Unmasking Sarcasm in the COVID-19 Twittersphere

Xingchen Li†
xingchen.li@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Geyu Shen†
gshen35@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Eric Yeon†
syeon31@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

With the advancement in technology, social media platforms have become the battleground for the dissemination of information. The sharing of information has slowly transitioned from traditional media to popular social media apps like Twitter. Twitter in particular has become a depository for real-time updates, opinions, and news. Thus, the authenticity of information especially in the Twitter ecosystem continues to be an obstacle in the battle for truth and accuracy where users are allowed to freely express their opinions. Our study focuses on sarcasm detection on COVID-19-related tweets and its subgroups [8]. We experimented with both TextGCN and GAT as our sarcasm detection classifier to effectively identify sarcastic content in tweets based on a 5,000-entry hand-labeled Twitter dataset. The classifier will then be trained and analyzed on a Twitter dataset that comprises of 2,400 COVID-19-related tweets, specifically about misinformation about 5G and COVID-19. Overall, the results show that GCN and GAT performed similarly in terms of F1 score (GCN: 0.687, GAT: 0.691). Yet, it is noticeable that GAT required less time for training (GCN: 2305 seconds, GAT: 1644 seconds). Furthermore, we delve into the trends, patterns, and relationships that emerge from the results to shed light on the dynamics of sarcasm in the context of a global health crisis. When analyzing sarcasm in misinformation tweets and in counter-misinformation tweets, we found the sarcasm-to-non-sarcasm ratios are very similar for both groups. However, for both groups, sarcastic tweets have more positive sentiments than non-sarcastic Tweets. In addition, although previous findings revealed that misinformation tweets increased as the crisis progressed, further analysis showed that a quarter of these were sarcastic and did not aid in the battle against misinformation. These insights discovered can help facilitate further research on the development of new tools and methods to combat the fight against global health misinformation.

ACM Reference Format:
Xingchen Li, Geyu Shen, and Eric Yeon. 2023. Unmasking Sarcasm in the COVID-19 Twittersphere. In *Proceedings of CSE 8803 DSN*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

CSE 8803 DSN, Sept 27, 2023, Atlanta, GA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

1.1 Aim

Our paper focuses on the intersection of sarcasm detection and misinformation identification on COVID-19 tweets, shedding light on the significance of developing detection models capable of discerning sarcastic intent within the vast repository of tweets. The objective is clear: to unveil the subtle patterns of tweets with sarcastic intent and to utilize this understanding as a tool to analyze trends containing misinformation. By doing so, we aim to mitigate the harmful consequences of false information and provide a more accurate representation of facts. Performing our experiments on an existing Twitter dataset with an improved sarcasm detection model will allow us to find additional patterns and criteria within the various subgroups of tweets that exist. The purpose of this paper is to answer the main research question:

How can we refine and advance current sarcasm detection models in tweets related to COVID-19 misinformation with improved accuracy?

Propagating from this main objective, we look to answer the following sub-issues:

1. What features and techniques can be utilized to improve sarcasm detection in the context of COVID-19?
2. What trends and patterns emerge in sarcastic tweets related to COVID-19, and what insights can be found from their analysis?

1.2 Challenges

Sarcasm is a form of communication where the intended meaning of the message is the opposite of the literal words used. It is often used to mock or criticize through a certain tone or context with a humorous periphery. Therefore, it is more difficult to identify when written down due to the lack of human intervention and context. Popular NLP tasks incorporate lexicon-based approaches, rule-based approaches, and supervised learning models to tackle this domain. Using a deep recurrent neural network to identify sarcasm within tweets has proven to be robust because it does not require manual feature engineering and additional methods such as sentiment lexicons [4]. Additional methods include a Bayesian RNN to capture the uncertainty of a black box neural network and provide more reliable probabilistic predictions [5]. Sarcastic remarks are extremely context-dependent, especially in text-formats and a BNN offers the capability to give an uncertainty estimate. Moreover, this framework allows for more robust outputs with limited data which can save computational resources and time.

In theory, a full BNN is quite costly to account for all the uncertainties for large numbers of successive layers. Therefore, in practice, the model should use a few probabilistic layers purposefully positioned in the network. However, its computational cost creates difficulties in implementation and its feasibility is questioned. A supplemental approach is a hybrid network with L total layers that consist of $L - P$ deterministic layers followed by P probabilistic layers. The few probabilistic layers are used for uncertainty estimation which comes from the nature of incorporating the Bayesian logic while the first $L - P$ layers are for representation learning. Ideally, no pre-training is required and a single flow calculation is all that is needed for the trained network [3].

