SARCOVID: A framework for sarcasm detection in Tweets using hybrid transfer learning techniques

Balaji TK[1], Annushree Bablani[1], Sreeja SR[1], and Hemant Misra[2]

[1] Indian Institute of Information Technology Sri City
Andhra Pradesh 517646, INDIA
{balaji.tk,annushree.bablani,sreeja.sr}@iiits.in
[2] Swiggy, INDIA
misrahemant@gmail.com

Abstract. The COVID-19 pandemic sparked a surge in online discussions, making sentiment analysis challenging due to the prevalence of sarcasm on social media. Identifying sarcastic expressions within the context of COVID-19 conversations poses a unique linguistic hurdle. To tackle this challenge, a novel framework called SARCOVID is proposed that leverages hierarchical transfer learning and ensemble techniques to detect sarcasm in the field. Through rigorous evaluation on a collected COVID-19 dataset, SARCOVID demonstrates superior performance in identifying sarcastic content with reduced bias compared to traditional methods. The findings reveal a significant presence of sarcasm in online COVID-19 discussions, underscoring the importance of robust sarcasm detection techniques. In a test, the framework outperforms other models with 0.61 accuracy on Sarcasm corpus V2. This approach not only advances sentiment analysis capabilities for evolving online conversations but also provides deeper insights into the nuanced expressions of sentiment on social media.

Keywords: Sentiment analysis · Transfer learning · Deep learning · Opinion mining · Neural networks

1 Inroduction

Sentiment analysis is a crucial aspect of understanding public opinion and emotions, particularly during times of crisis such as the COVID-19 pandemic. The pandemic has led to a surge in discussions and opinions on various social media platforms, where people share their thoughts and feelings about the situation. As a result, sentiment analysis has become increasingly important for governments, businesses, and researchers to understand the public's concerns, reactions, and attitudes towards the pandemic.

However, sarcasm poses a significant challenge to sentiment analysis. Sarcasm, a form of irony, can completely change the meaning of a sentence, making it difficult for sentiment analysis models to classify the sentiment accurately. For

example, a sentence like "I'm so glad, I am not vaccinated" may seem positive at first glance, but it is actually sarcastic and expresses the opposite sentiment. Therefore, accurately detecting sarcasm is essential for improving the accuracy of sentiment analysis, especially in the context of COVID-19.

It is crucial to address sarcasm in sentiment analysis during the COVID-19 pandemic due to the potential impact on the accuracy of sentiment classification. Sarcasm, an irony often used in online communication, can alter the sentiment expressed in text, leading to misinterpretation by sentiment analysis models. In the context of the pandemic, where emotions and opinions are heightened, accurately capturing sentiments is essential for understanding public reactions, concerns, and attitudes towards COVID-19. Failure to detect sarcasm can result in misleading analyses and misrepresenting public sentiment, highlighting the significance of effectively dealing with sarcasm to ensure the reliability and precision of sentiment analysis during this critical period.

Numerous approaches are available for detecting sarcasm, transfer learning is one such effective method among them when a domain-specific dataset is limited. Transfer learning is a machine learning technique that allows models to learn from pre-trained models and adapt them to new tasks [1]. In the case of sarcasm detection, transfer learning can be used to train models on large datasets of sarcastic and non-sarcastic text, allowing the models to learn the patterns and nuances of sarcasm. This can lead to more accurate sarcasm detection, which in turn can improve the overall performance of sentiment analysis models. Addressing this challenge of detecting sarcasm in COVID-19 tweet sentiment analysis, a novel framework called SARCOVID is introduced. This framework leverages hierarchical transfer learning and ensemble methods to enhance the accuracy of sarcasm detection. The effectiveness of SARCOVID is evaluated using the SENSECOR dataset, which revealed the prevalence of sarcasm in COVID-19 discussions.

The subsequent sections unfold as Section 2, Section 3 explains the methodology of this study, section 4 discusses the results, and section 5 concludes the study.

