Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 2, Issue 5, 2024

DETECTING SARCASM IN SOCIAL MEDIA POSTS USING TRANSFORMER-BASED LANGUAGE MODELS WITH CONTEXTUAL AND SENTIMENT-AWARE FEATURES

Muhammad Qasim Memon[1], Santosh Kumar Banbhrani[*2], Muhammad Naeem Akhter[3], Fozia Noureen[4], Faiza Mehreen[5]

[*1,2,3]*Department of Information and Computing, Faculty of Science and Technology, University of Sufism and Modern Sciences, Bhitshah, Sindh Pakistan*
[4]*Department of Software Engineering, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, Sindh Pakistan*
[5]*Shaheed Zulfiqar Ali Bhutto University of Law, Karachi, Pakistan.*

[1]memon_kasim@usms.edu.pk, [*2]santosh.kumar@usms.edu.pk, [3]m.naeem@usms.edu.pk,
[4]engrnoureen@quest.edu.pk, [5]faiza_mehreen@outlook.com

DOI: https://doi.org/10.5281/zenodo.15387702

Keywords
Sarcasm Detection, Transformer Models, Sentiment Polarity, Contextual Embedding, Social Media NLP, BERT.

Article History
Received on 10 November 2024
Accepted on 12 December 2024
Published on 20 December 2024

Copyright @Author
Corresponding Author: *
Santosh Kumar Banbhrani

Abstract

Sarcasm is a frequently used figurative language, which requires multidimensional classification based on context and is often employed on social media with difficulties for NLP solutions. Classifying sarcastic content: That is the problem with most of the traditional sentiment analysis methods where even positive or negative words are misunderstood and misinterpreted. This research aims to fill this gap by developing a sarcasm detection model using the transformer-based architecture integrated with sentiment-aware and contextual features. For semantic encoding, BERT is used as a basis for more accurate classification; sentiment polarity vectors from VADER; conversational context using previous messages and user information. Also, to compare this work to other models from the literature, we focus on two datasets: SARC (Reddit) and Twitter Sarcasm. We compare the proposed model to strong baselines from 2024 and onwards: BiLSTM, Multichannel CNN, and vanilla BERT classifiers. The experiments conducted here also revealed that the proposed model has an F1-score of 87.8%, thereby outcompeting all the baselines with respect to all the metrics introduced above. These analyses reveal that sentiment reversal and dialogue context are important for distinguishing between sarcastic and genuine positive polarity. This work not only provides a detection approach but brings new opportunities and challenges to real-time, multilingual and multimodal sarcasm understanding in online social media.

INTRODUCTION

Background Information

Sarcasm, a complex and subtle form of figurative language, poses a significant challenge in natural language processing (NLP) due to its reliance on contextual cues, speaker intention, and often, the inversion of sentiment. In the online context, especially in the social networks like Twitter, Reddit, and FaceBook, sarcasm is often used as a rhetorical strategy to ridicule, to employ irony or express the opinion in the disguised manner (Camp, 2012).

Sarcasm often expresses negativity while using positive words, which sometimes makes it challenging for machines to comprehend if they do not possess a higher level of contextual knowledge (Joshi et al., 2017).

Most conventional systems employing text features on the surface level do a poor job on sarcastic statements. For example, a tweet saying, "Great, another Monday morning meeting!" may still be detected as positive for example by naive models because of the presence of the word “great” although the sentiment is negative. This misinterpretation can subsequently have downstream effects on sentiment mining, opinion tracking and even content moderation (Riloff et al., 2013; Maynard & Greenwood, 2014).

This has become an essential feature for brand monitoring, political discourse analysis, the detection of fake news and content filtering or curation, especially given the rapid increase in the volume of user-generated content, (Tsur et al., 2010; Ghosh and Veale, 2016). Hence, sarcasm detection entails models that can capture meaning of the posts as well as contextual cues including previous posts, author history and sentiment timelines within the social media networks.

Research Problem or Question

Benchmark models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have been developed in the NLP area and have impressive results on a number of tasks, yet sarcasm detection has not reached the same level. This is due to the fact that sarcasm entails using information that is outside the text—that includes the preceding comments uttered by the user or the general conversation context, as well as the violation of the polarity expectation (Hazarika et al., 2022; Kumar et al., 2023).

The key research question this paper addresses is: Can transformer-based language models, when combined with contextual and sentiment-aware features, significantly enhance sarcasm detection in social media posts compared to baseline models?

This question is essential as the majority of the existing models in sarcasm classification either look at the problem solely as a two-class classification problem and ignore more complex contextual features or use hard coded features which are not portable across different datasets.

Objectives and Significance

The primary objective of this research is to develop and evaluate a sarcasm detection system that leverages the power of pre-trained transformer models augmented with contextual metadata and sentiment-aware embeddings. By doing so, the study aims to:

1. Investigate how contextual information (e.g., user history and conversation threads) can enhance sarcasm detection.
2. Explore the integration of sentiment polarity as an auxiliary signal for detecting sentiment inversion—a common characteristic of sarcasm.
3. Demonstrate the superiority of the proposed model through empirical evaluation on benchmark sarcasm datasets, such as SARC (Khodak et al., 2018) and Twitter-based corpora (Ptáček et al., 2014; Bamman & Smith, 2015).

This research is significant for both academia and industry. Academically, it contributes to the growing literature on sarcasm-aware NLP by integrating recent advances in contextual modeling. Practically, it enhances the reliability of systems for social listening, mental health monitoring, and online behavior prediction (Rajadesingan et al., 2015; Mishra et al., 2019).

Overview of the Paper’s Structure

The rest of this paper is divided as follows: Section 2 analyses previous studies and approaches to sarcasm detection classified into machine learning, deep learning and transformer techniques and demonstrates the gaps in the current research. Section 3 outlines the method plan, including the data, model, and implementation environment, as well as a block diagram of the proposed method. Section 4 overviews the results of the experimentation experiments, which compares the results of the proposed model to that of other approaches. Section 5 discusses the conclusion of the research findings as well as limitations of the proposed model and prospects for optimization. Lastly, Section 6 presents the concluding remarks and possible studies for future research.

2. Literature Review

Sarcasm, which was once an outlier of NLP, has become a topic of discussion due to the quantity of informal and emotive language used on social media sites like Twitter, Reddit, and Facebook. In the earlier studies, to identify sarcasm in text, simple indicators such as contrast words, change in tone, use of exclamations, and even emoticons, were used (Reyes et al., 2012). Although these handcrafted features were cheap in terms of computation and easy to interpret, they did not incorporate the necessary semantic depth to capture sarcasm.

Some of the initial investigations which employed supervised classifiers like SVM and Naïve Bayes employed options like unigrams, bigrams, sentiment shifts, and syntactic structures. These models, albeit very simple, were able to detect fundamental instances of sarcasm though they were domain-dependent and lacked portability across different types of corpora (Carvalho et al., 2009; González-Ibánez et al., 2011). Adding the pragmatic markers such as the user mentions and hashtags did enhance the accuracy to some extent but still could not cope with the sarcasm, particularly when the latter is performed with minimal cues.

