Advancements in Sentiment Analysis: A Methodological Examination of News using multiple LLMs

Muhammad Ali Mahmood[1], Iffat Maab[2], Muhammad Sibtain[1], Asima Sarwar[1],
Muhammad Arsalan[3], Masroor Hussain[1]
FCSE, GIK Institute of Engineering Sciences and Technology, Topi[1]
National Institute of Informatics, Tokyo[2] Technische Universität Braunschweig, Germany[3]
{u2021346, u2021459, asima.sarwar, hussain}@giki.edu.pk[1]
{maab}@nii.ac.jp[2], {muhammad.arsalan}@tu-braunschweig.de[3]

Abstract

As the digital news consumption continues to grow, sentiment analysis has become essential for understanding public opinion and the impact of media. Traditional NLP methods often fail to capture the in-depth emotions in news content, especially when taken from various media sources. This work identifies gaps in the use of fine-tuned deep learning models and large language models (LLMs) without fine-tuning in the sentiment analysis of news articles, offering enhanced insights across various model families. In our work, we collect news from BBC, and annotate BBC dataset using the proprietary OpenAI GPT-3.5-turbo model, and fine-tune models such as DistilBERT, BERT, and RoBERTa-large. We also compare fine-tuned models with LLM variants including Llama-3 and Qwen-2 models without any model fine-tuning through crafted prompts. Our results show that RoBERTa-large achieved the highest performance, delivering an accuracy of 86%.

1 Introduction

Given the rise of misinformation, polarized reporting, and the complexities of contemporary global crises, it becomes crucial to develop methods that can accurately interpret the language used in today's digital media. Political orientations, economic movements, and societal attitudes are all shaped by the media, which often carry subtle emotional undertones, whether intentional or not [1]. Tools that can unveil these hidden sentiments enhance transparency and promote media literacy [2].

Sentiment analysis using NLP has demonstrated significant importance across diverse fields such as market analysis, political opinion tracking, and social media analytics [3, 4]. With the growth in online news and its various formats, understanding how news articles express emotional tone and sentiment becomes increasingly important for stakeholders such as businesses, policymakers, and the general public [4]. Moreover, incorporating large language models (LLMs) into this domain presents promising avenues to improve the accuracy and interpretability of sentiment classification [5].

This paper investigates sentiment analysis on news articles collected from high-quality source i.e., BBC using NewsAPI. Traditional NLP models often struggle to capture the nuanced emotions present in complex texts, especially in news reporting, where the language used is often subtle and context-dependent [6]. While various sentiment analysis tools exist, they frequently underperform in the context of news articles due to limitations in their training data and lack of domain-specific knowledge. In our study, we introduce a framework to enhance the accuracy of sentiment classification while also offer deeper insights into the narrative elements that shape public perception.

In our work, we utilize models such as BERT and RoBERTa [7, 8], which have been fine-tuned to determine sentiment classification across news content. We also utilize instruction-tuned open-source LLMs such as Llama-3 (1B, 3B, 8B) and Qwen-2 (1.5B, 3B, 7B) without any fine-tuning to examine how they process sentiments compared to fine-tuned approaches. In addition, we identify the dominant words frequently used in news sources that influence sentiments, drawing on methodologies such as those described attention-based models by [9]. The significance of this work lies in its response to the growing demand for

advanced analysis of contemporary and real-time media landscapes.

2 Related Work

Sentiment analysis, particularly focusing on news content, has significantly advanced due to innovations in machine learning (ML), deep learning (DL), and generative AI (GenAI) approaches. Each approach has contributed uniquely to the evolution of sentiment analysis techniques.

Popular **classical techniques** such as Naive Bayes, SVM, and the hybrid approaches have been adopted earlier for analyzing the sentiment from news articles. [10] proposed Naive Bayes approach tailored for social and news sentiment analysis to achieve higher performance on small datasets. Similarly, [11] combined Naive Bayes and SVM for text classification, highlighting its effectiveness in identifying sentiment trends in news media. [12] took sentiment analysis a step further by incorporating effective computing techniques with SVM to facilitate more nuanced sentiment categorization across a variety of news sources. Building on these approaches, [13] utilize a hybrid SVM model to achieve greater accuracy, showing significant improvements over traditional machine learning methods in sentiment classification.

Table 1: Summary of dataset distribution for three-class sentiment analysis.

