Evaluating Text Classification Models for Sentiment Analysis on Tweets

Abstract—**The advent of social media websites led to a tremendous surge in the need for accurate and efficient sentiment analysis techniques, particularly for short and casual messages like tweets. Deep learning algorithms—specifically transformer-based algorithms like BERT (Bidirectional Encoder Representations from Transformers)—have proved to show outstanding performance in a vast majority of natural language processing applications. However, due to their excessive computational cost and vulnerability to sparse data, they tend to fail in low-resource environments. In this research work, the performance of six different classification models—Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Transformer, Temporal Convolutional Network (TCN), and Extreme Gradient Boosting (XGBoost)—is compared under an extensive range of experimental setups. These include combinations of various preprocessing types (none, nouns only), and tokenizers (TF-IDF embeddings and BERT embeddings). Out of all these experiments, GRU with TF-IDF resulted in a high accuracy of 0.792, an AUC-ROC of 0.889, and an F1 Score of 0.789.**

Keywords— BERT, GRU, Sentiment Analysis, TF-IDF, XGBoost

I. INTRODUCTION

The development of social media platforms has changed how people convey opinions and sentiments in real time. Among these platforms, Twitter (now known as X) is particularly notable due to its concise format and rapid information flow. Though tweets are straightforward, they typically have undertones of opinions and identifying the sentiments is a more challenging task. Their informality, and contextual dependence make more sophisticated tools necessary for proper interpretation. Companies use this ability to monitor sentiment towards their brands and respond to customer views in real time. Government institutions use it to analyze public sentiments in response to policies or issues such as natural disasters and health outbreaks. It also has use in finance to predict market trends. More recent developments in Natural Language Processing (NLP) with deep learning have accelerated sentiment analysis. LSTM, GRU, and CNN were some of the models that performed well on text sequences, but transformer models such as BERT provided better results. These transformer models use attention mechanism and large-scale pretraining to better capture meaning and tone, thereby improving the text classification, especially on tweets. These approaches, however, have trade-offs. They need big, labeled data sets and heavy compute resources, and thus are expensive and unrealistic for small teams or real-time use.

Conventional machine learning techniques, such as Logistic Regression, SVM (Support Vector Machines), XGBoost with TF-IDF (Term Frequency-Inverse Document Frequency) features, provide a simpler and lower-resource alternative. If supplemented with light weight neural layers like GRU, they can provide good performance without the overhead and thus are suitable for low-resource projects or projects with strict latency constraints.

Most prior studies limit their evaluation to a small number of models, thus limiting the scope of the results. In this view, this paper makes an attempt to identify the most effective methods to use for Twitter sentiment analysis with real, balanced tweet data. This research conducts a comprehensive comparative evaluation with six architectures, namely, GRU, LSTM, CNN, Transformer, TCN, and XGBoost. Furthermore, model improvement techniques have also not been comprehensively studied in the past. This research particularly investigates the impact of the additional variables on the performance of the model, thus making good contributions to understanding their true impact on the accuracy of sentiment classification. The results of this work can guide researchers and practitioners to build systems that are not only accurate but also efficient, fast, and deployable in real-world social media applications.

II. LITERATURE REVIEW

Sentiment analysis has been significantly revolutionized by technological advancements in machine learning and deep learning particularly when there are social media sites like twitter. Leading this brand-new wave of revolution is the Transformer model, which introduced a paradigm shift in sequence transduction models [1]. In this paper the authors demonstrated that attention-only models' mechanisms can do better than models constructed with advanced recurrent neural networks.

XGBoost has been found to produce outcomes on a wide set of machine learning problems, and it is an extremely important method applied in most predictive modeling tasks which can ease processes related to sentiment analysis [2]. In addition, the evolution of NLP has been informed by distributed word and phrase representations, also known as word embeddings [3]. These embeddings are an important mechanism of capturing nuanced relationships among words and thus ensuring greater context-related information, which is required for analytical models to function. These basic principles are the foundations where deep learning systems are developed, which have revolutionized the area of sentiment analysis and NLP. One such addition introduced to this area is

BERT [4]. By conditioning on the context around in all its depths, pre-training is performed to generate rich bidirectional representations of unannotated text. Then, with little task-specific moderation, pre-trained BERT can be fine-tuned to attain state-of-the-art performance. RNN Encoder-Decoder architecture has shown to have benefits in learning semantic representations of phrases, which is one of the primary approaches for a few future areas like statistical machine translation and sentiment analysis sequential processing data [5]. Moreover, researchers investigated convolutional and recurrent networks on general sequence modeling problems, providing significant architectural alternatives to attention-based-only approaches to textual sequence processing [6]. By optimizing pretraining, RoBERTa (A Robustly Optimized BERT Pretraining Approach), which is constructed upon BERT architecture, has produced transformer models that are much more effective on different NLP benchmarks [7].