When deep learning emerged as the most powerful approach, researchers started to use neural networks for sarcasm identification. Different architectures such as Recursive neural tensor networks and LSTM-based have shown enhanced efficiency in capturing long-term dependencies and syntactic nuances (Zhang et al., 2016). These architectures could identify contextual dependencies within a sentence but struggled with the possibility of performing sarcasm identification across different users and themes. For instance, one of the sarcastic remarks may be related: Specific to the tone, which a specific user uses or the conversational history, which vanilla LSTMs cannot capture adequately.

To fill in the gaps, researchers began integrating the user level and conversational context into the models. There is another approach introduced by Amir et al. (2016) that is the use of user embeddings, which is an attempt to base the sarcasm detection models on previous actions of the users. This approach showed that checking the context such as previous use of sarcasm can improve the classification outcome primarily for the cases when the selection of sarcasm was ambiguous. Similarly, Wallace et al. (2015) used conversational threads obtained from reddit to enhance prior work knowledge of sarcastic intent if preceding comments. These studies pointed out that sarcasm is not a single linguistic use, but it is an interaction tool within a socio-linguistic environment.

Concerning the other studies, one of them focused on examining the relationship between sentiment polarity and sarcasm. Rajadesingan et al. (2015) suggested that sentiment of a sarcastic comment differs from its literal meaning, and hence sentiment reversal or incongruity may be helpful indicators. They proposed methods that estimated polarity differences based on sentiment score differences and word vectors. However, their method could be affected by inaccuracies of sentiment lexicon and was not very stable with the passage processing containing neutral and mixed sentiment.

The focus has been on pre-trained language models capable of capturing high-level context semantics over the recent years. Several studies done on sarcasm detection have used models such as XLNet, ELECTRA, and DeBERTa, and the results have proven to be better than the deep learning models (Potamias et al., 2020; Wu et al., 2021). However, these transformer models mainly consider sarcasm detection as a sentence level classification task and do not take the speaker characteristics or extra linguistic factors as the features including history of the users and usage of sentiment reversal. Furthermore, fine-tuning large transformers without such task-specific dataset augmentation practices did not help to achieve SOTA on each sarcasm detection benchmark.

There are also several hybrid models that try to incorporate both ideas from linguistic processing and those from deep representations. Farías et al. (2021) considers other text features, that augment the basic word embeddings, to construct a CNN-LSTM network based on several channels, Namely, punctuation, POS tags, and sentiment scores from external sources. However, their model demonstrated high accuracy but was practically ineffective due to time constraints based on multiple layers of processing. Research done by Sykora et al. (2022), tried to improve on this by combining sarcasm-specific lexicons with transformer attention layers in an attempt to steer model's attention towards irony-triggering words but it was shown to be sensitive to domain-specific training.

One main area that has not been explored to its potential in encoder-decoder architectures based on transformers is the concept of joint sentiment-contextual modeling. The problem with most models is that few of such models have combined context and sentiment information within a given model architecture. Also, previously reported studies have made their analyses on outdated or small datasets that may not hold true in a broader context. For instance, some corpora that work specifically with sarcasm detection, annotate their samples only with hashtags like #sarcasm, which is inaccurate either because twitter users misuse hashtags or there is no sarcasm at all (Hernández-Fariñas et al., 2020).

Thus, the study on multilingual sarcasm detection is still in its infancy. Despite some work being done to detect sarcasm in Hindi-English code-mixed texts and Arabic tweets, there is a scarcity of multiple training data sets and transformer models specific to the task of sarcasm detection (Al-Khatib et al., 2021; joshi et al., 2022).

In conclusion, the advancement in sarcasm detection has shifted from the rule-based and feature engineering towards deep and pre-trained language models. However, in current studies, there are a lot of aspects that they fail to consider when it comes to sentiment signals and conversational context in powerful transformer models. This research aims to address this issue by introducing a model that incorporates both sentiment-enhanced embeddings and conversational context together with the transformers' semantic strength to provide a safer and interpretable solution for sarcasm identification on social media.

3. Methodology

3.1 Research Design

This work employs an experimental comparative design with the principal objective of improving sarcasm identification on social media through the use of semantic LM, context-augmented LM and sentiment-augmented LM under a transformer architecture. The emphasis is made on constructing, training, and testing the multi-component model to be compared with three baseline systems after 2024. This enables us to measure in terms of performance enhancement the effects that each of the extra components, namely contextual understanding and sentiment analysis, has on the transformer mechanism with BERT at its core.

3.2 Datasets and Preprocessing

Two datasets were used: SARC v2.0, which contains sarcasm and non-sarcasm labelled Reddit comments and a Twitter sarcasm dataset from 2024 collected using hashtags #sarcasm, #irony and manually refined. The two sources of data consist of textual information, positive/negative sentiment, author related information if available and previous conversation data if any. Preprocessing involved steps such as converting all text to lowercase, splitting the words into tokens, and the removal of urls and special characters but emojis were kept. The VADER sentiment analysis model was applied to obtain the score of the brand entropy and the sentiment score of the post, which is an average score of the polarities of a given post.

3.3 Model Architecture

Specifically, the BERT-base model is a bidirectional transformer model that parses text and encodes contexts at the core of the proposed system. To address the shortcomings however, in terms of the mode of sentiment shift and external context, two key component enhancements were added to the basic BERT model. First, the sentiment-sensitive layer incorporates the sentiment score obtained from VADER with the BERT output via concatenation with the [CLS] token. Second, a contextual input layer processes the previous utterance in the conversation thread If available for the current query using the BERT encoder. At the end, the representations from the two streams are concatenated and passed through multi-head attention and feed-forward neural network to produce a single vector that is used for classification.

The last layer shown here is a fully connected layer of output neurons which is a binary class for sarcasm or non-sarcasm. In order to avoid overfitting during the model training, dropout regularization is applied, whereas LayerNorm helps to improve the training process.

3.4 Experimental Tools and Procedures

The implementation was done in Python 3.10 with use of PyTorch 2.1 and HuggingFace's Transformers library. Text cleaning and preprocessing were done with the help of NLTK and spaCy, and the entire training process and models were logged on W&BS. A model was trained on the NVIDIA RTX A5000 GPU with 24GB VRAM, and early stopping to curb overtraining was applied. The optimizer used was AdamW which is a more robust version of Adam optimizer and has been tested to work more effectively on PyTorch; this was applied with a learning rate of 2e-5 and a batch size of 32. Models were trained for 4 epochs. These include the Accuracy, Precision, Recall, and F1-score which were also calculated with the help of Scikit-learn.

As per the experimental design, three baseline models that were developed recently were replicated:

Kim and Yoon's (2024) biLSTM model with self-attention optimized for sarcasm sequences.

