Detecting COVID-19 Misinformation Using Natural Language Processing

Peter Chika Ozo-Ogueji
Department of Mathematics and Statistics
American University
Washington, DC, USA
po3783a@american.edu

Pham Hoang
Department of Computer Science
American University
Washington, DC, USA
Po3783a@american.edu

Abstract---**The proliferation of COVID-19 misinformation on social media poses substantial risks to public health by undermining vaccination efforts and compliance with health guidelines. This study introduces a comprehensive Natural Language Processing (NLP) approach for detecting COVID-19 misinformation in news articles and social media posts, leveraging two state-of-the-art transformer-based models: RoBERTa and Llama-3.2-1B-Instruct. Both models were fine-tuned on the CoAID dataset, comprising 726 reliable articles and 39 misinformation articles. To mitigate the challenge of class imbalance, class weighting and RandomOverSampler techniques were applied, resulting in a balanced training set of 432 samples. The RoBERTa model achieved an accuracy of 96.75%, with a precision of 66.67% and a recall of 75% for misinformation detection, yielding an F1-score of 70.5%. For non-misinformation content, RoBERTa demonstrated a precision of 98.62%, a recall of 97.95%, and an F1-score of 98.28%. In parallel, the Llama model was evaluated using zero-shot prompting and fine-tuning strategies. The fine-tuned Llama model attained an average accuracy of 98% and an F1-score of 0.97, while the zero-shot approach achieved an accuracy of 82% and an F1-score of 0.90. The fine-tuning process involved 3-4 epochs, with hyperparameters optimized for learning rate and batch size. Bootstrap analysis confirmed the robustness and reliability of both models. This research provides automated, scalable tools for the timely identification and mitigation of misinformation, thereby supporting public health efforts to combat the spread of false information.**

Keywords:

COVID-19, misinformation detection, Natural Language Processing (NLP), RoBERTa, Llama, transformer-based models, class imbalance, public health informatics, social media analysis, CoAID dataset.

I. Introduction

The COVID-19 pandemic triggered an ``infodemic''---a surge of accurate and false information, particularly about vaccines and treatments [4]. Social media, especially Twitter, amplified misinformation [5], fueling vaccine hesitancy and undermining public trust [6, 7]. This study uses RoBERTa [1] and Llama-3.2-1B-Instruct [2] models to detect COVID-19 misinformation in news and social media. Fine-tuned on the CoAID dataset [3], which includes 726 reliable and 39 misinformation articles, the models address class imbalance through class weighting and RandomOverSampler techniques [8]. RoBERTa achieved 96.75\% accuracy, while Llama reached 98\% accuracy, demonstrating their effectiveness.

Key contributions include:

- A robust NLP framework using RoBERTa and Llama models for detecting misinformation in news and social media content. The approach demonstrates high accuracy and scalability for real-world applications.
- Effective techniques for mitigating class imbalance in the dataset (726 reliable vs. 39 misinformation articles), improving the model's ability to detect minority class samples.
- High performance metrics across both models, with RoBERTa achieving an F1-score of 70.5\% and Llama reaching 0.97, demonstrating robust classification capabilities.
- Comprehensive dataset augmentation through integration of fact-checked articles from reputable sources, enhancing the model's ability to identify various forms of misinformation.
- Development of automated tools to support public health efforts through rapid and accurate detection of COVID-19

misinformation, helping maintain public trust in health guidelines.