The real-world applications of sentiment analysis are very diverse and are spread across different fields of study such as public health, political science, and sports. For example, in public health, a study used Twitter to measure public opinion of the Monkeypox outbreak and gave us the facts which established fear of panicking, disseminating false information, and stigmatization [8]. This shows the importance of sentiment analysis in measuring public perception during times of crisis in the health sector. Sentiment analysis has provided real-time intelligence, as seen in studies that analyzed political tweets during the 2019 Spanish Elections to gauge people's opinion of political events and political leaders [9]. Likewise, in the ever-changing sports landscape, sentiment analysis provides tremendous insights; football-oriented tweets research attempts to identify rapid shifts in fan sentiment during a live match on the basis of events like goals or penalties, thus creating an in-depth understanding of fan engagement and emotional investment over the course of a match [10].

While sentiment analysis has traditionally focused on less complex dual or triple-class labeling (e.g., positive/negative or positive/negative/neutral), the higher complexity of online argument has encouraged the move towards more complex multi-class or penta-class labeling [11]. The set of analytical tools used in tweet sentiment analysis is constantly being refined, consistent with the need for higher granularity and accuracy. In the context of the COVID-19 pandemic, researchers have used sentiment analysis accompanied by deep learning techniques to detect cultural differences in polarity and emotional responses in tweets and thereby reveal differentiated societal responses to global crises and government interventions [12].

A notable advancement in tweet sentiment analysis involves an advanced Aquila Optimizer combined with an ensemble Bi-LSTM-GRU and fuzzy emotion extractor [13]. This innovation combines deep learning recurrent models and fuzzy decision support systems and nature-inspired optimization techniques to attain better detection of sentiment. In some cases, the employment of traditional classification techniques short of capturing the intended user sentiment, and researchers have turned to multi-class sentiment analysis to detect the very sentiment conveyed by the user, rather than evaluating an overall polarity [14].

Apart from the discussed uses and base models, a number of techniques enhance sentiment analysis. Ordinal regression, for example, has been used in Twitter sentiment analysis to determine the inherent ordered nature of the classes of sentiment (e.g., very negative, negative, neutral, positive, very positive) rather than classifying them as discrete classes [15]. Comparative studies are necessary to understand the model landscape, with studies between transformer-based models and conventional models for sentiment analysis in social media corpora, giving insights into their strengths and weaknesses [16]. In addition, the creation of a dual-perspective fusion network addresses aspect-based multimodal sentiment analysis, showing a trend towards the fusion of multiple data modalities, such as text and image, while concentrating on sentiment regarding specific aspects conveyed in the content [17].

Deep learning methods for sentiment analysis of tweets have also been used directly to contribute to the critical decision-making process, such as decisions regarding the COVID-19 pandemic [18]. To improve textual analysis, a weight-distributing method has been created that intelligently integrates sentiment dictionaries with TF-IDF to effectively emphasize words of high sentiment content while retaining crucial textual context, thus enhancing overall accuracy [19]. Through knowledge application, the SSK-DNN (Semantic and Sentiment Knowledge for Incremental Sentiment of Text Classification) approach enables incremental text sentiment classification in dynamic datasets in a way that the models adjust to evolving and enhancing knowledge over time [20].

The growing importance of decoding complex human sentiments online is clearly highlighted by the recent advances in sentiment analysis models that are increasingly more advanced and diverse. Zhang et al.'s research discusses the effectiveness of large language models (LLMs) sentiment analysis in software development. Small Language Models (SLMs) are found to perform very well in few-shot and zero-shot situations, particularly in low resource scenarios [21]. It means that, if sufficient labeled data is given, SLMs are capable of performing better than LLMs.