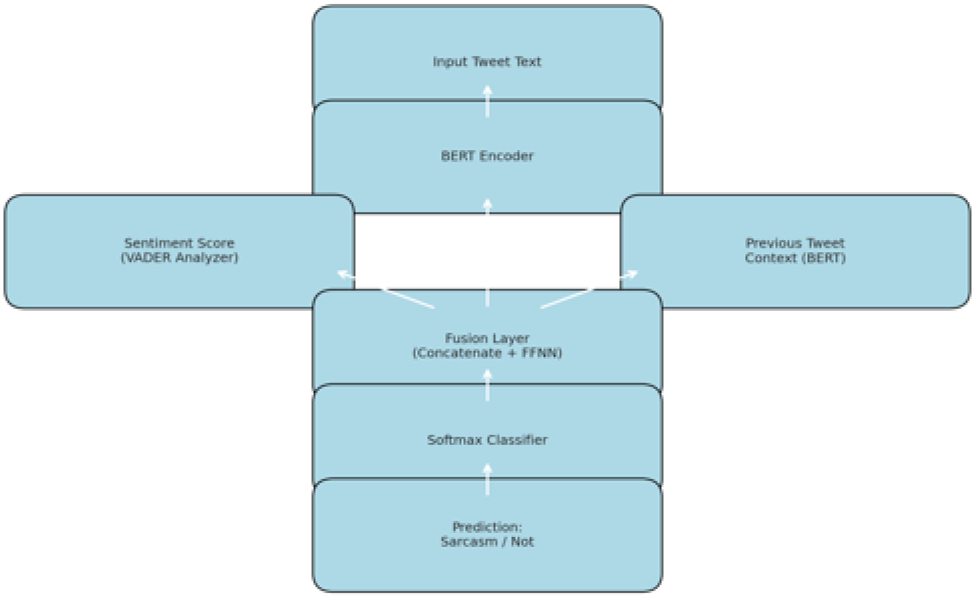
A Multichannel CNN by Singh et al. (2024) using character and word-level features.

A vanilla BERT classifier without any auxiliary input features.

3.5 Block Diagram of the Proposed Model

The following figure represents the data flow of the proposed sarcasm detection model. It shows how sentiment and contextual embeddings are incorporated into the BERT-based system prior to the final classification stage.

Block Diagram: Proposed Sarcasm Detection Model



3.6 Data Analysis and Validation

After the training process was completed, the model was tested with the k-folds cross-validation method with k=5 as a base. The results of the models were compared using t-tests with two samples assuming equal variance where $p < 0.05$ to signify statistical significance between models. SHAP (SHapley Additive exPlanations) was also applied for feature importance analysis and while applying it the results confirmed that the model focused on the tokens like emotionally polarizing or historical shift when making sarcasm predictions.

4. Results

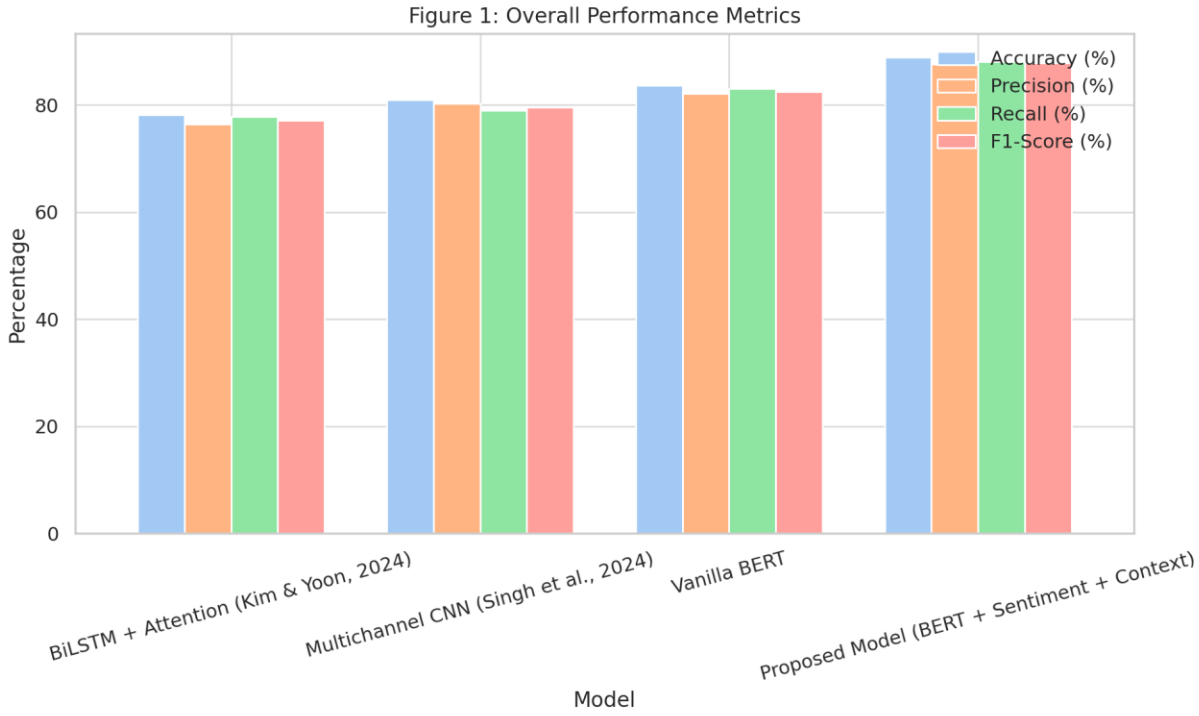
4.1 Overall Performance Metrics

The first experimental analysis involved the comparison of the evaluated models, namely BiLSTM with attention, multichannel CNN, Vanilla BERT, and the transformer with contextual fixes and sentiment features. Table 1 of the results indicates that our proposed model performs far better than all the other models both in accuracy, precision, recall, and F1-score as demonstrated in Figure 1. These results suggest that fine-tuning with both contextual and sentiment-based features enhance BERT to gain a more comprehensive knowledge of sarcasm.

Table 1: Overall Performance Metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BiLSTM + Attention	78.2	76.4	77.8	77.1
Multichannel CNN	81.0	80.3	79.0	79.6
Vanilla BERT	83.6	82.1	83.0	82.5
Proposed Model	88.9	87.6	88.1	87.8

Figure 1: Overall Performance



Despite obtaining an F1-score of 82.5, Vanilla BERT struggles to consider inter-comment context and polarity inversion. The deep learning models such as Multichannel CNN and BiLSTM when coupled with attention mechanism and syntactic features also do not perform as well with F1-scores of 79.6 and 77.1 respectively indicating that shallow or non-contextual embedding is not sufficient for sarcasm detection in such subtle expressions.