2 Literature work

Several studies have explored sentiment analysis of COVID-19, revealing valuable insights into public opinion. Twitter data has become a popular resource for sentiment analysis [2]. In [3], authors explored public sentiment on Twitter in India during the early stages of the COVID-19 pandemic (December 2019 to May 2020). They used TextBlob to analyze the emotional tone (polarity) of tweets and NLTK to identify frequently used words. Their analysis, visualized by state and month, revealed some surprising findings. Despite the pandemic's challenges, the dominant sentiment among Indian Twitter users was positive. This positivity coincided with announcements of lockdowns, with higher tweet volumes coming from states hit hard by COVID-19. While there were negative tweets, the positive sentiment suggests a general trust in the government's response.

The study [4] evaluated various machine learning classifiers across different datasets for sentiment analysis of COVID-19-related Twitter data. Traditional methods like TF-IDF with SVM showed strong performance, with accuracy scores ranging from 0.829 to 0.845. Embedding-based models, particularly fastText, outperformed others due to their effective handling of out-of-vocabulary words. Deep learning approaches, such as using GloVe embeddings with deep convolutional neural networks (DCNN), demonstrated superiority over bidirectional long short-term memory (BiLSTM). Hybrid models like hybrid ranking outperformed IWV, emphasizing the importance of incorporating sentiment and context information. BERT stood out among transformer-based language models, surpassing all others with performance scores exceeding 0.85 across all datasets.

Authors in [5] investigated sentiment trends across eleven heavily affected countries during the pandemic. Analyzing over 50,000 tweets, the study revealed nuanced emotional responses, with some nations displaying predominantly positive sentiments while others showed a balance between positive and negative expressions. Emotion analysis highlighted shifts over time, from initial fear to growing trust as recovery rates improved. Utilizing the Syuzhet package in R, the research compared sentiment analysis algorithms to unveil the complex emotional dynamics amidst the global crisis.

Traditional sentiment analysis approaches primarily focus on surface-level sentiment in tweets, neglecting the crucial layer of sarcasm. This research introduces SARCOVID, a novel framework that handles sarcasm in COVID-19 tweets. SARCOVID leverages a hybrid approach, combining hierarchical transfer learning for improved knowledge transfer and an ensemble majority voting [6] technique to achieve more accurate sarcasm detection. By adjusting sentiment analysis based on identified sarcasm, SARCOVID aims to provide a more nuanced understanding of public opinion within COVID-19 discussions on social media.

3 Methodology

The SARCOVID framework development comprises the fusion of two methodologies. Initially, a hierarchical evaluation of tweets is conducted for sarcasm detection, employing transfer learning techniques in the first phase. Subsequently, in the second phase, the tweets undergo classification utilizing an ensemble majority voting technique. This dual approach ensures comprehensive analysis and robust sarcasm detection within COVID-19-related discourse on social media platforms.

3.1 Dataset

The datasets used for this study to train, evaluate and test the models are:

1. **News Headlines (26,709 samples) [7]:** This dataset is dedicated to identifying sarcasm in concise texts such as headlines, ensuring an equal distribution of sarcastic and non-sarcastic examples, denoted by 1 and 0 respectively.

2. **Reddit Sarcastic (1 million reviews) [8]:** With a vast collection of reviews, this dataset serves to analyze sarcasm within online discussions, with clearly labeled instances of sarcasm (1) and non-sarcasm (0).
3. **SemEval (5,735 samples) [9]:** Sourced from the iSarcasmEval GitHub repository, this dataset provides additional data with definitive yes/no labels for sarcasm.
4. **Twitter**[3] **(2,000 samples):** This dataset is instrumental in uncovering sarcasm within tweets, offering a balanced mix of both sarcastic and non-sarcastic instances.
5. **Sarcasm Corpus V2(9116 samples) [10]:**The Sarcasm Corpus comprises three balanced types of samples—Generic, Rhetorical Questions, and Hyperbole—each containing an equal number of sarcastic and non-sarcastic samples.

This study evaluates the proposed SARCOVID framework on the SENSECOR [11] dataset, which we previously collected for research on the COVID-19 Omicron variant [11]. The SENSECOR dataset comprises 160,000 tweets related to this specific variant.