III. Methodology

The experiment was carried out with various combinations using models, tokenizers, add-ons, and preprocessing methods. The dataset contained a total of 250 tweets, 129 of which were negative and 121 of them were confirmed positive [22]. The objective of the Twitter Sentiment Analysis dataset is to label hate speech in social media updates. The tweets are grouped, with each such instance indicating the incidence of hate speech (labeled as '1' for racist or sexist language) or not (marked as '0'). Due to a concern for privacy, usernames within tweets are anonymized as @user. This dataset is employed as a baseline for comparison for model building and assessment that defines risky or offensive language on social media. The sentiment analysis system depicted in Fig. 1 proposes a methodology for sentiment classification of tweets. The model is apt for empirical research on various aspects of sentiment classification since it can accommodate a very broad range of text processing methods and learning frameworks. Raw tweets comprise the first stage of the operation, which runs through a few stages, such as tokenization, model choice, possible architectural level, and sentiment prediction at the end. The initial step in the pipeline is text preprocessing. There are two environments. In the first environment, the original tweets don't evolve, so the model can learn from the entire context. In the next arrangement, a filter keeps only the nouns from each tweet. The choice is based on the theory that assumes nouns will have dense meaning and can better capture the sentiment. Comparing both methods, the framework enables observers to examine how much information is maintained or lost through minimal preprocessing. Upon preprocessing, the tweets are translated into numerical formats (tokenization) in two various ways. The first one is the standard TF-IDF, which measures how important each word is across all documents by balancing frequency with uniqueness. This method is efficient and works well with regular machine learning methods.

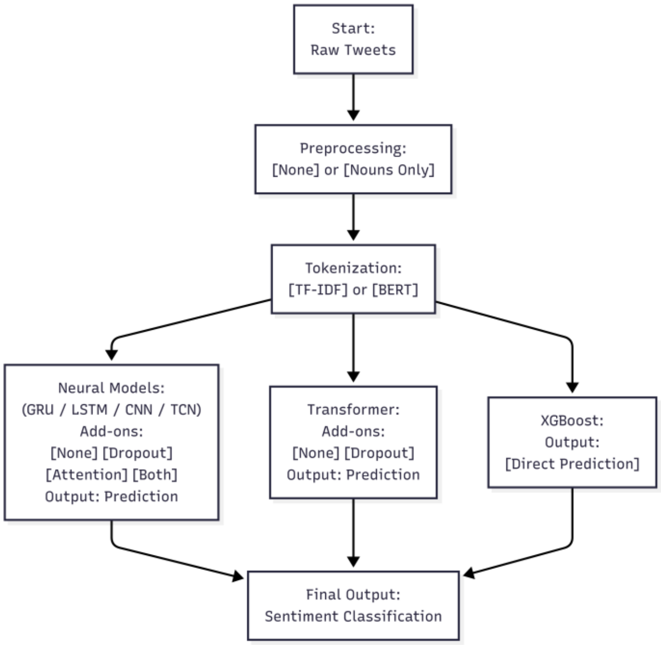

Fig. 1. Proposed model architecture.

The second approach is built on the BERT language model, which uses subword tokenization and captures context around each word by analyzing the entire sentence. BERT offers deeper language understanding, making it particularly effective for interpreting tweets that often contain sarcasm, abbreviations, or informal expressions. After this, the vectorized tweets are processed through one of the models available. These options include diverse set of models such as GRU, LSTM, CNN, TCN, Transformer, and XGBoost. Each model has its unique strengths and disadvantages. GRU and LSTM are recurrent neural networks which are generally ideal for sequential data and are good at capturing word correlations.

CNNs are used for shorter messages like tweets since they are good at observing brief patterns in text. Transformers are the foundation of many cutting-edge models because they use self-attention to determine the importance of each word in a sequence. By employing convolutions to capture long-range dependencies, TCNs provide an alternative to recurrent networks. A gradient-boosting approach that works well with TF-IDF features is called XGBoost. Despite not being a neural network, its accuracy and quickness with structured data make it very valuable. The framework enables the addition of additional layers that can improve learning or lessen overfitting in addition to choosing a basic model. These improvements include an attention layer that assists the model in focusing on pertinent textual elements for sentiment and a dropout layer that, for all models except XGBoost, randomly disables neurons during training to keep the model from becoming unduly specialized. To combine the advantages of targeted feature weighting and regularization, certain research might employ both layers. It is simpler to modify the architecture for various jobs or datasets with these options. The prediction layer, a fully connected layer that generates the sentiment categorization for every tweet, is the pipeline's last stage. Binary output is used to distinguish positive and negative sentiment. With an importance on binary classification and evaluation metrics such as accuracy, F1-score, and AUC-ROC, the system is carefully and clearly designed.