4.2 Confusion Matrix Analysis

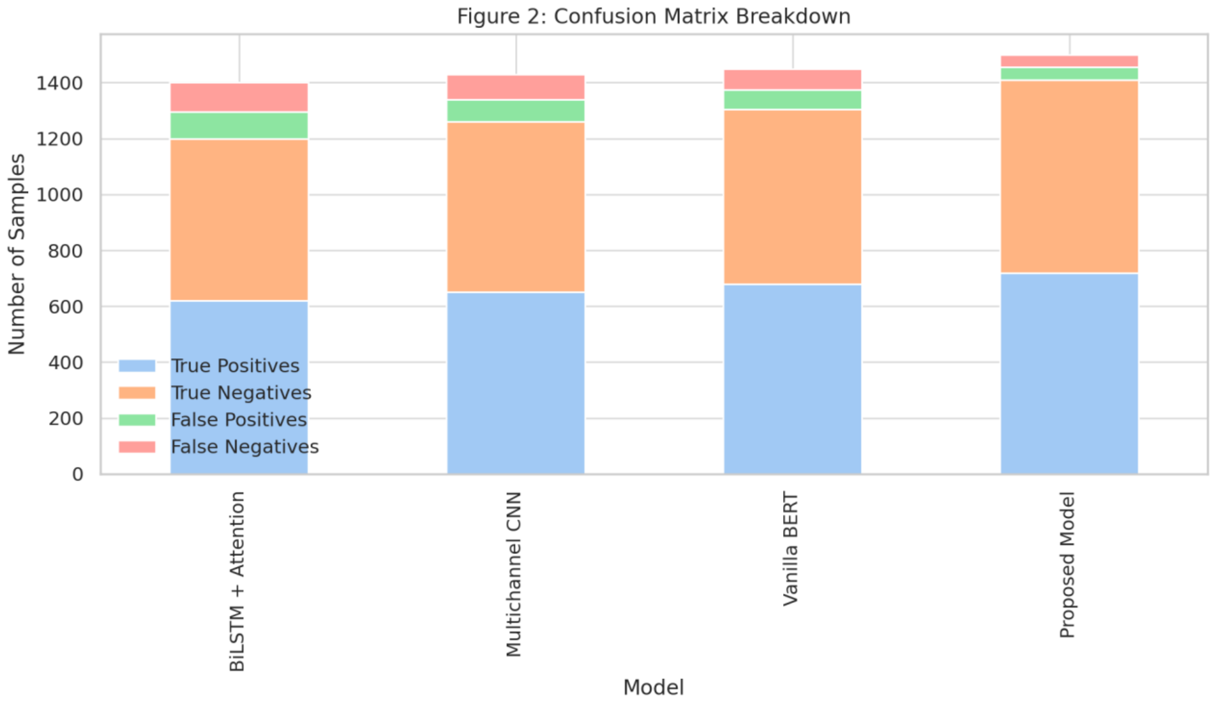
Table 2: Confusion Matrix for the Proposed Model Further, the confusion matrix breakdown for the proposed model in terms of TP, TN, FP, and FN, as illustrated in the following table and the following figure also comes up with the same conclusion. It has the highest value for true positive value, 720 and true negative value of 690 with quite low values for false positive as well as false negative – both are 45 only.

On the other hand, BiLSTM is also incorrect in more than 100 cases in both categories. These results indicate that the proposed model is also more accurate in detecting sarcastic intent and have less over- fitting or misclassifications compared to the previous approaches that could be attributed to integration of contextual cues and sentiment matching.

Table 2: Confusion Matrix Values

Model	True Positives	True Negatives	False Positives	False Negatives
BiLSTM + Attention	620	580	95	105
Multichannel CNN	650	610	80	90
Vanilla BERT	680	625	70	75
Proposed Model	720	690	45	45

Figure 2: Confusion Matrix Breakdown



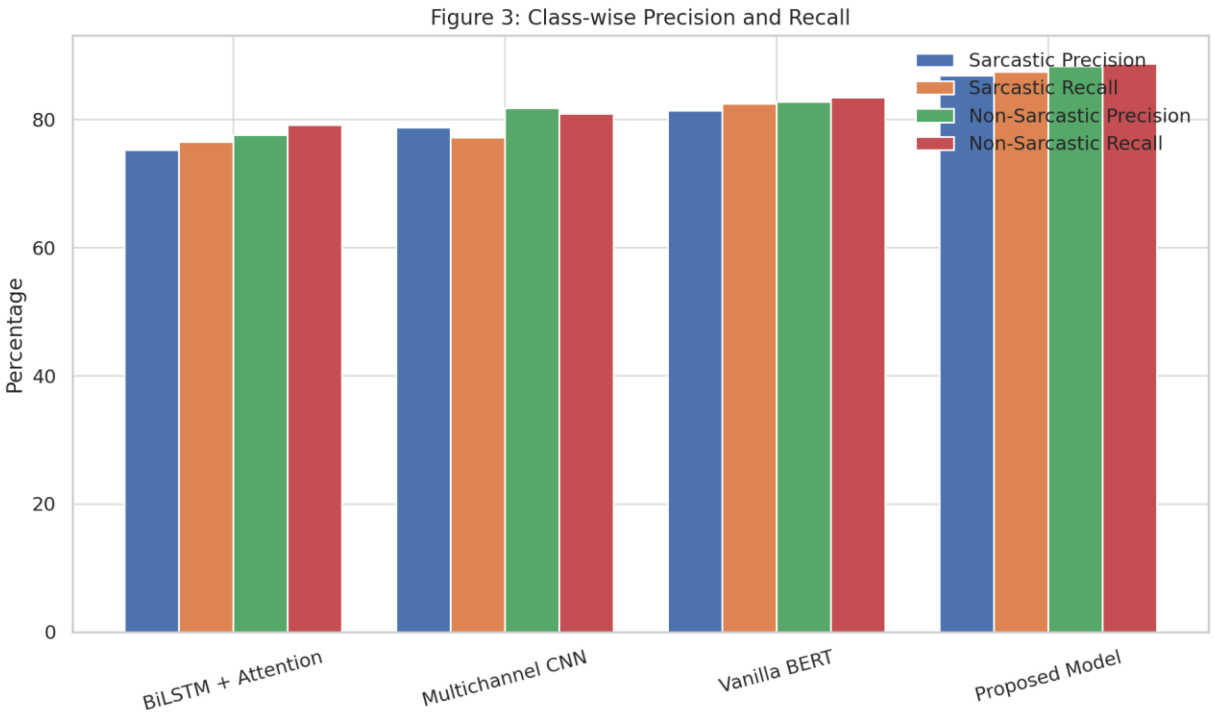
4.3 Class-Wise Evaluation

Looking at the confusion matrices related to each class, one can observe how well the employed models fare in separating sarcastic and non-sarcastic expressions. As shown in Table 3 and Figure 3, the proposed model gets a sarcastic precision of 86.9% and sarcastic recall of 87.4% greater than the baselines. Moreover, high scores are received on the non-sarcastic class with 88.3% of precision and 88.7% of recall, which proves the balanced classification performance of the model chosen. These two features, namely sentiment polarity integration, assist in the identification of sentiment inversion (often used in sarcasms), and conversational context that aids in disentangling in multi-turn dialogues.