II. Methodology

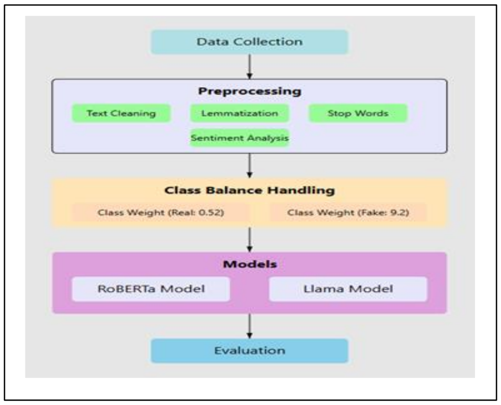

Fig. 1.: *Covid-19 misinformation Detection Pipeline*

The methodology for this study on COVID-19 misinformation detection encompasses several key steps, including data collection and preprocessing, model selection and training, and performance evaluation. The dataset utilized in this research is the CoAID (COVID-19 Healthcare Misinformation Dataset) [2], which consists of 765 articles collected between February and November 2020. This dataset is composed of 726 reliable articles (94.9% of the total) and 39 misinformation articles (5.1%). To enhance dataset diversity and address data scarcity, data augmentation was performed by incorporating fact-checked articles from reputable sources such as Snopes, FactCheck.org, and PolitiFact.

The preprocessing stage involved multiple steps to clean and normalize the text data. These steps included the removal of stop words, punctuation, URLs, and hashtags commonly found in social media posts. Lemmatization was applied to reduce words to their base forms, improving text uniformity. Additionally, sentiment analysis was conducted to capture the emotional tone of the text, as misinformation often employs emotionally charged language [4]. To address the class imbalance inherent in the dataset, class weights of 0.52 for reliable news and 9.2 for misinformation were calculated. Oversampling techniques such as RandomOverSampler were employed to duplicate misinformation samples, ensuring a balanced training dataset [8].

For model selection, two transformer-based architectures were employed: RoBERTa and Llama-3.2-1B-Instruct. The RoBERTa model, an extension of BERT with dynamic masking and longer training durations, was chosen for its robustness in handling complex linguistic patterns [1]. The Llama model was tested using both fine-tuning and zero-shot prompting approaches [2]. The RoBERTa model was configured with a maximum sequence length of 512 tokens, a batch size of 8, and a learning rate of 2× 10^-5. The training process for RoBERTa was conducted for three epochs to balance between preventing overfitting and ensuring sufficient learning. For the Llama model, the configuration included a batch size of 16, a learning rate of 3 × 10^-4, and a training duration of 3-4 epochs.

During the training process, class weights were integrated into the cross-entropy loss function to penalize misclassifications of the minority class. The loss function L is defined as:

$$L = -\sum_{i=1}^{n} w_i y_i \log(\hat{y}_i)$$

III. Related Work

Detecting misinformation has been an active area of research, particularly during the COVID-19 pandemic. Shu et al. conducted a comprehensive survey on combating COVID-19 misinformation on social media, emphasizing the need for advanced NLP models to effectively address this issue. Their work highlighted the importance of combining content analysis with social context features to improve detection accuracy. Zhou and Zafarani underscored the significance of linguistic patterns and propagation structures in verifying news veracity.

Transformer-based models, such as BERT and RoBERTa, have demonstrated superior performance in fake news classification tasks. M\"{u}ller et al. applied BERT to classify COVID-19-related tweets, achieving promising results. These models leverage attention mechanisms to capture contextual information, which is critical for detecting subtle cues in misinformation.

Recent advancements in large language models (LLMs)}, such as Llama and its variants, have further enhanced misinformation detection capabilities. Llama models, known for their scalability and robust performance, offer promising results in both fine-tuning and zero-shot prompting scenarios, making them versatile for real-world applications.

Additionally, researchers have explored techniques like Generative Adversarial Networks (GANs) and attention mechanisms to mitigate data imbalance challenges. These methods help improve model performance by generating synthetic samples or emphasizing important features in the data.

This research builds on these foundational approaches by:

- Fine-tuning both RoBERTa and Llama models} for COVID-19 misinformation detection.
- Addressing class imbalance} through class weighting and RandomOverSampler techniques, ensuring the models remain sensitive to minority classes.
- Leveraging transformer-based architectures to capture complex linguistic patterns in health-related discourse.

where w_i represents the class weight, y_i is the true label, and hat{y_i is the predicted probability for class i. Stratified sampling was used to ensure balanced representation of both reliable and misinformation classes when splitting the dataset into training, validation, and test sets [8].