Generally, by changing preprocessing techniques, tokenization schemes, model topologies, and enhancement layers, allow for a wide variety of unique settings. Standard approaches like, combining TF-IDF features with classical or recurrent models, are depicted in the picture. Modern deep learning methods are also emphasized, particularly those that use attention processes and BERT-based embeddings. By encouraging methodical experimentation, this dual-path configuration enables researchers to evaluate the ways in which various elements affect model performance. This framework is a useful tool for both exploratory analysis and production-level sentiment classification on Twitter data because it supports a large variety of setups.

IV. Results and Discussion

The comparison of tokenization methods shown in Table 1 indicated that traditional TF-IDF consistently performed better than BERT-based tokenization across all major evaluation metrics. Minimal preprocessing, specifically none, and simple architecture resulted in the best performance. F1-score is the harmonic mean of precision and recall, capturing a balance between false positives and false negatives. AUC-ROC (Area Under the Receiver Operating Characteristic Curve) measures the model's ability to distinguish between classes across different thresholds.

TABLE I. Best Configurations

Embedding	Model	Add-on Layer	Accuracy	F1-Score	AUC-ROC
TF-IDF	GRU	none	0.792	0.789	0.889
TF-IDF	GRU	both	0.772	0.769	0.863
TF-IDF	XGBoost	both	0.768	0.768	0.820

On average, TF-IDF achieved an accuracy of about 58%, while BERT models only reached around 50%. The F1 score followed a similar pattern, with TF-IDF scoring around 0.48 compared to BERT's 0.36. Notably, TF-IDF models were also much more efficient, training over five times faster than BERT. The average time per run was about 4 seconds for TF-IDF versus 22 seconds for BERT. These results show that simpler, sparse vector representations like TF-IDF can be more practical and effective than complex contextual embeddings, especially when working with short and limited data like tweets.

Further, this paper also examined the impact of widely used deep learning additions such as dropout and attention mechanisms for a few models. The overall effect was relatively modest. Dropout provided the most consistent improvement, resulting in a slight increase in both F1 and AUC-ROC metrics, while attention alone had little impact. For instance, dropout configurations achieved an AUC-ROC of around 0.61, compared to about 0.58 when only using attention. Combining both did not yield a significant advantage over dropout alone, suggesting that in data-scarce situations, simpler regularization methods may be enough. Finally, we evaluated different preprocessing strategies. Using raw, unfiltered text without any part-of-speech filtering led to noticeably better results than extracting only nouns. Specifically, the full-text approach reached higher accuracy and F1 scores, around 55% and 0.44, respectively, while the "nouns-only" method slightly underperformed with values closer to 53% and 0.40. This indicates that even less important words in tweets, like adjectives or adverbs, may provide key sentiment clues that help models make better predictions. Table 2 and Fig. 2 present a summary of performance averaged across configurations for each model type.

TABLE II. MODEL-WISE PERFORMANCE SUMMARY

Model	Accuracy	F1	AUC-ROC	Training Time (s)
XGBoost	0.606	0.600	0.639	6.74
GRU	0.592	0.497	0.668	13.74
LSTM	0.545	0.415	0.655	13.95
CNN	0.506	0.340	0.636	13.80
Transformer	0.498	0.331	0.554	15.67
TCN	0.498	0.331	0.523	15.49

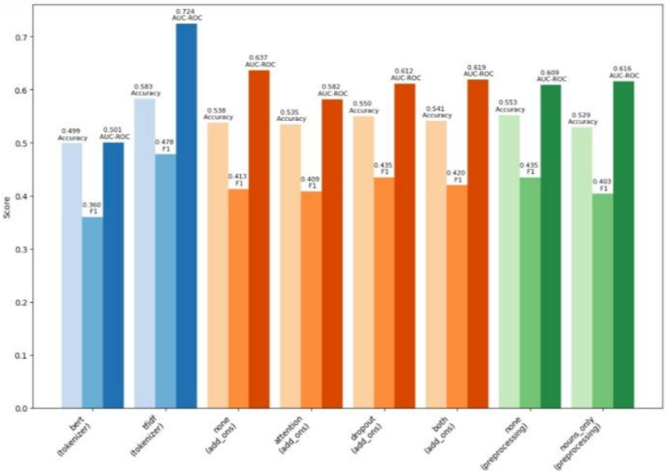

Fig. 2. Model performance summary.