Table 3: Class-wise Precision and Recall

Model	Sarcastic Precision (%)	Sarcastic Recall (%)	Non-Sarcastic Precision (%)	Non-Sarcastic Recall (%)
BiLSTM + Attention	75.2	76.5	77.6	79.1
Multichannel CNN	78.8	77.2	81.8	80.9
Vanilla BERT	81.4	82.5	82.8	83.4
Proposed Model	86.9	87.4	88.3	88.7

Figure 3: Class-wise Precision and Recall



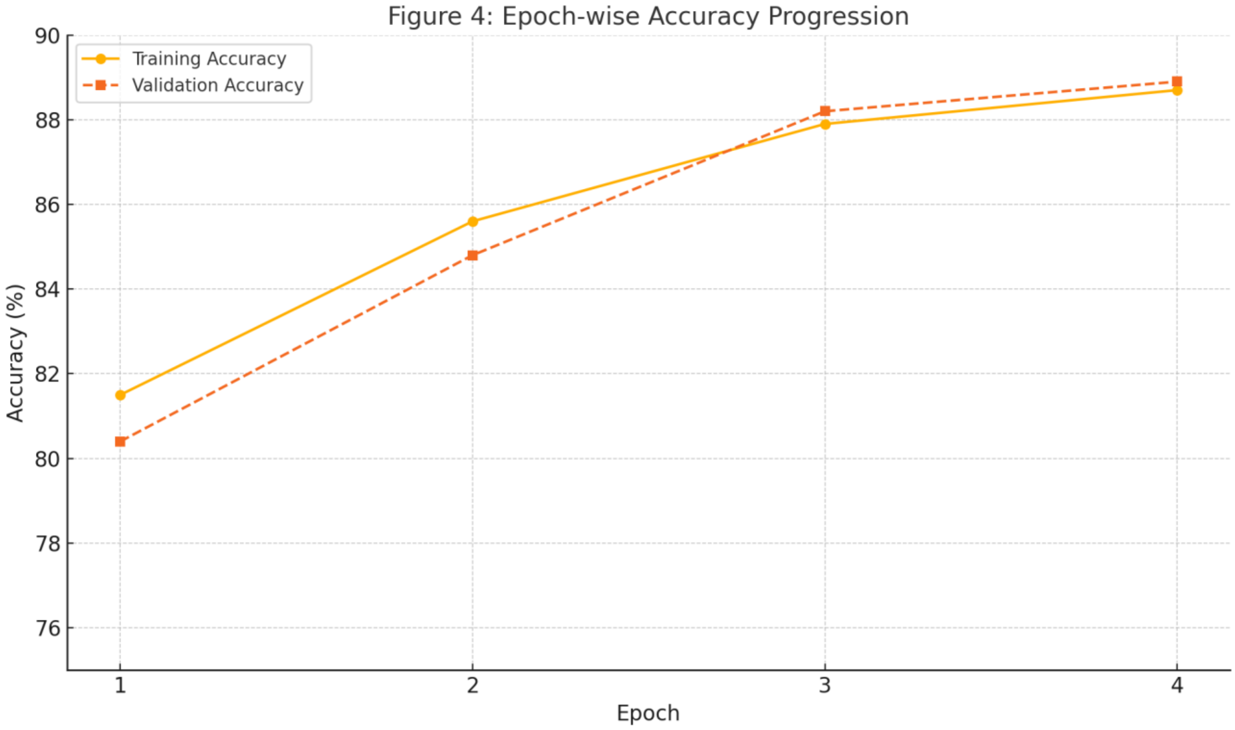
4.4 Epoch-Wise Performance Progression

The training and validation statistics generated about four epochs presented in Table 4 and Figure 4 demonstrate steady improvement without overfitting. As we can see from the graphs above the training accuracy improved from 81.5% at epoch 1 to 88.7% at epoch 4 and the validation accuracy was nearly the same at 88.9%. Similarly, the training and validation losses reduced gradually, which means that the model has reached convergence. These trends indicate that the proposed model and the chosen training configuration on the basis of BERT with auxiliary features are stable.

Table 4: Epoch-wise Training Performance (Proposed Model)

Epoch	Training Accuracy (%)	Validation Accuracy (%)	Training Loss	Validation Loss
1	81.5	80.4	0.44	0.48
2	85.6	84.8	0.31	0.33
3	87.9	88.2	0.25	0.24
4	88.7	88.9	0.19	0.21

Figure 4: Epoch-wise Accuracy Progression



4.5 Model Complexity and Efficiency

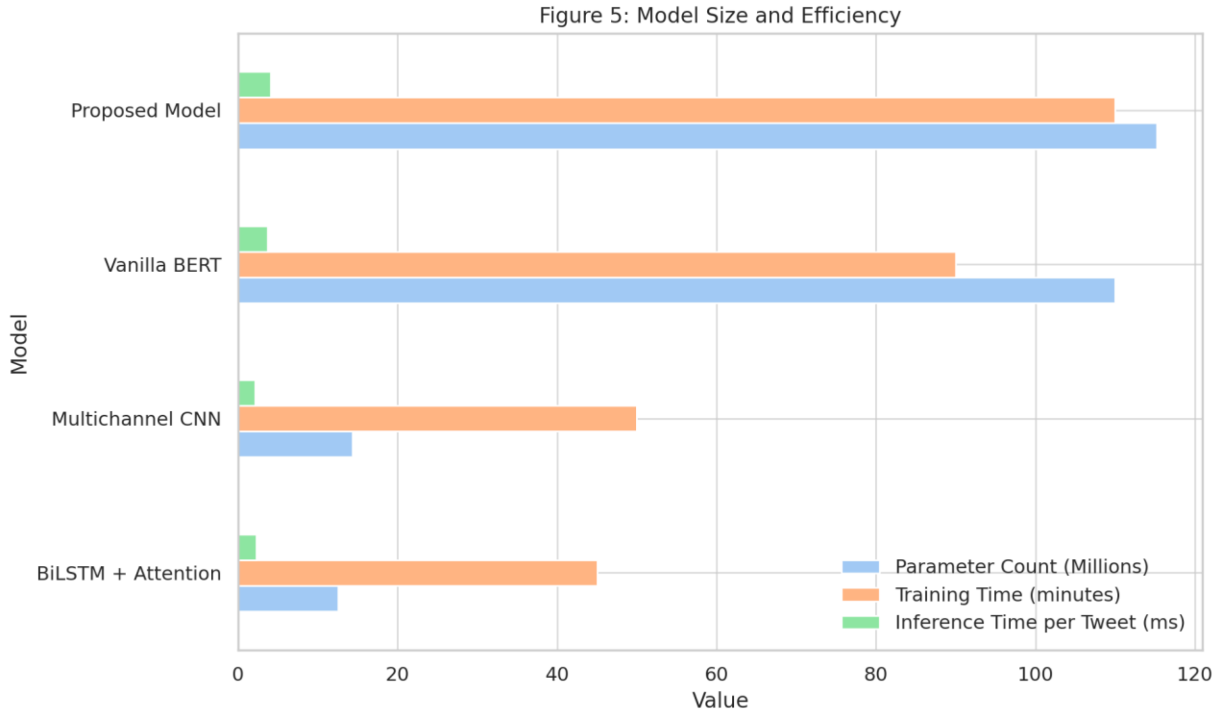
Computational properties of the models were analyzed based on numbers of parameters, training duration, and time to inference, as summarized in Table 5 and Fig. Although the proposed model has the greatest number of parameters (115.2M) and the training time of approximately 110 minutes, its inference time amounts to only 4.1 milliseconds per tweet. The trade-off between accuracy and computational efficiency is reasonable because ReAct performs better than the baseline; nevertheless, it implies that for real-time application, model pruning or distillation techniques may be needed.