Labels	Number of Sentences	Percentage
Positive	3,500	35%
Negative	3,500	35%
Neutral	3,000	30%
Total	10,000	100%

Deep learning models like LSTM and BERT have proven more effective in sentiment analysis, particularly excelling with long-form news articles. In this context, [14] showcased the strength of LSTM networks in identifying temporal patterns, making them particularly suitable for analyzing sequenced news data. Another work by Yang et. al. in [15] highlighted the exceptional capabilities of BERT, a pre-trained transformer model, in delivering nuanced sentiment classification, achieving state-of-the-art performance in news sentiment tasks. Similarly, research by [16] highlights the superior capability of LSTMs over CNNs in capturing contextual sentiment in text analysis. [17] find strong efficacy of RoBERTa model for sentiment prediction, especially in the domain of politically sensitive news content. [18] use DistilBERT for real-time sentiment classification, achieving notable accuracy while significantly reducing computational costs compared to larger models.

Recent advancements in **generative AI models**, including GPT-3, have led to the development of robust frameworks for sentiment classification through the training of large-scale language models. As shown by [19], detecting social and news sentiments using GPT-3 also has transfer learning abilities for adapting to diverse datasets. [20] investigated the integration of GPT-2 and RoBERTa for multi-source sentiment analysis, tackling issues of cross-domain adaptability and providing insights into effective sentiment classification across varied news environments.

While sentiment analysis of news has advanced significantly, particularly with recent progress in machine learning, deep learning, and generative AI models, many existing approaches still struggle to accurately analyze news content. Traditional machine learning techniques, such as Naïve Bayes and SVM, failed to capture the complexity of emotions in news articles because they lack contextual understanding to interpret sentiments. Although deep learning models are more adept at handling sequential data, they remain computationally expensive to train and may not address the contextual issues unless specifically fine-tuned on news datasets. While models like GPT-2 and GPT-3 have shown promise for sentiment analysis, they are computationally costly and require domain-specific fine-tuning, presenting challenges in practical applications.

3 Proposed Approach

Our methodology consists of a multiple-stage process for analyzing the sentiment of news articles. First, we utilize NewsAPI[1)] to collect data from BBC News corpus, selecting specific sources and keywords to filter news articles. We filtered the news articles using specific keywords such as politics, technology, science, sports, etc. to remain relevant to the targeted sentiment themes.

Second, we preprocess the data through text cleaning and tokenization. The dataset was initially collected in the JSON format, after which it was cleansed and processed using the Beautiful Soup library. This initial step involved removing URLs, special characters, and stop words, effectively eliminating extraneous information and allowing