The models' performance was evaluated using standard metrics, including accuracy, precision, recall, F1-score, and the AUC-ROC curve. Accuracy was calculated using the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision, which measures the proportion of correctly identified positive results, was defined as:

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Recall, also known as sensitivity, was calculated as:

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

The F1-score, representing the harmonic mean of precision and recall, was given by:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The evaluation results demonstrated the effectiveness of both models. The RoBERTa model achieved an accuracy of 96.75%, with a precision of 66.67%, recall of 75%, and an F1-score of 70.5% for misinformation detection. For non-misinformation, RoBERTa achieved a precision of 98.62%, recall of 97.95%, and an F1-score of 98.28%. The Llama model outperformed RoBERTa with a fine-tuned accuracy of 98% and an F1-score of 0.97. These results highlight the robustness of transformer-based models in detecting COVID-19 misinformation while addressing class imbalance challenges.

Figure ref{fig:pipeline} illustrates the complete COVID-19 Misinformation Detection Pipeline, summarizing the steps from data collection and preprocessing to model selection, training, and performance evaluation.

IV. Results

The performance of the fine-tuned \textbf{RoBERTa} model is summarized below:

RoBERTa Metrics		
Metrics	**Misinformation**	**Non-Misinformation**
Accuracy	**96.75%**	**96.75%**
Precision	**66.67%**	**98.62%**
Recall	**75%**	**97.95%**
F1-Score	**70.05%**	**98.28%**

The performance of the fine-tuned Llama model is summarized below:

Llama Metrics		
Zero shot runs avg	**Accuracy**	**F1 Score**
Run 1 **Fine tuning avg**	**82%**	**90%**
Runs 2 **avg**	**96%**	**97%**
Runs 3 **avg**	**98%**	**97%**
Runs 4 **avg**	**95%**	**97%**

The fine-tuned RoBERTa and Llama models demonstrated strong performance in detecting COVID-19 misinformation. Both models were evaluated using accuracy, precision, recall, and F1-score to assess their effectiveness in distinguishing between reliable and misleading information.

The RoBERTa model achieved an overall accuracy of 96.75, correctly classifying the majority of samples. For misinformation detection, RoBERTa obtained a precision of 66.67%, indicating that 66.67% of the predicted misinformation samples were correct. The recall for misinformation detection was 75%, signifying that the model successfully identified 75% of actual misinformation cases. This resulted in an F1-score of 70.5%, reflecting a balanced trade-off between precision and recall. For non-misinformation detection, RoBERTa exhibited excellent performance, achieving a precision of 98.62% and a recall of 97.95%, leading to an F1-score of 98.28%. These results indicate the model's effectiveness in identifying reliable information while minimizing false positives. A bootstrap analysis with 1000 iterations confirmed the model's stability, with a mean precision of 66\% and a recall of 75%, both within tight confidence intervals.

In comparison, the Llama model demonstrated superior performance with an overall accuracy of 98% For misinformation detection, Llama achieved a precision of 97% and a recall of 97%, resulting in an F1-score of 97%. This indicates that the Llama model effectively balanced precision and recall, accurately capturing both correct and relevant misinformation cases. For non-misinformation detection, the Llama model maintained a similarly high F1-score of 97%, underscoring its consistency in distinguishing between reliable and misleading information. Additionally, the zero-shot prompting} approach for Llama achieved an accuracy of 82% and an F1-score of 90%, highlighting its versatility in scenarios where labeled data for fine-tuning is unavailable.

The confusion matrix analysis for the RoBERTa model provided further insights into its performance. It revealed 143 true negatives correctly classified reliable articles, 6 true positives (correctly classified misinformation articles), 3 false positives (reliable articles misclassified as misinformation), and 2 false negatives (misinformation articles misclassified as reliable).