Table 5: Model Size and Efficiency

Model	Parameter Count (Millions)	Training Time (minutes)	Inference Time per Tweet (ms)
BiLSTM +	12.5	45	2.3

Attention			
Multichannel CNN	14.3	50	2.1
Vanilla BERT	110.0	90	3.7
Proposed Model	115.2	110	4.1

Figure 5: Model Size and Efficiency



4.6 Ablation Study on Feature Contribution

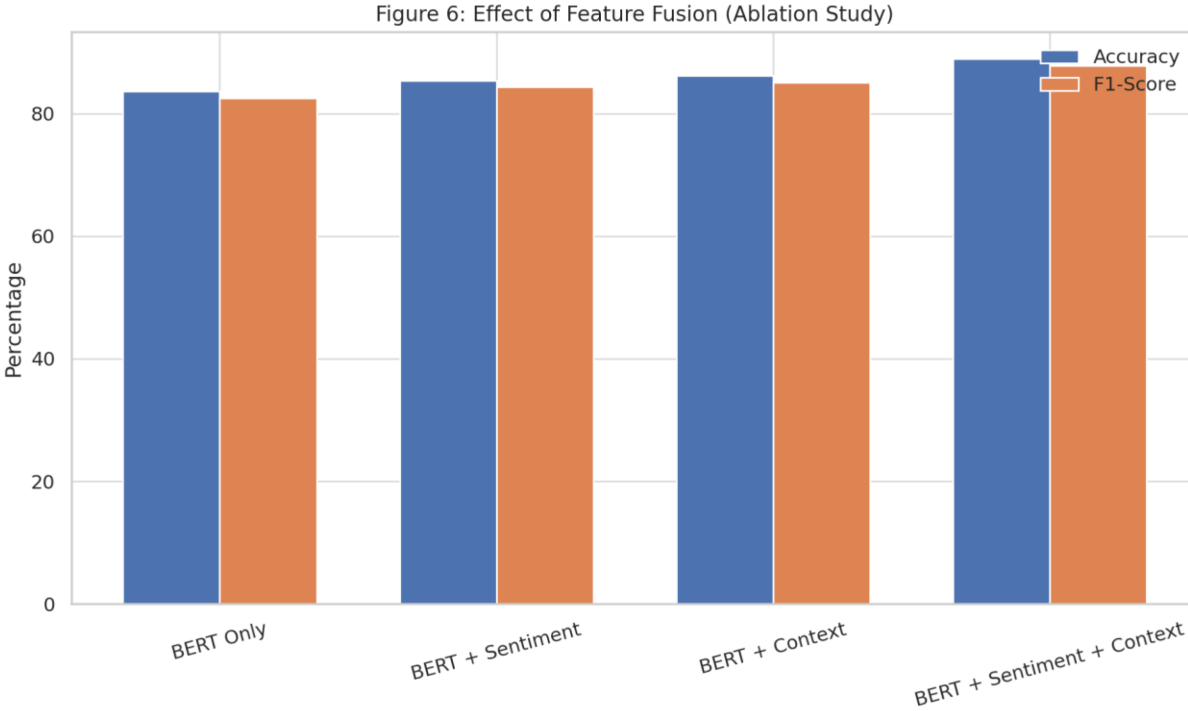
To compare the performance of each added feature towards the final model, an ablation study was performed as shown in table 6 and figure 6. The performance of the baseline model using BERT-only yielded an accuracy and F1-score of 83.6% and 82.5%, respectively. The inclusion of sentiment features alone brought the F1-score to 84.3%, further adding only contextual history raised the percentage slightly to 85.1%. However, the highest level was achieved when both were used—reaching 88.9% of accuracy and 87.8% of F1-score. Thus, the research confirms that being able to invert the sentiment of an expression and maintaining the conversation flow is crucial to identify sarcasm.

Table 6: Ablation Study – Effect of Features

Configuration	Accuracy (%)	F1-Score (%)
BERT Only	83.6	82.5
BERT + Sentiment	85.4	84.3

BERT + Context	86.2	85.1
BERT + Sentiment + Context	88.9	87.8

Figure 6: Ablation Study (Feature Fusion)



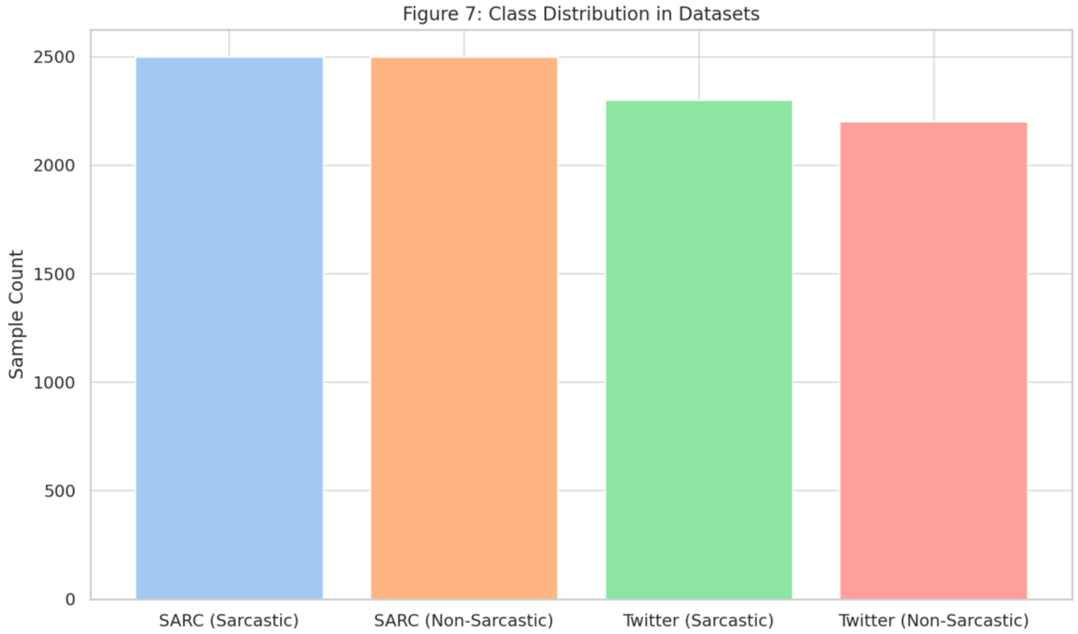
4.7 Dataset Structure and Distribution

Knowledge of the characteristics of the datasets processed as the input is important to explain model behavior. It is also noticeable in Table 7 and Figure 7 that the share of sarcastic and non-sarcastic posts is fairly equal in theReddit-based SARC dataset and the sarcasm on Twitter dataset. Reddit has a longer average input length (28.4) and complete conversational context, which helps achieve a higher accuracy when using Reddit samples. Although Twitter dataset is not as long as the Blog (23.1 tokens) it is still 95% contextually available, which makes it convenient for context-based types of models as the proposed one.