1) https://newsapi.org/

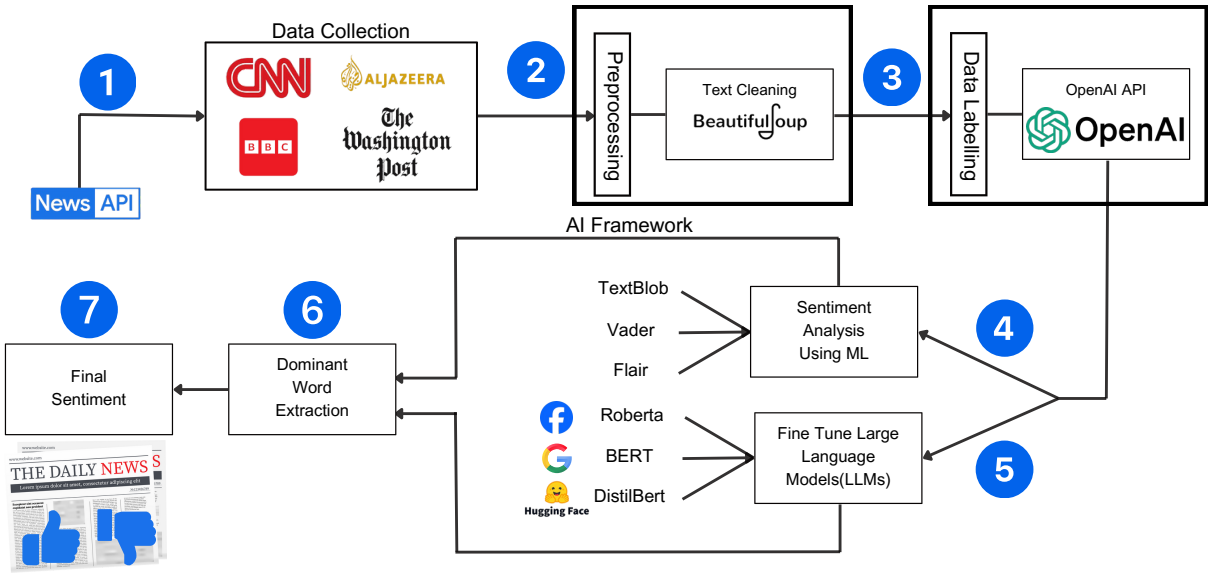

Figure 1: Workflow pipeline for enhanced sentiment analysis of news articles, showing stages from data collection to final sentiment scoring.

us to concentrate on the relevant text. We then annotate the collected samples, ensuring an equal distribution of positive, negative, and neutral sentiments, using widely recongnized GPT-3.5-turbo model to guide the labeling process.

In the final stages, we experiment with various traditional sentiment analysis models, including TextBlob, Vader, and Flair. Flair emerges as the top performer. Flair's performance served as as a baseline for experimenting with more advanced models, i.e., we also fine-tune deep learning models such as DistilBERT, BERT, and RoBERTa-large on the annotated dataset. RoBERTa-large demonstrates superior performance. Furthermore, zero-shot experiments using LLMs are performed with variants of Llama-3 and Qwen-2 model families. Table 3 shows the template used for zero-shot experiments. We also configure RoBERTa model to extract dominant words associated with each sentiment label, adding an interpretative layer to the results by highlighting influential words that drive the sentiment classification.

The integration of our fine-tuned models in sentiment analysis pipeline show improved performance in discerning context-aware sentiment scores, which enhanced the predictions compared to those from traditional models and LLMs used without any fine-tuning. The workflow is illustrated in Figure 1, detailing each phase from data collection through to the final sentiment classification step. Table 1 shows dataset distribution. See Appendix A.1 for details on dataset, A.2 for experimental setup and implementation details, and A.3 for evaluation metrics.

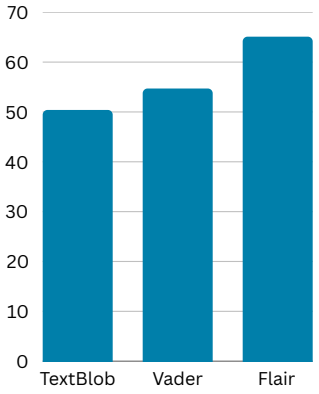

Figure 2: Comparison of ML models in terms of accuracy.

4 Results and Experiments

In our first set of experiments, we use ML models for sentiment classification on our dataset, using TextBlob, Vader, and Flair. The accuracy of each model was evaluated, with the results summarized in Figure 2. TextBlob achieved an accuracy of 50.36%, while VADER performed slightly better with an accuracy of 54.66%. The best performance among the ML models was from Flair, with an accuracy of 65.09%.

We also compare fine-tuned DL models against LLM

Table 2: Sample showing Sentiment Dominant (SD) scores across each sentiment type.

News Sample	Sentiment Type	SD Scores
Federal Reserve hints at potential rate hike amid inflation concerns.	NEGATIVE	hints: 5.99, hike: 4.13, inflation: 3.41, rate: 3.25, amid: 2.56
Tech giant announces revolutionary smartphone with groundbreaking features.	POSITIVE	groundbreaking: 2.85, revolutionary: 1.54, smartphone: 1.34, announces: 1.31, features: 0.98
Global survey reveals shifting trends in consumer behavior.	NEUTRAL	shifting: 2.27, trends: 1.28, reveals: 0.99, survey: 0.94, global: 0.85

Prompt	Labels
You are a sentiment analysis detector. User: Classify the following "$sentence$" as x_a, x_b, or x_c without providing additional details System: The sentence is:	x_a: Positive x_b: Negative x_c: Neutral

Table 3: Zero-shot prompt for the Llama-3 and Qwen-2 models involves using the labels x_a, x_b, and x_c, which represent different classes within the dataset.

