Comparative Analysis of Traditional Large Models and Large Language Models (LLMs) for Sentiment Analysis: A Case Study on Sentence Length and Processing Efficiency

Shaurya Vardhan
Independent Research

Abstract— This paper explores the comparative effectiveness of traditional large models, such as BERT, and contemporary Large Language Models (LLMs) like GPT -3.5 and Llama2 in performing sentiment analysis on a Google Reviews dataset. Our study finds that for sentences with up to 40 keywords, traditional models perform on par with LLMs. However, for sentences exceeding 40 keywords, LLMs are essential for maintaining accuracy and efficiency. This finding highlights the potential for optimizing sentiment analysis tasks by strategically selecting models based on sentence length, thereby ac hieving a balance between cost -effectiveness and processing speed.

Index Terms—BERT, Large Language Models, Sentiment Analysis, Comparative Study

I. INTRODUCTION

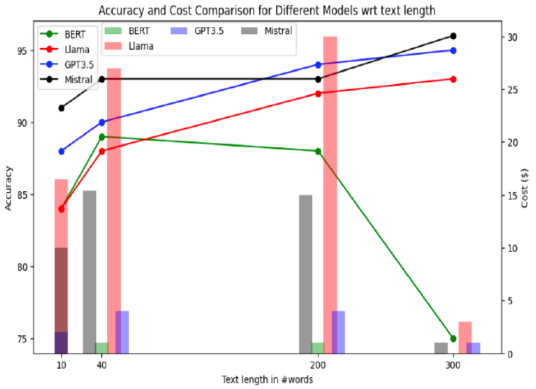
Sentiment analysis has become an integral part of understanding customer feedback, social media sentiment, and user -generated content. With the evolution of natural language processing (NLP), various models have been developed to enhance the accuracy and efficiency of sentiment analysis. Traditionally, models like BERT (Bidirectional Encoder Representations from Transformers) have been widely used due to their robust performance in various NLP tasks. However, the emergence of Large Language Models (LLMs) such as GPT -3.5 and Llama2 has prompted a reevaluation of the tools used for sentiment analysis, particularly concerning their efficiency in handling varying sentence lengths.

This research investigates whether the newer LLMs offer significant advantages over traditional models like BERT in processing sentences of different lengths. By examining the performance on a dataset of Google Reviews, we aim to provide insights into the optimal use cases for each model type, thereby guiding more cost -effective and efficient sentiment analysis practices.

II. METHODOLOGY

Dataset: The study utilized a dataset consisting of Google Reviews, which were pre -labeled based on their star ratings. Sentenc es were categorized into two groups based on length: those with up to 40 keywords and those with more than 40 keywords. Models Evaluated: BERT: A traditional large model, known for its bidirectional training and effectiveness in various NLP tasks. GPT -3.5 Turbo, Llama2 (including quantized variants). Evaluation Metrics: Accuracy/F1 Score: To assess the correctness of sentiment classification. Processing Time: Time taken to analyze sentences. Cost Efficiency: Measured in terms of computational resources requ ired. Scalability: Ability to handle large datasets effectively. Experiment Design: Sentences from the dataset were divided into two groups based on their length. Each group was processed using both BERT and the selected LLMs. The performance was measured across the defined metrics, with a specific focus on comparing

III. RESULTS



Group 1: Sentences with Up to 40 Keywords For sentences with up to 40 keywords, BERT performed remarkably well, matching the accuracy and F1 scores of LL Ms like GPT -3.5 Turbo and Llama2. The processing time and computational costs were also comparable, suggesting that for shorter sentences, traditional models are just as effective as modern LLMs. This indicates that using BERT for such tasks can result in significant cost savings without sacrificing performance. Group 2: Sentences with More Than 40 Keywords In contrast, for sentences exceeding 40 keywords, LLMs demonstrated a clear advantage. While BERT struggled with longer sentences, often misclassifying sentiments or taking significantly longer to process, LLMs like GPT -3.5 Turbo and Llama2 maintained high accuracy and efficiency. The processing time for these models was notably faster, and despite the higher computational requirements, the overall cost -effectiveness was better due to the reduced need for reprocessing or manual intervention.

IV. DISCUSSION

The results of this study provide a nuanced understanding of when to deploy traditional models like BERT versus when to utilize LLMs for sentiment analy sis. While LLMs have been celebrated for their superior performance, our findings indicate that for sentences with fewer than 40 keywords, BERT remains a viable, cost -effective option. This is particularly relevant for organizations looking to optimize the ir computational resources without compromising on accuracy. However, for more complex tasks involving longer sentences, LLMs are indispensable. Their ability to handle extensive text inputs with higher accuracy and speed justifies their use, despite the higher initial computational costs. This strategic deployment of models based on sentence length can lead to more efficient sentiment analysis pipelines, saving both time and resources.

V. CONCLUSION

This study underscores the importance of selecting the right tool for the right task in sentiment analysis. While traditional large models like BERT are sufficient for shorter sentences, the use of LLMs becomes crucial for longer sentences, where their advantages in accuracy and processing speed become evident. By adopting this selective approach, organizations can achieve significant cost savings and efficiency gains. Future research should explore the impact of other factors such as sentence complexity, domain -specific language, and real -time processing needs on t he performance of traditional models versus LLMs. Additionally, further studies could investigate hybrid approaches that combine the strengths of both model types for more comprehensive sentiment analysis solutions.

REFERENCES

[1] Devlin, J., et al. (2019). BE RT: Pre -training of Deep Bidirectional Transformers for Language

[2] Understanding. NAACL.

[3] Brown, T., et al. (2020). Language Models are Few -Shot Learners. NeurIPS.

[4] Meta AI. (2024). Llama2 Model Card.

[5] OpenAI. (2024). GPT -3.5 Turbo. OpenAI Pricing.

[6] Hugging Face. (2024). Performance and Scalability in LLMs.