1.3 Dataset

We utilized two datasets: one is a 5,000-entry Twitter dataset with hand-labeled sarcasm label, and the other one is a 2,400-entry 5G-COVID-19 Twitter misinformation dataset with hand-labeled misinformation label (misinformation, counter-misinformation, irrelevant). Additionally, we annotated 100 tweets in the second dataset with sarcasm labels (0.759 Cohen's Kappa Agreement Score). We trained both GAT and GCN models based on the first dataset and a small portion of the second one for generalization. After the training, we used the trained model to generate sarcasm labels for the entire 5G-COVID dataset and performed further analysis based on it.

1.4 Models and Performance

We experimented both TextGCN and GAT as our sarcasm detection models. We first generated a heterogeneous graph based on the entire dataset. Then, we applied TextGCN and GAT models to such graph structures for training. The result shows that both models performed similarly in terms of precision, recall, and F1 score, but GAT model requires less training time (Exact results are in Table 3). When compared to baseline models, logistic regression and DistilBERT, our models outperformed logistic regression in the F1-score and DistilBERT in their running time (Exact results are in Table 2).

1.5 Impact

Even though the model's performance is similar to our baseline models, to the best of our knowledge, our research is the first to explore the possibility of graph methods in sarcasm detection. We believe that the performance would be further improved with better computing resources and more attention to parameter tuning. Additionally, our analysis of sarcasm in COVID-19 misinformation discovers many interesting aspects both categorically and temporally. Such findings will help misinformation detection with sarcasm as a new facilitating indicator for future use.

2 LITERATURE SURVEY

There have been studies focusing on sarcasm detection, and they could be roughly classified into two categories: feature-based methods and contextualization-based methods.

Feature-Based: These models mostly rely on extracting linguistic cues and training the detection algorithm based on features such as sentiment, syntactic patterns, and other textual clues. For example, a linguistic-pattern-based model was proposed for sarcasm detection [1]. For this model, researchers focused on four kinds of features: sentiment-related features, punctuation-related features, syntactic and semantic features, and pattern-related features. The key idea is to use tags (Part-of-Speech tags) to extract the underlying patterns of Tweet sarcasm. Based on the accuracy performance of those four features, it is found that pattern-related features achieved the highest score. A major drawback of this approach is its limitation on utilizing auxiliary information, such as user engagement and authors' background information. Additionally, it is unclear if this model could be applied to more specific topics since its dataset is only Tweets with hashtag *sarcasm*. Yet, this study is highly related to our project in the way that a major part of the project is to detect sarcasm based on the content of Tweets. Thus, we could possibly use the innovative pattern-based feature extraction as our backbone model and add adjustments to accommodate the COVID-19 datasets.

Contextualization-Based: This kind of model utilizes features other than purely textual features. For example, the model proposed in [2] also investigates information like author features, audience features, and other environmental features. By incorporating more information into the model, it is shown that the accuracy gradually increases. This approach is similar to the way that researchers explore auxiliary and graph information in the space of misinformation detection. Yet, this model still has things that could be improved upon. For example, it only investigated the one-to-one relationship between the author and comment users, but it has shown in the field of misinformation detection that a network-based analysis could bring more underlying features into our sight. Therefore, constructing a network-based model is also an option for us on our project, but this model is a good baseline for incorporating additional information into the model. Additionally, it is also possible for us to explore multimodal approaches by analyzing images in the Tweet since a large portion of sarcasm is reflected in the form of memes or other visual cues.

Another major part of our project is to apply our trained sarcasm detection model in the context of misinformation. We will utilize the misinformation model in [9] and do a sarcasm analysis of its resulting categories: misinformation, professional fact-checks, opinion-based citizen responses, and evidence-based citizen responses. The paper itself has already done certain analyses on these categories, such as sentiment and politeness, but very little research, including this one, focused on sarcasm in social crises like COVID-19. Therefore, we hope that our work could expand upon this study and investigate the sarcasm trend among different groups and the temporal behavior of people's sarcasm as a whole society when facing social crises.