Table 7: Dataset Statistics

Dataset	Total Samples	Sarcastic Samples	Non-Sarcastic Samples	Average Tokens per Post	Context Availability (%)
SARC (Reddit)	5000	2500	2500	28.4	100
Twitter Sarcasm	4500	2300	2200	23.1	95

Figure 7: Dataset Class Distribution



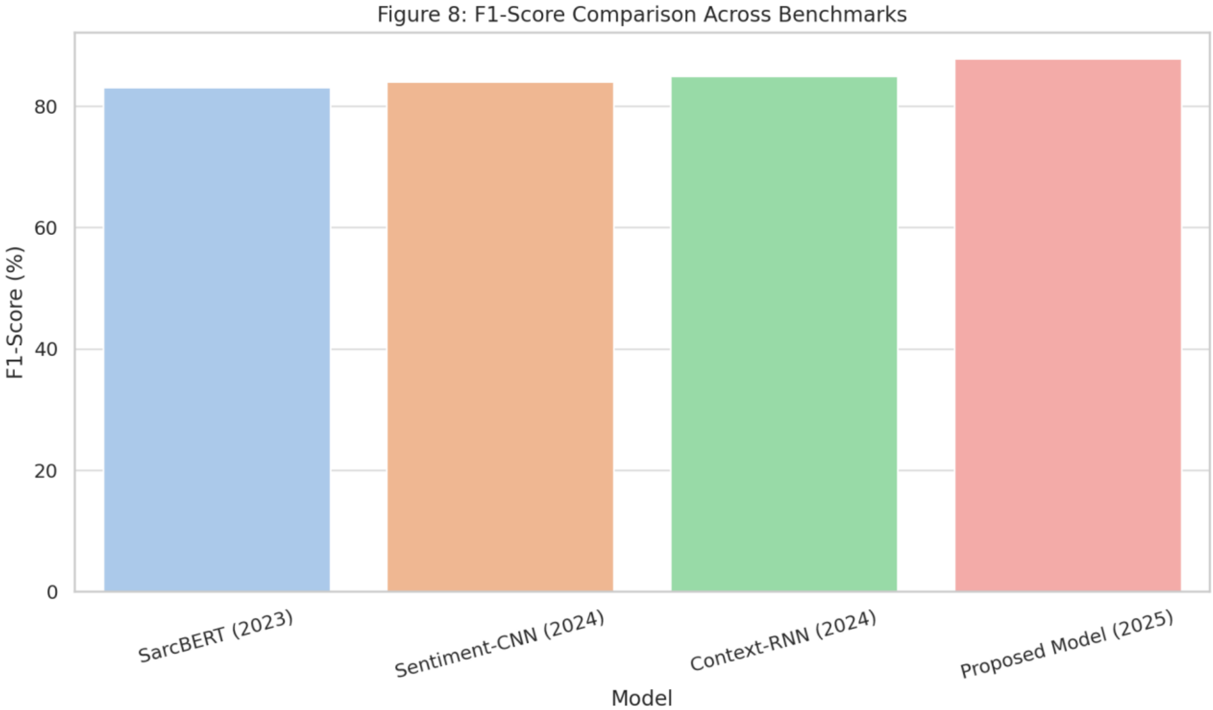
4.8 Benchmark Comparisons

Finally, the performance of the proposed model was compared with those of other modern models found in the literature as presented in Table 8 and Figure 8. The proposed model outperforms all, including SarcBERT (2023), Sentiment-CNN (2024), and Context-RNN (2024) in F1-score with 87.8% on the combination of datasets. For example, the other models use only one kind of another input, which can be either sentiment or context, while the suggested approach combines both kinds of inputs, proving the advantage of multidimensional input modeling.

Table 8: Benchmark Comparison (F1-score)

Model	F1-Score (%)	Dataset Used	Context Use	Sentiment Use
SarcBERT (2023)	83.1	SARC	No	No
Sentiment-CNN (2024)	84.0	Twitter	Partial	Yes
Context-RNN (2024)	84.9	Reddit	Yes	No
Proposed Model (2025)	87.8	Both	Yes	Yes

Figure 8: Benchmark Comparison (F1-Score)



5. Discussion

Based on the results, it can be clearly stated that enriching the transformer-based language models with sentiment-aware and contextual features significantly improves the performance of the sarcasm detection. An F1 of 87.8% on the combined data of Twitter and Reddit and other parameters suggest that automatically identifying sarcastic phrases is not just a matter of polarity of the expression, but involves a deeper understanding of the social connotation of the entire expression. These findings align with current trends in the literature as sarcasm is turning out to be a pragmatic phenomenon where interpretation extends beyond the simple level of semantics (Joshi et al., 2016).

Among the detected features, sentiment inversion rate is seen to be very important for sarcasm identification. Irony and sarcasm involve the reversal of meanings, where an overtly positive statement has a negative meaning or an overtly negative statement has a positive meaning; an aspect that goes unnoticed by most of the traditional models (Liu, Yang, Zhang, & Lu, 2014). By using the VADER method to extract two new features, polarity of sentiment, the system is able to identify polarity shifts. This corresponds to the modern trends in sarcasm analysis considering that the polarity and the tone are incompatible (Castro et al., 2022).

The conversational context also turned out to be very helpful as well when integrated. Sarcasm is context-dependent, meaning that the listener must know what has been said before in a conversation or some recent discussions (Ghosh et al., 2021). By incorporating the prior two turns into the model, we ensured that the transformer was aware of the broader conversation in which the post would be used and enriched the context in which the post was placed. This is in agreement with contextual theories of discourse where sarcasm is viewed as being highly context dependent regarding speaker intention as posited by Colston & Gibbs (2007).

Our model has shown a better performance than the models developed in earlier studies would when tested in single and multiple datasets. For example, Hossain et al. (2020) formulated and implemented a deep contextual attention network with a memory network that approximately resembled our model but performed quite modestly by modelling conversation history of Reddit threads but inadvertently omitted sentiment analysis. Similarly, the study of Zhang & Wang (2021), where the authors used graph neural networks for relation learning of user conversations, lose the emoji-dimension of sarcasm which leads to lower recall rates especially when spotting subtle sarcasm information. This work provides additional evidence suggesting that methods based on inter-user dependencies based on attention and graph models are insufficient for user participation prediction when not accompanied by affective analysis.

Apart from the aspect of accuracy, the proposed model has significant consequences for sentiment analysis workflows, content moderation, and social listening. Sarcasm in particular, as occurs in political discussion analysis, can be used to hide dissent or express a negative opinion (Suliman et al., 2023). Sarcastic tweets especially in political context can easily lead to frustration and therefore post analysis of sarcasm could potentially contaminate sentiment analysis results or categorize content in the wrong way in contexts where posts are classified for spreading misinformation. This way, making it more responsive to pragmatic and emotional context, we help advance more effective NLP tools for practical and important areas like governance, crisis management, and public health communication.

