al. (2021) explore the use of deep neural networks to enhance aspect based sentiment analysis. The proposed approach is able to capture complex relationships between aspects and sentiments. Similar to other proposed solutions, their is a high computational complexity.

Thet et al. (2010) method effectively handles informal and varied language typical of discussion boards. However, it requires significant preprocessing and manual feature selection, which can be time-consuming and less adaptable

Utilising techniques like topic modelling and sentiment lexicon construction, Rybakov and Malafeev (2018) attempts to extract the sentiment in Russian hotel reviews. The performance metrics for the negative sentiment reviews indicate room for improvement compared to existing literature.

Pannala, Nawarathna, Jayakody, Rupasinghe, and Krishnadeva (2016) leveraged supervised learning techniques to create structured model framework allows precise control over features. However, it requires extensive labelled data and feature engineering, making it less flexible and scalable.

Through the combination of LDA and BERT, Lohith, Chandramouli, Balasingam, and Arun Kumar (2023) leverages LDA's ability to extract aspects and BERT to capture contextual information and classify the extracted aspects. Combining these methods is processing-intensive which can weaken its efficiency.

3 Methods

Our study, drawing on research conducted in this domain, integrates both traditional unsupervised and supervised deep learning methods to extract and analyze sentiments associated with specific aspects of movie reviews. The methodology is segmented into several key components: data preprocessing, tokenizer utilization, model management, and sentiment classification.

3.1 Data Preprocessing

The raw movie reviews are initially subjected to a cleaning process that removes unnecessary elements such as HTML tags, URLs, and non-standard characters. For aspect extraction, we employed traditional NLP approaches, including sentence clause extraction and text lemmatization, to refine the text for further analysis

3.1.1 Subword Tokenization

This method involves breaking words into smaller, more manageable units (subwords), which helps in handling the morphological richness of words in the reviews. The tokenizer builds a vocabulary of subwords based on the frequency of their occurrence within the dataset. This vocabulary is then used to encode the text data into sequences of subword indices, facilitating more efficient handling of unknown words and morphological variations.

3.2 Model Management and Training

The tokenized data is processed through a series of neural network models, managed by a custom model manager that handles tasks such as model loading, data input preparation, and training orchestration:

- **CNN Model**: A Convolutional Neural Network (CNN) is employed to analyze the spatial structure of the text data. By using convolutional layers, the model can capture local dependencies and patterns within the text, such as phrases and common word sequences that are indicative of sentiment. This feature extraction is critical for understanding the overall sentiment expressed in the reviews, especially when analyzing complex emotional expressions.

- **LSTM Model**: A Long Short-Term Memory (LSTM) network is used to leverage the sequential nature of text data, allowing the model to maintain contextual information over longer text sequences. This is particularly beneficial for capturing the nuances and dependencies of sentiments expressed across sentences in the reviews.

Both models are trained on labeled sentiment data from the IMDB review dataset, utilizing binary cross-entropy loss functions and optimization algorithms to minimize error and enhance prediction accuracy.

3.3 Aspect-Based Sentiment Analysis

Aspect extraction is performed using Latent Dirichlet Allocation (LDA) to identify prevalent topics or aspects within the reviews. Each aspect is then associated with a sentiment score derived from the sentiment classification models. This involves:

- **Topic Modeling**: Clauses from the reviews are extracted and fed into the LDA model to identify significant aspects.
- **Sentiment Classification**: Each extracted aspect is then classified into one of four categories—actor, plot, music, and director—based on the keywords present in the topic. The text is subsequently evaluated by the trained models to determine whether the sentiment is positive or negative.

3.4 Integration and Final Evaluation

The final stage integrates the outputs from both CNN and LSTM models to deliver a final judgment on the polarity of sentiment. The effectiveness of the combined approach is evaluated using standard metrics such as accuracy and precision, providing a comprehensive view of model performance in distinguishing and correctly classifying sentiments in movie reviews.