Finally, it is worth noting that there exists another paper focusing on COVID-19 Tweet sarcasm analysis [7]. In this paper, researchers hand-labeled 27,547 Tweets and trained three models (libSVM, Naive Bayes, and Decision Tree) entirely based on textual features. These three simple models performed very well in their dataset, and their study is highly related to our project. Yet, there exist so many things that we could improve upon. For example, their dataset and models are not public, and it is hard to see the reason why these simple models achieved such good performance. Therefore, after building our model, which possibly includes a lot

Table 1: Comparison of Sarcastic and Non-Sarcastic Tweets

Metrics	Sarcastic	Non-sarcastic
Sentence per Tweet	1.6476	2.098
Word per Sentence	24.3492	30.3016
Sentiment Score	0.0079	0.0160
Reading Ease	61.22	69.58

more additional information/representation, we would compare our model to this study's models as baselines and analyze sarcasm detection in the context of COVID-19 specifically.

3 DATASET DESCRIPTION AND ANALYSIS

For datasets, we utilized two datasets: one is a 5,000-entry Twitter dataset [6] with a hand-labeled sarcasm label and the other one is a 2,400-entry 5G-COVID-19 Twitter misinformation dataset [8] with a hand-labeled misinformation label (misinformation, counter misinformation, irrelevant).

3.1 Sarcasm Dataset

We utilized a sarcasm dataset from another previous research[6]. The dataset contains 5,000 entries, each with tweet content, hand-labeled sarcasm label, and some other information, such as its context. The dataset is evenly split between sarcastic tweets and non-sarcastic tweets. We chose this dataset as our sarcasm model training dataset because it shares the most similarity with the following COVID-19 misinformation dataset: both are tweet-based and share common attributes too.

For dataset cleaning and preprocessing, we followed a rather standard procedure: removing HTML, applying existing stop word methods, and embedding through tokenization with batch encoding.

After cleaning the dataset, we performed some preliminary analysis on it. First, sarcastic and non-sarcastic tweets have similar average sentences per tweet and average words per sentence. Additionally, the sentiment analysis reveals shocking similarities between the two groups. One key difference is their reading ease. Sarcastic tweets are more difficult to read and comprehend than non-sarcastic tweets on average.

3.2 COVID-19 Dataset

In addition to the sarcasm datasets in which the model will be trained, the **COVID-19 Tweet Dataset** [8] will be shed light on to investigate sarcasm contents within COVID-19-related tweets. The dataset is established on a keyword-based method based on a misconception dataset, the IFCN dataset, by leveraging the work of fact-checkers. Keywords considered in the dataset include COVID-19 keywords and Misconception entity keywords, and a manually labeled professional fact-checking tweet dataset is standing by as ground truths.

The dataset contains 2,400 hand-labeled misinformation entries, which are evenly split among misinformation, counter-misinformation, and irrelevant categories. We performed similar preprocessing steps to Sarcasm dataset. The analysis reveals that counter-misinformation tweets show more negative sentiment than misinformation tweet on average. Exact Results are in Figure 5.

Additionally, for the purpose of testing the sarcasm model's performance on the COVID-19 dataset, we also hand-labeled 100 tweets from this dataset and annotated sarcasm labels to it. Overall, we reached 0.759 Cohen's Kappa Agreement Score, indicating substantial agreement among the annotators. For disagreement entries, we discussed the different opinions and labeled them with the final decision.

4 EXPERIMENT SETTING AND BASELINES

The proposed task can be designated into two parts: train the model on sarcasm twitter and use the trained model to predict the sarcasm contents in COVID-19 related tweets. To accomplish these sub-tasks, a classifier for sarcasm contents on twitter need to be developed first.

Existing classifiers based on word2vec can be categorized into Bag-of-Word (BoG) model and transformer model. Herein, logistic regression is taken as a instance for BoG model and Bidirectional Encoder Representations from Transformers (BERT) is taken as a instance for transformer model. This work takes these two models as the baselines for the sarcasm classifier task.

To solidify the utilization of logistic regression, Fig. 1 shows a sarcasm/non-sarcasm score based on the prepared sarcasm tweet dataset. In the distribution, sarcasm and non-sarcasm contents show a degree of separation, suggesting a potential linear distinguishable feature. Even though overlapping exist in the displayed features, logistic regression still can perform as classifier for such a corpus. The major tuned hyper-parameter in logistic regression training is regularization parameter λ.