3.2 Preprocessing

In the text preprocessing phase, a series of steps are implemented to enhance the data quality before the text given to the model to process. It involves

- Lowercasing all text
- Eliminating punctuation
- Tokenizing
- Removal of stopwords
- Stemming and lemmatization
- Removing special characters
- Handling Emoji and acronyms (if available)
- Removal of URLs, mentions and hashtags (if available)

3.3 Methodology implementation

The process involves four deep learning models (M1, M2, M3, and M4), each trained on individual sarcasm datasets (Reddit, News Headlines, SemEval, and Twitter). Each model is tested on the SENSECOR dataset for sarcasm identification to make the sentiment analysis free from biased opinions.

hierarchical evaluation

The hierarchical evaluation process outlined involves a multi-stage approach to sarcasm detection in COVID-19-related tweets using transfer learning

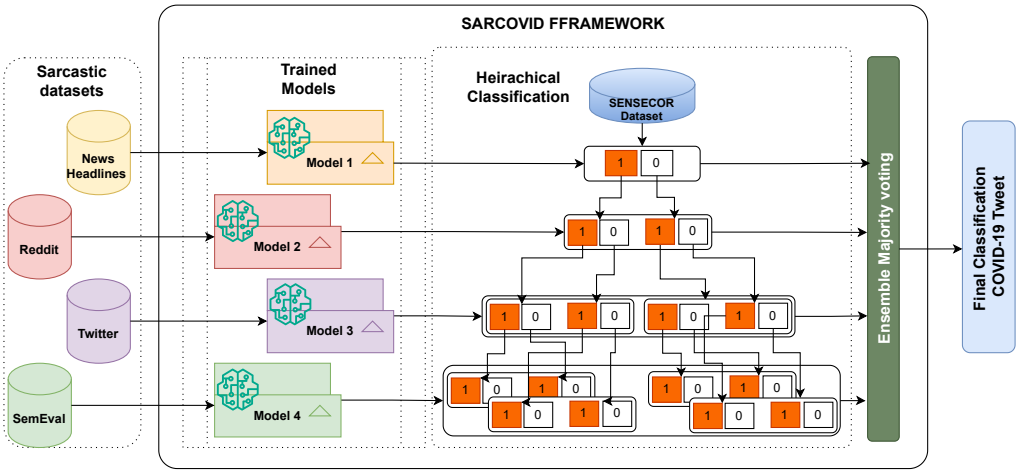


Fig. 1: Detailed architectural representation of SARCOVID framework.

techniques. The architectural representation of the proposed SARCOVID framework is presented in Fig. 1.

In the initial classification, model M1 is used to classify the tweets in the SENSECOR dataset, where each tweet is classified into two categories: those identified as sarcastic (Class 1) and non-sarcastic (Class 0).

In the subsequent evaluation phase, M2 conducts further analysis of the tweets based on M1's classifications. For tweets initially identified as sarcastic by M1, M2 evaluates them to either confirm or refute their sarcastic nature, resulting in two new categories: tweets confirmed as sarcastic and tweets reclassified as non-sarcastic. Similarly, for tweets initially labeled as non-sarcastic by M1, M2 evaluates them to detect any sarcastic content, leading to the creation of two additional categories: tweets flagged as sarcastic and tweets confirmed as non-sarcastic. Thus, the initial two results generated by M1 are now categorized into four.

This process continues with M3 and M4 analyzing the outputs from previous models, generating outputs at different levels for classifications based on M1, M2, M3, and M4. Consequently, M1 generates two outputs, M2 generates four outputs, and finally, M4 generates 16 outputs, totalling 30 outputs in a hierarchical fashion, comprising 15 for sarcastic and 15 for non-sarcastic tweets.

Ensemble majority Voting:

In the final evaluation phase, all model outputs are considered to determine the final classifications for each tweet. An ensemble majority voting approach

[3] https://github.com/surajr/SarcasmDetection/tree/master/Data

is then applied to assess the final classifications. Tweets that receive at least three votes(three model's votes) as sarcastic are categorized as sarcastic, while those receiving at least three votes(three model's votes) votes as non-sarcastic are categorized as non-sarcastic. Overall, this multi-stage hierarchical evaluation process leverages transfer learning techniques to effectively identify sarcasm in COVID-19-related tweets, allowing for a comprehensive analysis of sarcasm prevalence within the SENSECOR dataset.