Rationale for Methodological Choices:

The integration of subword tokenization and LSTM models is designed to harness the strengths of both morphological analysis and deep sequential context understanding. This hybrid approach aims to enhance the model's ability to accurately reflect the complex and nuanced expressions found in movie reviews, which are essential for effective aspect-based sentiment analysis. The choice of neural network strategies and LDA for aspect extraction is based on their proven effectiveness in handling large datasets and complex linguistic structures, as demonstrated in previous research within the field of natural language processing.

3.5 Architecture

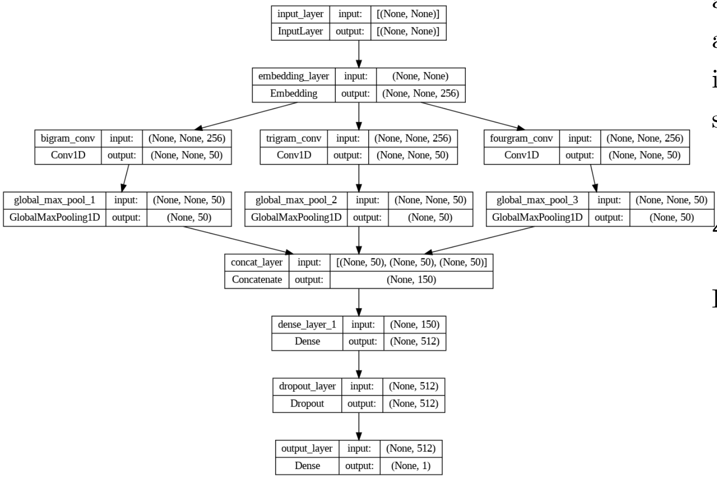
input_layer | input: | [(None, None)]
InputLayer | output: | [(None, None)]
embedding_layer | input: | (None, None)
Embedding | output: | (None, None, 256)
bigram_conv | input: | (None, None, 256)
Conv1D | output: | (None, None, 50)
trigram_conv | input: | (None, None, 256)
Conv1D | output: | (None, None, 50)
fourgram_conv | input: | (None, None, 256)
Conv1D | output: | (None, None, 50)
global_max_pool_1 | input: | (None, None, 50)
GlobalMaxPooling1D | output: | (None, 50)
global_max_pool_2 | input: | (None, None, 50)
GlobalMaxPooling1D | output: | (None, 50)
global_max_pool_3 | input: | (None, None, 50)
GlobalMaxPooling1D | output: | (None, 50)
concat_layer | input: | [(None, 50), (None, 50), (None, 50)]
Concatenate | output: | (None, 150)
dense_layer_1 | input: | (None, 150)
Dense | output: | (None, 512)
dropout_layer | input: | (None, 512)
Dropout | output: | (None, 512)
output_layer | input: | (None, 512)
Dense | output: | (None, 1)

Figure 1: CNN Architecture

The Convolutional Neural Network (CNN) architecture, shown in Figure 1, is built with three 1D convolutional layers that capture bi-gram, tri-gram, and four-gram features, respectively, with kernel sizes of 2, 3, and 4. A Max Pool layer is applied to down-sample the feature maps, reducing their dimensionality and highlighting the most significant features.

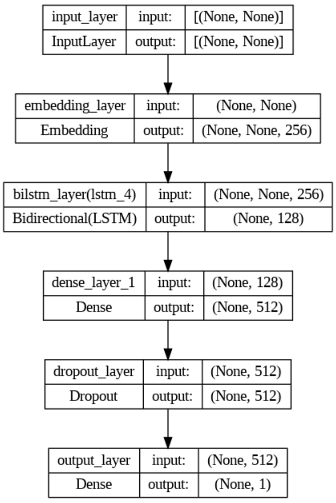
input_layer | input: | [(None, None)]
InputLayer | output: | [(None, None)]
embedding_layer | input: | (None, None)
Embedding | output: | (None, None, 256)
bilstm_layer(lstm_4) | input: | (None, None, 256)
Bidirectional(LSTM) | output: | (None, 128)
dense_layer_1 | input: | (None, 128)
Dense | output: | (None, 512)
dropout_layer | input: | (None, 512)
Dropout | output: | (None, 512)
output_layer | input: | (None, 512)
Dense | output: | (None, 1)

Figure 2: LSTM Architecture

The Long Short-Term Memory (LSTM) network, depicted in Figure 2, utilizes a single LSTM layer to capture long-term dependencies in the sequential data, which is crucial for maintaining context over long passages of text.