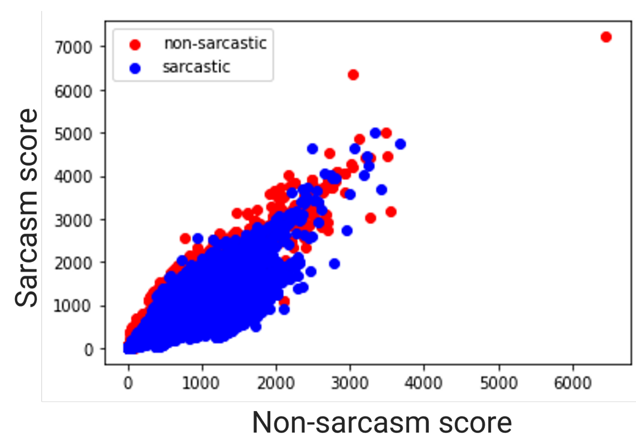

Figure 1: Sarcasm feature scores of data points in Sarcasm Tweets Dataset.

On the other hand, a BERT model is consisted of embedding (converts an array of one-hot encoded tokens into an array of vectors representing the tokens), stack of encoders (Transformer encoders to perform transformations over the array of representation vectors) and un-embedding (converts the final representation vectors into one-hot encoded tokens again). Since pre-trained BERT can be

Table 2: Performance Comparison of Baseline Models

Metrics	Precision	Recall	F1-score	Runtime
Logistic Regression	0.72	0.54	0.62	355 s
DistilBERT	0.83	0.82	0.82	33137 s

directly applied to the training task, no special concerns need to be considered and the major tuned hyper-parameters are Learning rate, epoch, batch size.

A constructing process including data processing, embedding and train/test set splitting is conducted before training. Html, bracket and stop words removal are applied, followed by embedding through tokenizer. WordPiece Tokenizer is utilized and truncating/padding are applied to finish the embedding process. For the dataset with 5,000 datapoint, both trainings take the same randomly selected 3750 pieces of data as training set and the rest 1,250 pieces as testing set.

In practice, the package of logistic regression in scikit-learn and distilBERT in TensorFlow are used for the training task. For the logistic regression, optimized result happens when $\lambda = 0.789$, yielding a precision of 0.72 and F1-score of 0.62. For the BERT, since distilBERT has 40% less parameters than bert-base-uncased and runs 60% faster while preserving over 95% of Bert's performances, only 3 epoches with batch size of 32 are enough to push the precision to 0.83 and F1-score to 0.82. Table 2 summarizes the performance comparison of these baseline models.

There are several conclusions can be derived from the baseline model comparisons. BoW model performs worse than Transformer model. The sarcasm corpus has huge overlapping in terms of sarcasm/non-sarcasm, bringing under capability for linear classifiers. Conversely, BERT learns contextual relations between words (or sub-words) in a text and is capable of understand the meaning of ambiguous language in text by using surrounding text to establish context. Based on the highly contextual nature of sarcasm, BERT stands out because of its content dependence.

5 PROPOSED METHOD

Two baseline models discussed in the last section have already shown some capability of being proper classifier as sarcasm contents. However, when facing the second RQ we proposed as applying model to COVID 19 Dataset and analyze trends, these two models might be obsolete on carrying such tasks. A model that can generate embedding within a larger corpus including both sarcasm and covid-19 tweets, and show the correlation between each datapoint is essential. In that case, a graph based model would be more suitable for such a task.

In this work, we propose a heterogeneous graph generation for dataset, and use GNN-based classifier for training and prediction. To accompolish this, we firstly combine labelled sarcasm data and covid-19 data together to form a larger dataset. Then, the covid-19 portion was shuffled and 100 pieces of tweets were picked on a random basis. These picked ones are manually labeled with sarcasm tags and passed the Cohen's Kappa Agreement test. After the processing, the whole dataset is consisted of three types of data: 1. Labeled sarcasm tweets, which will be used for training; 2. Covid-19 tweets with hand-labeled sarcasm tags, which will be used for testing; 3. Covid-19 tweets without sarcasm tags, which will be predicted with sarcasm tags later.