Algorithm 1 Hierarchical Sarcasm Detection

Input: SENSECOR dataset with n tweets T_i where $i > 0$
Output: Sarcastic and non-sarcastic tweet classifications
Run M_1 on all n tweets ▷ Classification using M1
$C_1 \leftarrow$ Tweets classified as sarcastic by M_2
$C_2 \leftarrow$ Tweets classified as non-sarcastic by M_2

Run M_2 on all C_1 tweets ▷ Classification using M2
$C_3 \leftarrow$ Tweets classified as sarcastic by M_2
$C_4 \leftarrow$ Tweets classified as non-sarcastic by M_2
Run M_2 on all C_2 tweets
$C_5 \leftarrow$ Tweets classified as sarcastic by M_2
$C_6 \leftarrow$ Tweets classified as non-sarcastic by M_2
Continue the classification using M3, and M4 models... ▷ assign tweets to c7,c8,...c30.
sarcastic_votes $\leftarrow \{0\}^n$
non_sarcastic_votes $\leftarrow \{0\}^n$
sarcastic_tweets $\leftarrow \{\}$
non_sarcastic_tweets $\leftarrow \{\}$
for $i = 1$ to n **do**
    for $j = 1$ to 30 **do**
        if T_i in C_j **then** ▷ T_i is tweet in a dataset
            if $j\%2 = 1$ **then**
                sarcastic_votes $\leftarrow$ sarcastic_votes + 1
            else
                non_sarcastic_votes $\leftarrow$ non_sarcastic_votes + 1
            end if
        end if
    end for
    if sarcastic_votes ≥ 3 **then** ▷ Threshold for classifying as sarcastic
        sarcastic_tweets.append(T_i)
    else if non_sarcastic_votes ≥ 3 **then** Threshold for classifying as non-sarcastic
        non_sarcastic_tweets.append(T_i)
    end if
end for
return sarcastic_tweets, non_sarcastic_tweets

The SARCOVID framework methodology is presented in algorithm 1. The algorithm takes the SENSECOR dataset with n tweets as input and initializes lists to store the final sarcastic and non-sarcastic tweet classifications. Model M1 is run on all tweets, and the initial sarcastic (C1) and non-sarcastic (C2) classifications are obtained. Then, M2 runs on (C1) and classifies sarcastic tweets as (C3) and non-sarcastic tweets as (c4). Later, M2 runs on (C2) and classifies sarcastic tweets as (C5) and non-sarcastic tweets as (c6). The same procedure follows for M3 and M4 models, which classify tweets and assign them to C7, C8, ..., and C30. Here, Odd classes such as C1, C3, and C5... are holding Sarcastic tweets and even classes such as C2.C4, C6,... hold non-sarcastic tweets. After the hierarchical classification, the algorithm counts the number of sarcastic and non-sarcastic votes for each tweet based on the model outputs by iterating over the models and incrementing the respective vote counts for each tweet based on its classification. Finally, ensemble voting is applied: tweets with at least three sarcastic votes are added to the sarcastic_tweets list, and tweets with at least three non-sarcastic votes are added to the non_sarcastic_tweets list. The algorithm returns these two lists as the final sarcastic and non-sarcastic tweet classifications. The sample tweet classification of the SARCOVID framework is presented in Table 1.

Table 1: Sample COVID-19 Tweets detected in SENSECOR dataset

1. 2 jabs taken, but tested positive for COVID-19. Meanwhile, my granny, who hasn't received the vaccine, tested negative. Thank you #FireFauci #COVIDsucks
2. I love lockdowns - no food, no job, no nothing! #pandemic
3. Can anyone play a song on COVID-19? My neighbours are disturbing me with their noise. #COVID19 #lockdown

4 Results and discussion

The performance of four distinct models BiLSTM [12], BERT [13], RoBERTa [14] and, DistilBERT [15] is evaluated across four diverse datasets: News Headlines, IsarcasmEval, Twitter, and Reddit. The performance evaluation of models is presented in table 2. On the News Headlines dataset, BiLSTM achieved a moderate accuracy of 0.81, while DistilBERT and BERT surpassed it with accuracies of 0.83 and 0.91, respectively. RoBERTa exhibited the highest accuracy of 0.93, indicating its superior performance in capturing nuanced linguistic cues associated with sarcasm in news articles. Transitioning to the IsarcasmEval dataset, BiLSTM performed moderately with an accuracy of 0.74, whereas DistilBERT,