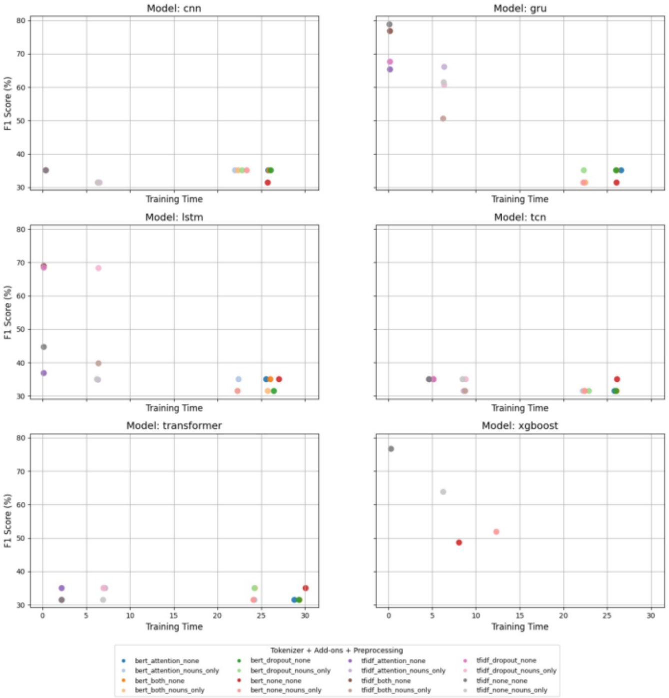

Fig. 3. Training time for all the models.

With an accuracy of 79.2%, the combination of TF-IDF and GRU is recommended for applications where accuracy is the top priority. As shown in Fig. 3, without additional preprocessing or add-ons, TF-IDF as the tokenizer and XGBoost as the model also performed well, with an accuracy of 76.8%, an F1-score of 0.768, and an AUC-ROC of 0.820. TF-IDF should be preferred over BERT in cases with comparatively limited resources.

V. CONCLUSION

To identify an effective and accurate model for sentiment analysis, this paper systematically evaluates various classifiers (GRU, LSTM, CNN, TCN, Transformer, and XGBoost) across two preprocessing types (raw vs. nouns-only) and two tokenization schemes (TF-IDF vs. BERT). The key findings are summarized as follows:

- Using raw, unfiltered text instead of limiting to nouns resulted in better performance, with noticeably higher accuracy and F1 scores.
- On average, TF-IDF achieved approximately 58% accuracy and an F1 score of 0.48, whereas BERT produced around 50% accuracy and an F1 score of 0.36. This clearly indicates that TF-IDF consistently outperformed BERT embeddings.
- The best-performing combination was TF-IDF with GRU (without add-ons), with an accuracy of 79.2%, F1 score of 0.789, and AUC-ROC of 0.889.
- Further, XGBoost was also performing relatively better with different configurations, which proves its robustness and its application in sentiment analysis.

Thus, the results demonstrate that the proposed model combination TF-IDF + GRU delivers superior performance for sentiment classification tasks.

References

[1] A. Vaswani *et al.*, “Attention Is All You Need,” in *Advances in Neural Information Processing Systems*, USA, 2017.

[2] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: https://doi.org/10.1145/2939672.2939785.

[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in *Advances in Neural Information Processing Systems*, USA, 2013.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: https://doi.org/10.18653/v1/N19-1423.

[5] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1724–1734. doi: https://doi.org/10.3115/v1/D14-1179.

[6] S. Bai, J. Z. Kolter, and V. Koltun, “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” Apr. 19, 2018, *arXiv*: arXiv:1803.01271. doi: https://doi.org/10.48550/arXiv.1803.01271.

[7] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Jul. 26, 2019, *arXiv*: arXiv:1907.11692. doi: https://doi.org/10.48550/arXiv.1907.11692.