Table 3: Performance Comparison of GNNs

Model	Precision	Recall	F1-score	Runtime
GCN	0.688	0.687	0.687	2305 s
GAT	0.692	0.691	0.691	1644 s

With the prepared dataset, a transformation to heterogeneous graph was conducted. The generated graph structure is shown in Fig. 2, in which T represents document nodes and W represents word nodes. Word nodes were generated based on word co-occurrence with context windows, and document-word edges were built upon doc-word frequency. PMI (Pointwise mutual information) was used as weights. All 7,488 texts are embedded within the same graph simultaneously.

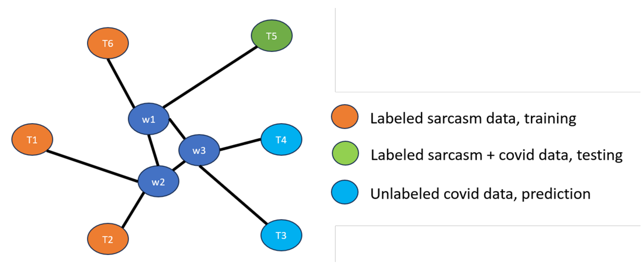

Figure 2: Applied graph consisting three different types of data nodes.

Two options of graphical models can be considered for training, which are TextGCN (text graphical neural network) [11] and GAT (graphical attention network) [10]. TextGCN is a graph convolution operation produces the normalized sum of the node features of neighbors, while GAT introduces the attention mechanism as a substitute for the statically normalized convolution. This work conducted both model for the training and predicting task, while using only one hidden layer. For the training process, our work took the 5,000 pieces of labeled sarcasm data, with 3515 used for training (in which 450 used for validation) and 1,485 used for testing.The training was supported by PyTorch and Dgl packages. Convolution layers defines GCN and GAT in the training with SimpleConv and MultiHeadGATLayer, respectively. Hyper parameter selections are hidden layer dimension of 800, learning rate of 0.2, epoch of 100 and early stopping of 20.

Result of training on sarcasm labelled data is shown in Table 3. It can be seen that GAT shows slight outperform comparing with GCN. The relatively low precision and F1-score is caused by time limitation of fine-tuning, and can be improved by changing hyperparameters like network hidden layer numbers and dimensions.

6 EXPERIMENTS AND RESULTS

To recall our RQ, we aim to find the trending of sarcasm within covid-19 tweets. In order to proceed the research, prediction accuracy on covid-19 dataset with pre-trained sarcasm classification

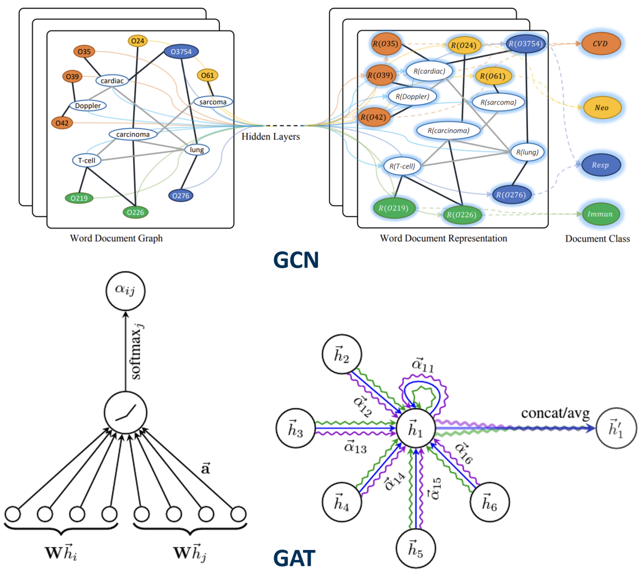

Figure 3: GNN structure applied in training. [11], [10]

Table 4: Prediction of pre-trained GNNs

Model	Precision	Recall	F1-score
pre-trained GCN	0.522	0.522	0.519
pre-trained GAT	0.601	0.601	0.600

model need to be tested. Followed by all 5,000 pieces of labeled sarcasm data are feed into each model, pre-trained GCN and GAT shown in Section 5 are applied on the picked 100 piece of covid-19 data with sarcasm labels for testing the prediction accuracy. Table 4 presents the testing results, in which indicates that pre-trained GAT performs better than pre-trained GCN, thus suggesting the selection of GAT for label prediction for unlabeled covid data.