For topic modeling, we chose Latent Dirichlet Allocation (LDA) and trained the model to identify 10 distinct topics in the text. Figure 3 illustrates the most salient terms identified by the model, highlighting the key terms that dominate the text corpus and providing insights into the main themes and topics present in the data.

Figure 3 presents the LDA topic graph, showing the distribution of topics across the text corpus. This provides a visual representation of how different topics are related and distributed within the data. By analyzing these topics, we can gain a deeper understanding of the underlying structure and themes of the text.

4 Experiments

In this study, we carried out the following experiments:

1. **Model Comparison:** Which Model is Best Suited for Sentiment Analysis?
   This experiment aims to determine the most effective neural network architecture for sentiment analysis by comparing the performance of Convolutional

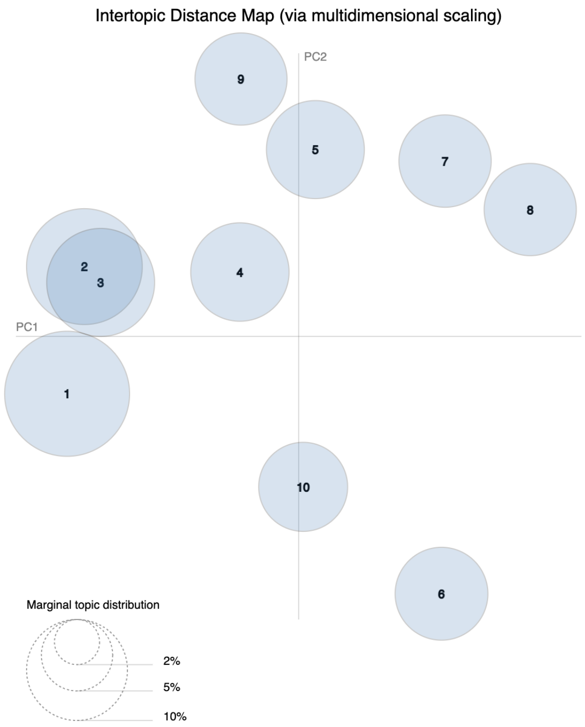

Figure 3: 10 Topic Graph Filtered using LDA

Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. We explore whether one model consistently outperforms the other across various metrics such as accuracy, precision, and recall, or if the superiority of a model depends on specific characteristics of the dataset.

2. **Impact of Tokenization:** What Impact Do Different Tokenization Techniques Have on Model Performance?
   We investigate the impact of various tokenization techniques on the effectiveness of sentiment analysis models. Initially utilizing subword tokenization, which segments text based on the frequency of subword units, we question whether more contextually aware methods like BERT tokenization can enhance model performance by preserving more semantic information in the text.

3. **Aspect Analysis:** Are the Chosen Aspects Reflective of the True Intent of IMDB Reviews?
   Focusing on four predefined aspects (actors, plot, music, and director), this experiment evaluates whether these aspects adequately capture the essence of sentiments expressed in approximately 50,000 reviews. We assess if these chosen aspects are sufficient for a comprehensive aspect-based sentiment analysis or if additional aspects should be considered to improve the depth and accuracy of the analysis.

4.1 IMDB movie review dataset

The IMDB movie review dataset is an extensive collection specifically designed for binary sentiment classification, featuring 25,000 movie reviews for training and an equal number for testing. This dataset also includes additional unlabeled data, suitable for unsupervised learning or further evaluation. Each review is labeled with binary sentiment polarity (positive or negative), making it ideal for training models in sentiment analysis tasks. Maas et al. (2011)

4.2 Data split

For the purpose of this study, the IMDB movie review dataset is divided into two distinct sets: 70% of the data is used for training the models and remaining 30% serves as the testing set, used to evaluate the models' performance and their ability to generalize to new, unseen data.