Model	Acc.	P	R	F1-score
Fine-tuned				
DistilBERT	80.40	76.35	76.25	76.29
BERT	72.22	69.94	69.81	69.85
RoBERTa-large	**86.12**	**81.29**	**81.36**	**81.31**
No fine-tuning				
Llama-3.2-1B-Instruct	39.03	39.03	39.03	39.03
Llama-3.2-3B-Instruct	51.97	54.51	51.97	51.97
Llama-3-8B-Instruct	56.61	58.14	56.61	56.61
Qwen-2-1.5B-Instruct	42.63	50.72	42.63	42.63
Qwen-2.5-3B-Instruct	56.83	57.78	56.83	56.09
Qwen-2-7B-Instruct	60.10	62.72	60.10	60.10

Table 4: Summary of results with the fine-tuned models against LLMs with no fine-tuning in terms of accuracy (Acc.), P, R, and F1-scores for three-class sentiments.

variants of Llama-3 and Qwen-2, noting that BERT, DistilBERT, and RoBERTa-large demonstrated robust performance over ML models and LLMs. Table 4 shows a summary of our results. Among fine-tuned models, it can be seen that RoBERTa-large outperformed with 86.12%, followed by DistilBERT with an accuracy of 80.40%, and BERT with 72.22%. In case of LLMs used without any fine-tuning, the Qwen-2-7B model demonstrated increased performance compared to a similarly sized Llama-3-8B model, achieving an F1-score of 60.91 versus 56.61, respectively. LLMs with smaller parameter counts, such as Llama-3 (1B, 3B) and Qwen-2 (1.5B, 3B), do not exhibit significant performance improvements. These findings indicate that fine-tuned transformer-based models outperform both traditional ML models and LLMs in handling sentiment classification tasks. We found that while, Vader and TextBlob provided some level of insight, their performance was limited, especially on sentences with subtle or compound sentiments, which RoBERTa handled more effectively.

In addition, we also integrate dominant word extraction, which identifies key terms that influence sentiment categorization. This approach adds interpretability to the model, an aspect often overlooked in previous studies. While most existing methods focus primarily on classification accuracy, our technique offers a deeper understanding of how specific words impact sentiment labels. For this purpose, we use our best performing model, RoBERTa-large. Refer to the results in Table 2, which illustrates an example showing the dominant words and their respective SD scores.

5 Conclusion

The objective of our study is to better understand the sentiment analysis research by synthesizing results of advanced proprietary LLMs, such as GPT-3.5-turbo, alongside various fine-tuned ML and DL models, against family of LLMs used without any gradient updates. As part of this study, we highlight open challenges associated with applying these models to sentiment analysis. Although the Llama-3 and Qwen-2 models yield less substantial results, they hold significance in resource-constrained environments where annotation can be costly. Overall, our findings indicate that DL models, when fine-tuned, exhibit substantial capabilities in accurately detecting sentiments, outperforming traditional ML and generative LLMs. Notably, the RoBERTa model has proven effective in sentiment classification within digital media contexts, although the limited availability of labeled data may impede the model's generalizability. Future work may involve sentiment interpretation by integrating hybrid DL approaches with LLMs, such as combining RoBERTa with XLNet or T5, to enhance the contextual analysis of sentiments.

Acknowledgements

The authors wish to express gratitude to the funding organisation as this study was carried out using the TSUBAME4.0 supercomputer at Institute of Science Tokyo.

References

[1] Alexander Ligthart, Cagatay Catal, and Bedir Tekinerdogan. Systematic reviews in sentiment analysis: a tertiary study. **Artificial intelligence review**, pp. 1–57, 2021.

[2] Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. A survey of sentiment analysis: Approaches, datasets, and future research. **Applied Sciences**, Vol. 13, No. 7, p. 4550, 2023.

[3] John Smith and Priya Patel. Exploring sentiment analysis techniques in natural language processing. **International Journal of NLP Research**, Vol. 12, No. 4, pp. 345–367, 2020.

[4] Andrew Lee and Mei Lin Wong. A guide to sentiment analysis using nlp. **Journal of Data Science and Applications**, Vol. 10, No. 2, pp. 221–240, 2019.

[5] Ethan Brown and Olivia Taylor. Advancements in nlp: Deep learning models for sentiment analysis. **Deep Learning and AI Applications**, Vol. 15, No. 6, pp. 501–518, 2021.

[6] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval**, Vol. 2, No. 1-2, pp. 1–135, 2008.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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: Human Language Technologies**, pp. 4171–4186, 2019.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[9] Zhenyu Li, Hua Zhang, and Yunfei Zhang. Attention-based models for sentiment analysis. **Computational Linguistics**, Vol. 47, No. 1, pp. 123–145, 2021.