The trained model was then used to analyze trends or patterns that existed during the COVID-19 pandemic between January 2020 to June 2020 on the hand-labeled dataset. Figure 4 visualizes the count of sarcastic as well as non-sarcastic tweets and the irrelevant tweets were removed for its inapplicability and unimportance. This chart shows that the sarcasm-to-non-sarcasm ratio is similar for both groups (misinformation and counter-misinformation) with a larger portion of tweets being sarcastic. Despite the consistent ratio, a noteworthy observation emerged in the form of positive sentiment dominance within sarcastic tweets across both categories (Figure 5). However, this is expected based off psychological and communicative factors since sarcasm provides individuals with a coping mechanism especially for distressing topics such as Covid-19. The pandemic's inherently negative subject matter induces individuals to express their frustrations or fears in a more palatable manner. Moreover, a negative sarcastic tweet would imply the user is conveying a positive message about the crisis which during the time was absent.

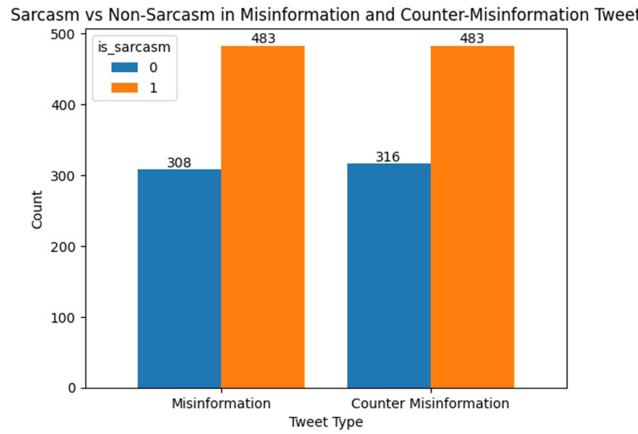

Figure 4: Sarcastic vs Non-Sarcastic Chart

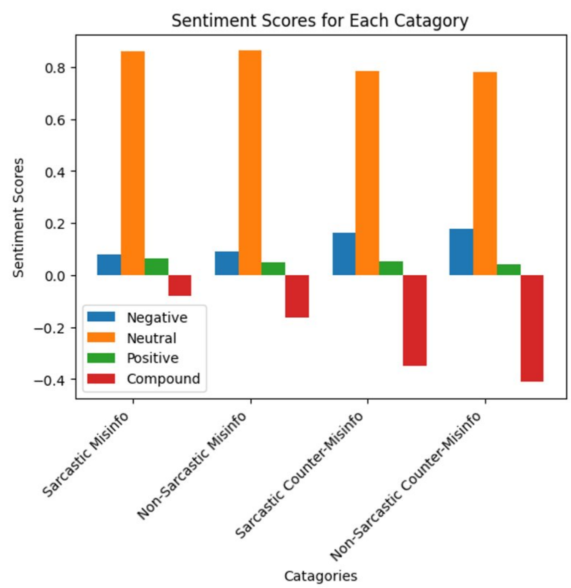

Figure 5: Sentiment Scores

Temporal examination on the predicted labels yield one compelling insight into societal discourse over time. In the month of April, the count of counter-misinformation tweets surpassed the number of misinformation and irrelevant tweets combined. Figure 6 further shows that the the number of sarcastic tweets were the highest in April as well. The findings revealed that the misinformation data did not show the entire reality: not all counter-misinformation-labeled tweets contributed to the fight against inaccurate information. A quarter of misinformation tweets in that month were labeled sarcastic indicating an inflated trend in diminishing misinformation.