4.3 Experimental setup

Each model is trained and tested under uniform conditions to ensure a fair comparison and reliable outcomes. The dataset is processed with a batch size of 32, and the same preprocessing functions are used to standardize and denoise the data for every experiment conducted.

4.4 Experiment 1: Model Comparison

This experiment evaluates whether Convolutional Neural Networks (CNNs) or Long Short-Term Memory networks (LSTMs) are more effective for sentiment analysis using the IMDB dataset. The models' effectiveness is assessed based on their precision in classifying sentiments.

1. **Model Setup:** Both the CNN and LSTM models are trained on same input sequences that are encoded using a subword tokenizer. This approach ensures that the models handle a variety of textual inputs efficiently.

2. **Hyperparameters:** The embedding dimension for both models is set to 256. Each model is trained over 5 epochs to balance between underfitting and overfitting.

3. **Testing and Evaluation:** The models are evaluated on their accuracy, precision, and recall. These metrics provide a comprehensive view of each model's performance and its capability to manage sentiment classification tasks.

4.5 Experiment 2: Impact of Tokenization

In this experiment, we explore the potential benefits of using BERT tokenization, a more contextually aware method that could enhance model performance by preserving richer semantic information in the text. We substitute traditional subword tokens with BERT tokens to generate BERT embeddings, which are then fed into an LSTM model.

1. **Model Setup:** The LSTM is configured to utilize BERT embeddings, integrating deep contextual insights into the sentiment analysis process.
2. **Testing and Evaluation:** The model's effectiveness is quantified through metrics such as accuracy, precision, and recall. These measures will help assess the improvement in performance attributable to the integration of BERT tokenization compared to standard methods.

4.6 Experiment 3: Aspect-Based Analysis

In this experiment, we evaluate the alignment between predefined aspects extracted via Latent Dirichlet Allocation (LDA) and the true sentiments expressed in the reviews. We aim to assess how effectively the LSTM and CNN model captures and classifies these sentiments.

1. **Model Setup:** Clauses and aspects extracted from the reviews using LDA are fed into the LSTM and CNN model, which is designed to predict the sentiment of each clause based on its contextual understanding.
2. **Testing and Evaluation:** The predicted sentiment for each clause is aggregated and compared to the overall sentiment of the sentence. This approach allows for a detailed evaluation of the model's ability to accurately reflect the sentiment of each aspect within the broader context of the review.

4.7 Results

This section detailed on the result obtained from our experiments. The result will be discussed in 5

4.8 Experiment 1 Result

In the Experiment 1 results, where we compared the performance of a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) network, both utilizing subword tokenization. The results indicate a closely matched performance between the two models, with the CNN slightly outperforming the LSTM.