BERT, and RoBERTa showcased improved performance with accuracies of 0.758, 0.774 and 0.805, respectively, with RoBERTa achieving the highest accuracy. Moving to the Twitter dataset, BiLSTM achieved a reasonable accuracy of 0.87, while DistilBERT, BERT, and RoBERTa further improved upon this with accuracies of 0.884, 0.902, and 0.91, respectively. Lastly, on the Reddit dataset, BiLSTM demonstrated moderate performance with an accuracy of 0.7, whereas DistilBERT, BERT, and RoBERTa exhibited superior performance with accuracies of 0.767, 0.782, and 0.79, respectively. These results underscored the varying capabilities of each model across different datasets. They highlighted the effectiveness of advanced transformer-based models like RoBERTa, BERT, and DistilBERT in sarcasm detection across diverse linguistic contexts.

Table 2: Models performances on different datasets

Model	Dataset			
	News Headlines	IsarcasmEval	Twitter	Reddit
BiLSTM	0.81	0.74	0.87	0.7
DistilBERT	0.83	0.758	0.884	0.767
BERT	0.91	0.774	0.902	0.782
RoBERTa	0.93	0.805	0.91	0.79

The SARCOVID framework incorporates the RoBERTa model as a key component in its construction. The RoBERTa model is selected for its superior performance compared to all other baseline methods examined in this study. Hence, Model M1 is RoBERTa trained on the News Headlines dataset, while Model M2 is RoBERTa trained on the Reddit Sarcastic dataset. Models M3 and M4 are RoBERTa trained on the SemEval and Twitter datasets, respectively. The proposed approach leverages transfer learning due to its adaptability to different domains. Therefore, the SARCOVID framework must be evaluated on a new labelled sarcastic benchmark dataset, which is not used for training these models. For this purpose, the evaluation of the proposed SARCOVID framework is conducted on the Sarcasm Corpus V2, utilizing 2000 randomly selected samples to assess the effectiveness of the transfer learning technique. The primary objective of the study is to detect sarcasm in COVID-19 tweets, which is accomplished using the SENSECOR dataset, a large unlabeled corpus. The analysis and findings are presented in the results section.

The performance evaluation of various models on the Sarcasm Corpus V2 dataset, as depicted in table 3, quantifies their effectiveness in sarcasm detection. BiLSTM, a traditional recurrent neural network model, achieved the lowest performance with a precision, recall, F1 score, and accuracy all at 0.50. DistilBERT exhibited a slight improvement, with a precision and recall of 0.54, an F1 score of 0.51, and an accuracy of 0.55. BERT further improved with a precision of 0.55, recall of 0.56, F1 score of 0.56, and accuracy of 0.561. RoBERTa outperformed the previous models with a precision of 0.57, recall of 0.55, F1

score of 0.59, and accuracy of 0.58. The SARCOVID model achieved the highest performance metrics, with a precision of 0.6, recall of 0.62, F1 score of 0.61, and accuracy of 0.61, highlighting its effectiveness in detecting sarcasm. Notably, the proposed SARCOVID framework outperformed all other models, boasting the highest accuracy of 0.61. This superior performance of SARCOVID underscores its efficacy in sarcasm detection, indicating its potential to advance the field of natural language understanding and sentiment analysis. The ROC curve of the RoBERTa model using various datasets is presented in Fig. 2.

Table 3: Test analysis of models performed on Sarcasm Corpus V2 dataset.