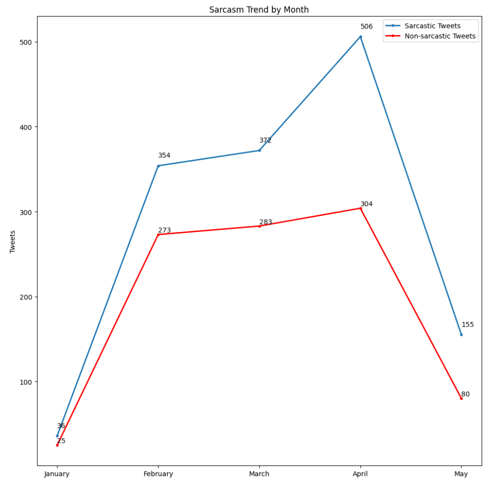

Figure 6: Sarcasm Trend by Month

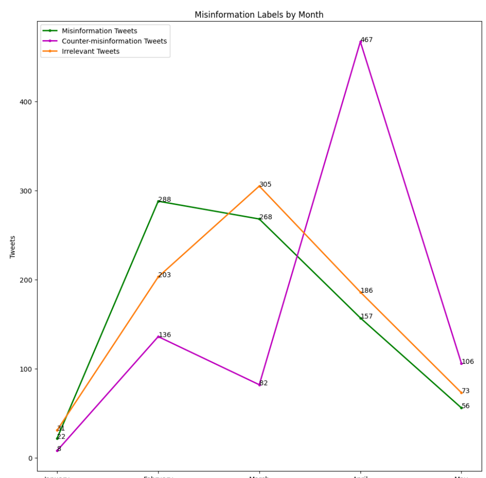

Figure 7: Misinformation Labels by Month

7 CONCLUSION

In conclusion, our study delves into the intricate intersection of sarcasm detection and misinformation identification within the realm of COVID-19-related tweets on Twitter. Leveraging advanced graph neural network models, TextGCN and GAT, we aimed to enhance sarcasm detection accuracy and shed light on nuanced patterns within the vast Twitter ecosystem. The results demonstrated comparable performance between GCN and GAT, with GAT exhibiting a faster training time. Notably, our analysis revealed consistent sarcasm-to-non-sarcasm ratios across misinformation and counter-misinformation tweets, emphasizing the pervasive nature of sarcasm. However, the positive sentiment prevalent in sarcastic tweets poses challenges in distinguishing their impact on the battle against misinformation. As we explored temporal trends and societal dynamics, our findings contribute valuable insights for refining detection models and devising strategies to combat global health misinformation. By addressing the challenges posed by sarcasm, our research advances the understanding of communication nuances on social media platforms, paving the way for more effective tools and methods in the ongoing fight for information accuracy.

8 CONTRIBUTIONS

All team members have contributed a similar amount of effort.

ACKNOWLEDGMENTS

REFERENCES

[1] David Bamman and Noah Smith. 2021. Contextualized Sarcasm Detection on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 1 (Aug. 2021), 574–577. https://doi.org/10.1609/icwsm.v9i1.14655

[2] Mondher Bouazizi and Tomoaki Otsuki Ohtsuki. 2016. A Pattern-Based Approach for Sarcasm Detection on Twitter. *IEEE Access* 4 (2016), 5477–5488. https://doi.org/10.1109/ACCESS.2016.2594194

[3] Daniel T. Chang. 2021. Hybrid Bayesian Neural Networks with Functional Probabilistic Layers. arXiv:2107.07014 [cs.LG]

[4] Christopher Eke, Azah Norman, Liyana Shuib, and Henry Nweke. 2020. Sarcasm identification in textual data: systematic review, research challenges and open directions. *Artificial Intelligence Review* 53 (08 2020). https://doi.org/10.1007/s10462-019-09791-8

[5] Meire Fortunato, Charles Blundell, and Oriol Vinyals. 2019. Bayesian Recurrent Neural Networks. arXiv:1704.02798 [cs.LG]

[6] Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A Report on the 2020 Sarcasm Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*, Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee, Anna Feldman, and Debanjan Ghosh (Eds.). Association for Computational Linguistics, Online, 1–11. https://doi.org/10.18653/v1/2020.figlang-1.1

[7] Raju Kumar and Aruna Bhat. 2021. An Analysis On Sarcasm Detection Over Twitter During COVID-19. In *2021 2nd International Conference for Emerging Technology (INCET)*. 1–6. https://doi.org/10.1109/INCET51464.2021.9456392

[8] Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020. The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. In *2020 IEEE International Conference on Big Data (Big Data)*. 748–757. https://doi.org/10.1109/BigData50022.2020.9377956

[9] Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020. The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. In *2020 IEEE International Conference on Big Data (Big Data)*. 748–757. https://doi.org/10.1109/BigData50022.2020.9377956

[10] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. 2017. Graph attention networks. *stat* 1050, 20 (2017), 10–48550.

[11] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 7370–7377.

Received 26 September 2023