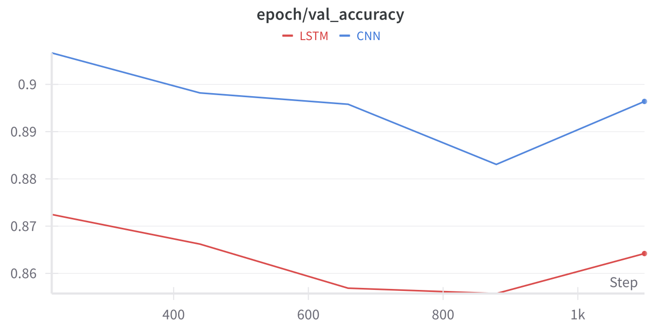

Figure 4: Accuracy comparison between CNN and LSTM

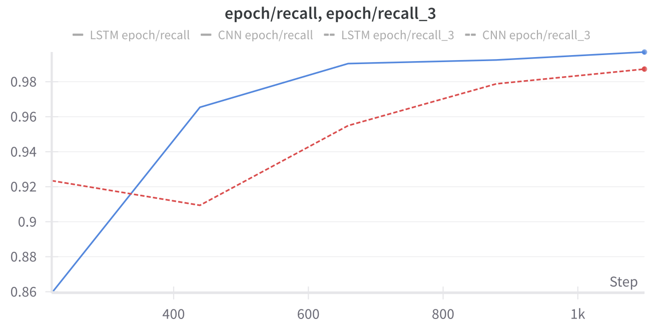

Figure 5: Recall comparison between CNN and LSTM

As shown in Figure 4, the accuracy rates achieved were 89% for the CNN and 87% for the LSTM. Figure 5 also demonstrates that both architectures are capable of effectively handling sentiment analysis on the IMDB dataset, albeit with a slight advantage in favor of the CNN.

4.9 Experiment 2 Result

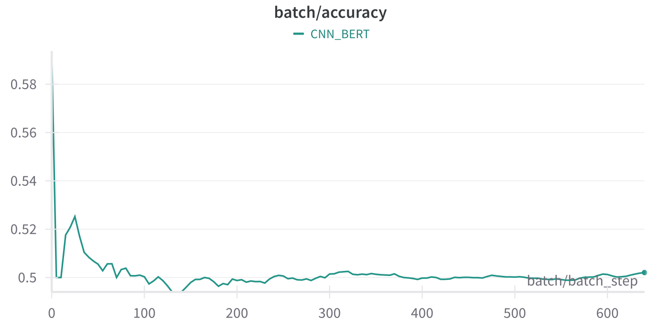

Figure 6: Performance of the Sentiment Analysis Model using BERT Tokenization

The performance results from utilizing BERT tokenization, as illustrated in Figure ??, indicate a substantial decrease in accuracy to 50%. This drop in performance suggests challenges in effectively integrating BERT for this specific sentiment analysis task.

To mitigate this issue, we decided to modify our BERT model by adding two more BiLSTM layers and concatenating the embeddings. This adjustment improved the model's accuracy, demonstrating the effectiveness of the enhanced architecture in addressing the challenges previously encountered. The graph in Figure 7 demonstrate the accuracy of bert after updating the model

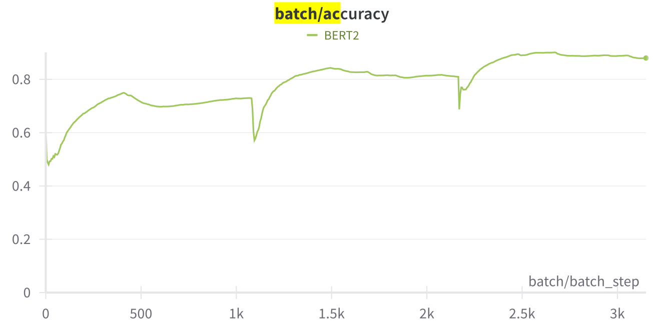

Figure 7: Performance of the Sentiment Analysis Model using BERT Tokenization

4.10 Experiment 3 Result

In Experiment 3, we focused on calculating the relevance of four predefined aspects (actor,plot,music,director) within the IMDB movie review dataset using two models, BILSTM and CNN. The aspects evaluated were actors, plot, music, and director. Below is a table comparing the accuracy of BILSTM and CNN for each aspect.

Aspect	LSTM (%)	CNN (%)
Actors	80	88
Plot	82	79
Music	75	72
Director	78	74

Table 1: Accuracy of BILSTM and CNN models for each aspect

Figure 8 below shows the occurrence of each aspect within the reviews, providing insight into which aspects are most frequently discussed.

The results indicate that BILSTM outperforms on CNN three out of one aspects, which may be attributed to its superior ability in handling sequence data and capturing the deeper contextual relationships necessary for accurate aspect-based sentiment analysis.