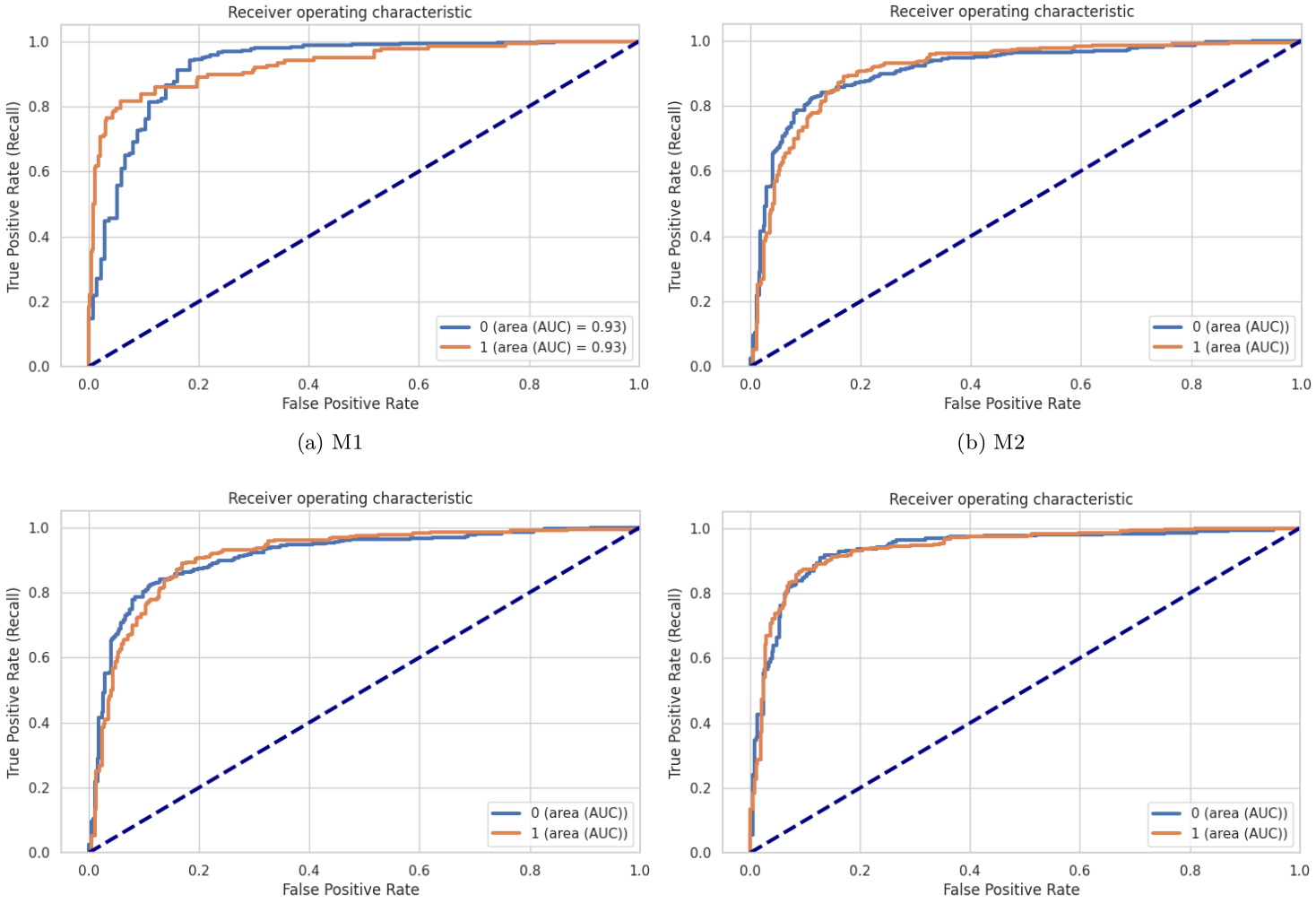
Model	**Precision**	**Recall**	**F1**	**Accuracy**
BiLSTM	0.49	0.50	0.50	0.5
DistilBERT	0.54	0.54	0.51	0.55
BERT	0.55	0.56	0.56	0.561
RoBERTa	0.57	0.55	0.59	0.58
SARCOVID	**0.6**	**0.62**	**0.61**	**0.61**

The performance of sarcasm detection models M1 to M4 on the SENSECOR dataset of 160,000 tweets is presented in Table 4. The effectiveness of each model is expressed as a percentage, indicating the proportion of tweets it classified as containing sarcasm. A lower percentage implies a more conservative approach, where only tweets exhibiting clear sarcastic cues specific to the SENSECOR dataset are identified as sarcastic. This selective identification helps reduce false positives, thereby improving the precision of sarcasm detection. Conversely, a higher percentage suggests a more liberal approach, increasing the risk of misclassifying non-sarcastic tweets as sarcastic, which could lead to lower accuracy in identifying truly sarcastic content.

Table 4: Sarcastic tweets detected in the SENSECOR dataset by models used in this study.

