INTRODUCTION

Sentiment analysis has emerged to be a crucial resource in Natural Language Processing (NLP), allowing companies and scholars to obtain valuable information through user-generated content, including reviews, social media posts, and customer reviews. Although classic machine learning classifiers such as Naive Bayes and Random Forest are able to predict sentiment well in binary cases, they do not work well in multi-class sentiment classification because they fail to register subtle language trends, contextual meaning, and multi-polar feeling, such as sarcasm or negation. Following improvements on deep learning, especially with transformer-based pre-trained model like BERT and RoBERTa, the models have been seen to perform at great capacity in comprehending complex linguistic patterns. Nevertheless, they can be quite complex to compute and less human friendly when it comes to application in the real world.

This paper will make a comparative review between classical machine learning, deep learning and transformer-based models to classify multi-sentence sentiment with respect to accuracy, computation and interpretability. The study will use the publicly available (under an open license) of a sentiment analysis dataset published by Kaggle, which is reproducible and does not violate ethical standards. This study aims at testing the best model to apply in multi-class sentiment analysis by benchmarking such models as Naive Bayes, Random Forest, LSTM, BiLSTM, BERT, and RoBERTa using a standardized dataset and considering main trade-offs between performance and complexity of the models. The results will not only have some contribution in the academic studies, but also in real-life practices that may present a useful piece of information to identify the best sentiment analysis methods in real life circumstances.

LITERATURE REVIEW

Sentiment analysis as branch and subfield of the Natural Language Processing (NLP) has also advanced greatly with the development of deep learning and machine learning. Initial strategies were based on word-based technique and conventional ML algorithm like Naive Bayes (NB) and Support Vector Machines (SVM), making use of TF-IDF and bag-of-words (BoW) representations (Pang et al., 2002). These models performed fairly well in binary sentiment classification yet failed when applied to multi-class tasks because they failed to express contextual relations and textual sequential dependencies (Liu, 2012).

The development of the models depending on neural networks, specifically, Recursive Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM), significantly widened the field of sentiment analysis, as they secretly read the text as a sequence and capture a long-term relationship (Hochreiter & Schmidhuber, 1997). They also came up with most of the pre-trained word embeddings (including GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017)) which deliver much richer semantics in the vectors. Nevertheless, those models still had some difficulties with processing complex linguistic trends, such as sarcasm, negation, and expressions of sentiment that are specific to a given domain (Zhang et al., 2018).

The paradigm shift took place when the models based on the transformer were introduced specifically BERT (Bidirectional Encoder Representations from Transformers) [1] and the RoBERTa (Robustly Optimized BERT Approach) [2] based on self-attention mechanisms to capture the relationship between words at a much deeper contextual level. These models, trained on huge corpora, depict state of the arts performance on several NLP tasks, including sentiment analysis. Experiments by Joshy & Sundar (2022) and Ojo et al. (2023) verified that transformer-based models are much more reliable in sentiment classification efforts (regarding multi-classification) than conventional ML-based methods and deep learning models, especially in dealing with complex sentiments. Nonetheless, their high cost in terms of computations and their lower interpretability serve as the major downfalls. [3]

The most recent works have considered hybrid methods using transformer models with classical ML methods to achieve increased efficiency without AI losing its performance quality. Moreover, attempts to ensure model interpretability have been launched by applying such methods as attention visualization and feature importance analysis. Although the advancements and progress have been made, the study that compares all three of these approaches deep learning, classical ML approaches, and transformers models on the same multi-class dataset is yet to be done. The present paper will attempt to fill this gap by benchmarking these methods systematically, and give an idea about their advantage and shortcomings in real-life scenarios of conducting sentiment analysis.

METHODOLOGY

*Reference:*

1. *Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).*
2. *Liu, Y., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.*
3. *Neelakanti, S. (2025). Comparative Study of Transformer-Based Models in Emotion Recognition. NORMA.*
4. *Joshy, A., & Sundar, C. (2022). Analyzing the Performance of Sentiment Analysis Using BERT, DistilBERT, and RoBERTa. Research Gate/ 2022 IEEE IPRECON (International Power and Renewable Energy Conference)*
5. *Ojo, E., et al. (2023). A Comparative Study of Transformer Models for Multi-Class Sentiment Analysis.*