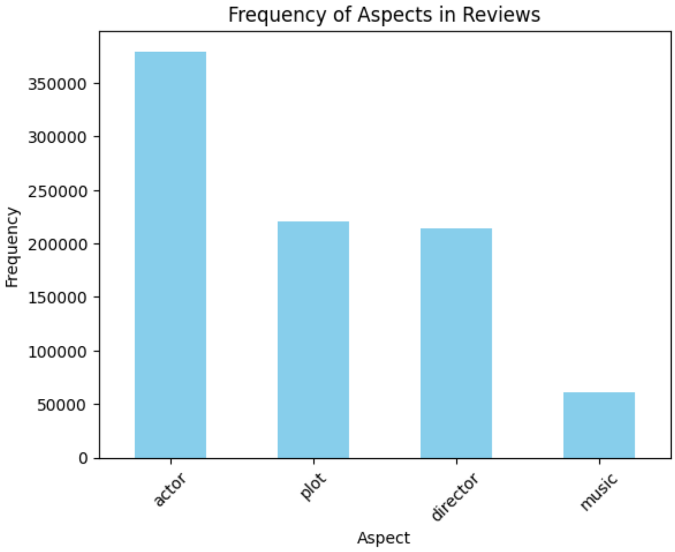

Figure 8: Occurrence of each aspect in IMDB movie reviews

5 Discussion

This study systematically evaluated the performance of different neural network architectures and tokenization methods in sentiment analysis using the IMDB movie review dataset. Experiment 1 revealed that CNN slightly outperforms LSTM in terms of accuracy, although both models showed closely matched results, suggesting that either model could be effective depending on the specific requirements of the sentiment analysis task. Experiment 2 demonstrated that while BERT tokenization is theoretically superior due to its deep contextual awareness, it did not enhance performance as expected and instead reduced accuracy significantly. This outcome might be attributed to the complexity of integrating BERT with existing models or perhaps the nuances of the dataset that do not align well with BERT's pre-trained model assumptions.

Experiment 3 focused on aspect-based sentiment analysis, comparing BILSTM and CNN accuracies across different movie review aspects. The results favored BILSTM, likely due to its capability to handle sequential data and maintain context over longer text spans, which is crucial for accurately linking sentiments to specific aspects.

These findings contribute to a nuanced understanding of how different AI technologies perform in real-world NLP applications and suggest that while advanced models like BERT offer promising capabilities, their integration into practical applications requires careful consideration and optimization.

6 Conclusion

The comparative analysis of CNN and LSTM models underscored the subtle yet important distinctions in how these architectures process and analyze textual data. Although CNN marginally outperformed LSTM in general sentiment analysis, the choice between these models should be guided by the specific characteristics of the text data and the computational resources available. On the other hand, the under performance of BERT in this setting highlights the challenges of deploying highly sophisticated NLP models in diverse settings. This serves as a cautionary note that more complex models, like BERT, are not always guaranteed to perform better than simpler models unless properly tuned and aligned with the task specifics.

References

Farhadloo, M., & Rolland, E. (2016). Fundamentals of sentiment analysis and its applications. *Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence*, 1–24.

Lohith, C., Chandramouli, H., Balasingam, U., & Arun Kumar, S. (2023). Aspect oriented sentiment analysis on customer reviews on restaurant using the lda and bert method. *SN Computer Science*, *4*(4), 399.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142–150). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/P11-1015`

Onalaja, S., Romero, E., & Yun, B. (2021). Aspect-based sentiment analysis of movie reviews. *SMU Data Science Review*, *5*(3), 10.

Pannala, N. U., Nawarathna, C. P., Jayakody, J., Rupasinghe, L., & Krishnadeva, K. (2016). Supervised learning based approach to aspect based sentiment analysis. In *2016 ieee international conference on computer and information technology (cit)* (pp. 662–666).

Phan, M. H., & Ogunbona, P. O. (2020). Modelling context and syntactical features for aspect-based sentiment analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3211–3220).

Rybakov, V., & Malafeev, A. (2018). Aspect-based sentiment analysis of russian hotel reviews. In *Ceur workshop proceedings* (pp. 75–84).

Thet, T. T., Na, J.-C., & Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, *36*(6), 823–848.

Wang, J., Xu, B., & Zu, Y. (2021). Deep learning for aspect-based sentiment analysis. In *2021 international conference on machine learning and intelligent systems engineering (mlise)* (pp. 267–271).