[10] H. Park, et al. Application of naive bayes in news sentiment analysis. **IEEE Transactions on Computational Social Systems**, 2020.

[11] B. Pang, et al. Sentiment analysis with naive bayes and svm. **Journal of Machine Learning Research**, 2019.

[12] E. Cambria, et al. Affective computing and sentiment analysis using ml. **IEEE Transactions on Affective Computing**, 2021.

[13] X. Liu, et al. Hybrid svm for news sentiment analysis. **International Journal of Data Mining**, 2021.

[14] L. Zhang, et al. Sentiment analysis using lstm. **Journal of Data Science**, 2020.

[15] Q. Yang, et al. Bert for news sentiment analysis. **IEEE Transactions on Neural Networks and Learning Systems**, 2021.

[16] S. Tang, et al. Comparative study of cnn and lstm for sentiment analysis. **Social Media Analysis Journal**, 2022.

[17] J. Sun, et al. Fine-tuned roberta for news sentiment. **Journal of Computational Linguistics**, 2022.

[18] M. Liu, et al. Real-time sentiment analysis with distilbert. **Journal of News Sentiment Analysis**, 2023.

[19] T. Ahmed, et al. Gpt-3 for sentiment analysis of social media. **Social Media Research Journal**, 2023.

[20] P. Brown, et al. Gpt-2 and roberta for advanced sentiment analysis. **Text Analytics Journal**, 2023.

[21] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[22] AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. **Meta AI**, 2024.

[23] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. **arXiv preprint arXiv:2407.10671**, 2024.

[24] Hugging Face. The ai community building the future. **URL: https://huggingface. co**, 2021.

A Appendix

A.1 Dataset

In this study, we used a dataset custom-labeled from a collection of news articles obtained through NewsAPI. The initial dataset comprised roughly 18,000 unlabeled samples from BBC News, chosen for its dependable and varied coverage of multiple topics ranging from politics to science, sports, and more. We utilized the GPT-3.5-turbo model to automatically assign sentiment categories: positive, negative, or neutral—to each sample in the dataset. Next, we downsample the data to achieve a balanced distribution of sentiments, resulting in a final dataset of 10,000 samples, with approximately 3,500 instances of each sentiment category. Having a balanced dataset allowed us to train and evaluate our sentiment analysis models with increased precision.

A.2 Experimental settings

We compare various fine-tuned models such as DistilBERT, BERT, and RoBERTa-large for the task of sentiment classification. In our experiments, a data split of 80/10/10 for training, validation, and testing, respectively, is used, ensuring that the samples do not overlap between sets. Initially, we test with various batch size, learning rate, and the number of training epochs to optimize the training efficiency of our models. Weights decay and frequency of logging was changed to prevent overfitting as well as to monitor the training progress. Moreover, for comparison of our models with classical methods, we utilize models including TextBlob, Vader, and Flair to evaluate the improvements offered by the fine-tuned models. The hyperparameters used in the fine-tuning of our models are summarized in Table 5. LLM variants include Llama-3 [21, 22] and Qwen-2 [23]. Moreover, we use PyTorch for model implementation, leveraging resources from HuggingFace [24] for DistilBERT, BERT, RoBERTa-large, Llama-3, and Qwen-2, and the NLP-Toolkit for TextBlob, Vader, and Flair.

A.3 Evaluation Metrics

The metrics included accuracy (*Acc.*), precision (*P*), recall (*R*), F1-score (*F1*), and Sentiment Dominance (*SD*), each tailored to the challenges inherent in analyzing sentiment in news content.

SD is a unique metric for the model ability to find words which are dominant, meaning words that carry majority weight to the sentiment category it belongs to is called Sentiment Dominance. In order to use a model to associate key terms to the correct sentiment during fine tuning, we extract influential terms specific to each sentiment as our metric to evaluate model capability.

Table 5: Hyper-parameter settings used for fine-tuning models.

Hyper-parameter Settings for LLMs	
Epochs	5
Learning rate	2×10^{-5}
Batch size (train)	16
Batch size (eval)	16
Weight decay	0.01
Logging steps	10
Model saving strategy	Epoch
Total checkpoints saved	2
Load best model at end	True