Models	**sarcasm detected (in %)**
M1	47.2
M2	46
M3	39
M4	48
SARCOVID	**24**

Among the evaluated models, Model M1 detected sarcasm in 47.2% of tweets, closely followed by Model M4, which identified sarcasm in 48% of tweets. Meanwhile, Models M2 and M3 exhibited lower detection rates, flagging sarcasm in



(a) M1 (b) M2

(c) M3 (d) M4

Fig. 2: ROC curves for RoBERTa model on various datasets.

46% and 39% of tweets, respectively. Interestingly, the SARCOVID framework showcased the most discerning performance, with a detection rate of only 24%. This suggests that SARCOVID employs a rigorous approach to sarcasm detection, prioritizing accuracy by minimizing false positives and ensuring precise identification of sarcastic content within the SENSECOR dataset.

The SARCOVID framework offers several advantages in its methodology for sarcasm detection in COVID-19 tweets. One notable advantage is its hierarchical evaluation approach, which allows for a comprehensive analysis by leveraging multiple deep-learning models trained on diverse sarcasm datasets. Most of the existing sarcasm detection techniques work on domain-specific labelled datasets to overcome false positive classification. This technique has a better approach to reducing the false positives without having a specific dataset. However, this methodology may present challenges regarding computational resources required for training and evaluating multiple models iteratively, as well as potential complexities in interpreting conflicting model outputs. Additionally, the reliance on pre-existing sarcasm datasets for transfer learning may introduce biases or limitations in detecting sarcasm in COVID-19 tweets.

5 Conclusion

Detecting sarcasm in COVID-19 tweets presents a significant challenge due to the scarcity of domain-specific datasets for model training. To address this issue, a novel framework called SARCOVID is proposed. SARCOVID employs a hybrid approach, combining hierarchical transfer learning and ensemble majority voting. The SARCOVID achieves high confidence in identifying sarcasm within the SENSECOR dataset, a collection of COVID-19 tweets. This framework exhibits a lower tendency for false positives in sarcasm detection, making it a robust solution for overcoming limitations in dataset availability for COVID-19 tweets. The model can be adapted to any application with limited dataset availability. A potential future direction could involve exploring the integration of additional modalities like visual data and employing continual learning techniques that could further enhance SARCOVID's ability to adapt and evolve with emerging linguistic patterns in this ever-changing domain.

References

1. C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pp. 270–279, Springer, 2018.
2. T. Ahmad, A. Ramsay, and H. Ahmed, "Detecting emotions in english and arabic tweets," *Information*, vol. 10, no. 3, p. 98, 2019.
3. T. Vijay, A. Chawla, B. Dhanka, and P. Karmakar, "Sentiment analysis on covid-19 twitter data," in *2020 5th IEEE international conference on recent advances and innovations in engineering (ICRAIE)*, pp. 1–7, IEEE, 2020.

4. U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis," *IEEE transactions on computational social systems*, vol. 8, no. 4, pp. 1003–1015, 2021.
5. M. A. Kausar, A. Soosaimanickam, and M. Nasar, "Public sentiment analysis on twitter data during covid-19 outbreak," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021.
6. L. I. Kuncheva and J. J. Rodríguez, "A weighted voting framework for classifiers ensembles," *Knowledge and information systems*, vol. 38, pp. 259–275, 2014.
7. R. Misra and P. Arora, "Sarcasm detection using news headlines dataset," *AI Open*, vol. 4, pp. 13–18, 2023.
8. "A large self-annotated corpus for sarcasm." 2017.
9. I. Abu Farha, S. V. Oprea, S. Wilson, and W. Magdy, "SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, (Seattle, United States), pp. 802–814, Association for Computational Linguistics, July 2022.
10. S. Oraby, V. Harrison, L. Reed, E. Hernandez, E. Riloff, and M. Walker, "Creating and characterizing a diverse corpus of sarcasm in dialogue," *arXiv preprint arXiv:1709.05404*, 2017.
11. T. Balaji, A. Bablani, S. Sreeja, and H. Misra, "Sensecor: A framework for covid-19 variants severity classification and symptoms detection," *Evolving Systems*, vol. 15, no. 1, pp. 65–82, 2024.
12. A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.
13. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
14. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
15. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.