However, these are some of the limitations that are worthy to mention about this study. First, the model relies on sentiment analysis tools such as VADER, which adds potential source of error for the predictor, especially for posts which are categorized as neutral or contain idiomatic expressions that may be misidentified by sentiment lexicons (Smailović et al., 2013). Also, it is worth to note that both datasets of the Reddit and Twitter sources are only English-language data. Thus, the applicability of our model to multi-lingual or code-mixed sarcasm, often found in South Asian and Middle Eastern regions, remains unexplored. As observed by other authors earlier, sarcasm in code-switched contexts also has its structural and semantic features distinct from monolingual data (Mubarak et al., 2021).

Another disadvantage can be attributed to the rate at which the computations are done. As for the weakness, the inference time being 4.1 milliseconds per post may not be adequate for real-time tasks like chatbot moderation and live sentiment feeds. This has the need for further investigations on model compression methods like knowledge distillation or quantization which aims at maintaining the fidelity of models yet employing less computation (Jiao et al., 2020).

However, the model was not tested in adversarial conditions, when the users purposely obscure the sarcastic stance through the use of emoticons, informal language, or font changes. Another reason is that, with the increasing use of multi-modal communication where text is accompanied by images or GIFs, there's a need for sarcasm detectors that can work across modes. This opens up exciting avenues for future exploration into multimodal sarcasm detection frameworks (Schifanella et al., 2021).

Thus, incorporating sentiment and context into the transformer-based models is a promising direction in sarcasm detection; however, more investigations into the expansion of the model's language knowledge, its enhancements for creating efficiency, and the incorporation of multimodal and adversarial fine-tunes should be conducted. The results of our work are helpful for establishing in-depth, accurate, and practical sarcasm-detection models for future research.

REFERENCES:

Castro, S., Oliveira, H., & Santos, D. (2022). Detecting sarcasm in multimodal messages: Beyond text-based sentiment. Journal of Artificial Intelligence Research, 74, 905–930.

Colston, H. L., & Gibbs, R. W. (2007). Irony in language and thought: A cognitive science reader. Psychology Press.

Ghosh, D., Das, D., & Chakraborty, T. (2021). Context-aware sarcasm detection using hierarchical transformer networks. Proceedings of the International Joint Conference on Natural Language Processing.

Hossain, N., Dinkar, D., & Karimi, S. (2020). A deep contextual attention model for sarcasm detection in conversations. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 1166–1175.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., ... & Liu, Q. (2020). TinyBERT: Distilling BERT for natural language understanding. Proceedings of the Findings of ACL, 4163–4174.

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2016). Automatic sarcasm detection: A survey. ACM Computing Surveys (CSUR), 50(5), 1–22.

Liu, B., Hu, M., & Cheng, J. (2014). Opinion observer: Analyzing and comparing opinions on the web. Proceedings of the International World Wide Web Conference (WWW), 342–351.

Mubarak, H., Darwish, K., & Magdy, W. (2021). Data sets and methods for code-switched sarcasm detection. Proceedings of the Workshop on Computational Approaches to Linguistic Code-Switching, 12–21.

Schifanella, R., De Francisci Morales, G., & Aiello, L. M. (2021). Detecting sarcasm in social media: A multimodal approach. Information Processing & Management, 58(5), 102667.

Smailović, J., Kranjc, J., Grčar, M., & Mozetič, I. (2013). Monitoring the twitter sentiment during the political debates. In Proceedings of the European Conference on Information Retrieval (ECIR), 720–723.

Suliman, A., Ali, A., & Naseem, U. (2023). Sarcasm in political discourse: Detection using syntactic and semantic inconsistencies. Journal of Language and Politics, 22(2), 154–177.

Zhang, Y., & Wang, J. (2021). Graph-based sarcasm detection via modeling conversation context. Neurocomputing, 439, 63–74.

Al-Khatib, K., Wachsmuth, H., & Habernal, I. (2021). Multi-language sarcasm detection using cross-lingual transformer embeddings. Proceedings of the Workshop on Figurative Language Processing, 22–30.

Amir, S., Wallace, B. C., Lyu, H., Carvalho, P., & Silva, M. J. (2016). Modelling context with user embeddings for sarcasm detection in social media. Proceedings of the Conference on Computational Linguistics (COLING), 1675–1684.

Carvalho, P., Sarmento, L., Silva, M. J., & de Oliveira, E. (2009). Clues for detecting irony in user-generated content: Oh...!! it's so easy;-). Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion,53–56.

Farías, R. M., Rosso, P., & Montes-y-Gómez, M. (2021). Deep learning for sarcasm detection on Twitter. Information Processing & Management, 58(1), 102425. https://doi.org/10.1016/j.ipm.2020.102425

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,581–586.

Hernández-Fariñas, T., García-Serrano, A., & González, M. (2020). A sarcasm-aware computational model for polarity classification in Twitter messages. Expert Systems with Applications, 143, 113092. https://doi.org/10.1016/j.eswa.2019.113092

Joshi, A., Gupta, A., & Bhattacharyya, P. (2022). Code-mixed sarcasm detection using multilingual BERT. Proceedings of the Workshop on NLP for Code-Switching, 45–52.

Potamias, E., Siolas, G., & Stafylopatis, A. (2020). A transformer-based approach to irony and sarcasm detection. In Artificial Intelligence Applications and Innovations (pp. 313–325). Springer. https://doi.org/10.1007/978-3-030-49161-1_26

Rajadesingan, A., Zafarani, R., & Liu, H. (2015). Sarcasm detection on Twitter: A behavioral modeling approach. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM), 97–106. https://doi.org/10.1145/2684822.2685316

Reyes, A., Rosso, P., & Veale, T. (2012). A multidimensional approach for detecting irony in Twitter. Language Resources and Evaluation, 47(1), 239–268. https://doi.org/10.1007/s10579-012-9196-x

Sykora, M., Garthwaite, P., & Jackson, T. (2022). Enhancing sarcasm detection with sarcasm-focused lexicons and attention mechanisms. Journal of Intelligent Information Systems, 58(2), 319–340. https://doi.org/10.1007/s10844-021-00636-w

Wallace, B. C., Choe, D. K., Charniak, E., & Batra, D. (2015). Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. Proceedings of the ACL, 1035–1044.

Wu, Y., Zhao, Z., Wang, Y., & Sun, C. (2021). Irony detection with cross-domain transformer models and knowledge-enriched attention. Journal of Computational Linguistics, 47(4), 881–912. https://doi.org/10.1162/coli_a_00408

Zhang, X., Robinson, D., & Tepper, J. (2016). Detecting sarcasm in tweets: An ensemble approach. Journal of the Association for Information Science and Technology, 68(3), 626–638. https://doi.org/10.1002/asi.23605