[8] S. Bengesi, T. Oladunni, R. Olusegun, and H. Audu, “A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets,” *IEEE Access*, vol. 11, pp. 11811–11826, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3242290.

[9] M. Rodriguez-Ibanez, F.-J. Gimeno-Blanes, P. M. Cuenca-Jimenez, C. Soguero-Ruiz, and J. L. Rojo-Alvarez, “Sentiment Analysis of Political Tweets From the 2019 Spanish Elections,” *IEEE Access*, vol. 9, pp. 101847–101862, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3097492.

[10] S. Aloufi and A. El Saddik, “Sentiment Identification in Football-Specific Tweets,” *IEEE Access*, vol. 6, pp. 78609–78621, 2018, doi: https://doi.org/10.1109/ACCESS.2018.2885117

[11] M. F. Almufareh, N. Jhanjhi, N. A. Khan, S. N. Almuayqil, M. Humayun, and D. Javed, “BertSent: Transformer-Based Model for Sentiment Analysis of Penta-Class Tweet Classification,” *IEEE Access*, vol. 12, pp. 196803–196817, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3515836.

[12] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, “Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets,” *IEEE Access*, vol. 8, pp. 181074–181090, 2020, doi: https://doi.org/10.1109/ACCESS.2020.3027350.

[13] A. Sherin, I. Jasmine Selvakumari Jeya, and S. N. Deepa, “Enhanced Aquila Optimizer Combined Ensemble Bi-LSTM-GRU With Fuzzy Emotion Extractor for Tweet Sentiment Analysis and Classification,” *IEEE Access*, vol. 12, pp. 141932–141951, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3464091.

[14] M. Bouazizi and T. Ohtsuki, “Multi-Class Sentiment Analysis in Twitter: What if Classification is Not the Answer,” *IEEE Access*, vol. 6, pp. 64486–64502, 2018, doi: https://doi.org/10.1109/ACCESS.2018.2876674.

[15] S. E. Saad and J. Yang, “Twitter Sentiment Analysis Based on Ordinal Regression,” *IEEE Access*, vol. 7, pp. 163677–163685, 2019, doi: 10.1109/ACCESS.2019.2952127.

[16] A. R. Lubis, Y. Fatmi, and D. Witarsyah, “Comparison of Transformer Based and Traditional Models on Sentiment Analysis on Social Media Datasets,” in *2023 6th International Conference of Computer and Informatics Engineering (IC2IE)*, Lombok, Indonesia: IEEE, Sep. 2023, pp. 163–168. doi: https://doi.org/10.1109/IC2IE60547.2023.10331232.

[17] D. Wang, C. Tian, X. Liang, L. Zhao, L. He, and Q. Wang, “Dual-Perspective Fusion Network for Aspect-Based Multimodal Sentiment Analysis,” *IEEE Trans. Multimedia*, vol. 26, pp. 4028–4038, 2024, doi: https://doi.org/10.1109/TMM.2023.3321435.

[18] K. Fathima and A. R. M. Shanavas, “COVID-19 Related Decision Making by Deep Learning Based Sentiment Analysis of Tweets,” in *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, Bengaluru, India: IEEE, Mar. 2024, pp. 1–5. doi: https://doi.org/10.1109/ICDCOT61034.2024.10515929.

[19] H. Liu, X. Chen, and X. Liu, “A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis,” *IEEE Access*, vol. 10, pp. 32280–32289, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3160172.

[20] J. Khan, N. Ahmad, C. Choi, S. Ullah, G. Kim, and Y. Lee, “SSK-DNN: Semantic and Sentiment Knowledge for Incremental Text Sentiment Classification,” in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, Shanghai, China: IEEE, Dec. 2023, pp. 52–59. doi: https://doi.org/10.1109/ICDMW60847.2023.00016.

[21] T. Zhang, I. C. Irsan, F. Thung, and D. Lo, “Revisiting Sentiment Analysis for Software Engineering in the Era of Large Language Models,” *ACM Trans. Softw. Eng. Methodol.*, vol. 34, no. 3, pp. 1–30, Mar. 2025, doi: https://doi.org/10.1145/3697009.

[22] Ali Toosi, “Twitter Sentiment Analysis,” https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech, last accessed on 28 June 2025.