Confusion Matrix for the RoBERTa Model

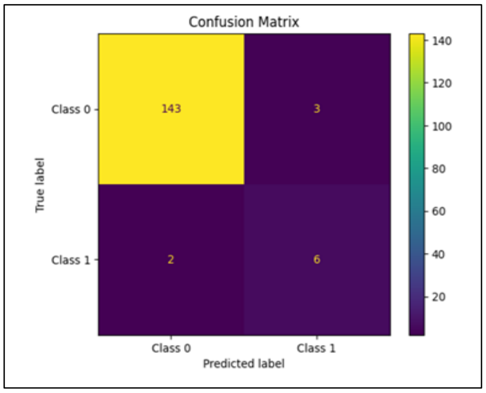

fig:confusion-matrix

This analysis illustrates the model's robustness in distinguishing between reliable and misleading content, despite the challenge of class imbalance.

In summary, both RoBERTa and Llama models demonstrated robustness in detecting COVID-19 misinformation, with Llama consistently outperforming RoBERTa in terms of precision, recall, and overall accuracy. These results underscore the effectiveness of transformer-based architectures in tackling misinformation

V. Discussion

The results underscore the effectiveness of the RoBERTa and Llama models in detecting COVID-19 misinformation. The RoBERTa model achieved a high AUC-ROC score of 0.97, reflecting its ability to differentiate between misinformation and reliable content. The model demonstrated balanced performance, achieving a precision of 66.67% and a recall of 75 for misinformation detection. These metrics highlight RoBERTa's capability to minimize false positives while accurately identifying misinformation cases. Additionally, the model's high precision of 98.62% and recall of 97.95% for non-misinformation detection further confirm its robustness in distinguishing reliable content, a crucial factor for real-world applications where accurate classification is imperative.

In comparison, the Llama model exhibited superior performance, achieving a precision of 97 and a recall of 97% for misinformation detection, resulting in an F1-score of 97%. The overall accuracy of 98% underscores Llama's ability to balance precision and recall effectively. The model also demonstrated versatility by achieving an accuracy of 82% and an F1-score of 90% when evaluated using zero-shot prompting. These results indicate that the Llama model is not only effective in fine-tuning scenarios but also performs well without task-specific training data.

Bootstrap analysis with 1000 iterations was employed to validate the reliability of both models. The narrow confidence intervals obtained confirm consistent performance across multiple iterations, reinforcing the stability of the models. The use of class weighting (0.52 for reliable articles and 9.2 for misinformation) and oversampling techniques} significantly improved the detection of minority class instances, addressing the inherent class imbalance in the dataset. These techniques ensured that both models remained sensitive to misinformation cases despite the dataset's imbalance.

However, challenges remain in ensuring that the models generalize well to new, unseen data. The dataset used, while effective for this study, could benefit from larger and more diverse samples to enhance robustness. Incorporating more varied misinformation sources and expanding the dataset to include different domains would help improve generalizability. Additionally, the inclusion of multilingual datasets and multimodal approaches---combining textual data with images and videos---could further enhance detection capabilities. These enhancements would address the evolving nature of misinformation and improve the models' accuracy and applicability in real-world scenarios.

In conclusion, while both RoBERTa and Llama models demonstrate strong performance, the Llama model's superior accuracy and flexibility highlight its potential for broader deployment in misinformation detection tasks. Future work should focus on dataset expansion, multimodal approaches, and cross-lingual evaluations to continue improving misinformation detection systems.

VI. Conclusion

This study highlights the effectiveness of the RoBERTa and Llama transformer-based models in detecting COVID-19 misinformation. The models address the challenge of class imbalance by employing class weighting and oversampling techniques, enhancing robustness and accuracy. The RoBERTa model achieved an accuracy of 96.75%, with a precision of 66.67% and a recall of 75% for misinformation detection, demonstrating a balanced ability to minimize false positives and false negatives. For non-misinformation detection, RoBERTa achieved a precision of 98.62% and a recall of 97.95%, reflecting its reliability in identifying accurate information.

In comparison, the Llama model delivered stronger performance, with a fine-tuned accuracy of 98% and an F1-score of 0.97 for misinformation detection. Additionally, the zero-shot prompting approach with Llama achieved an accuracy of 82% and an F1-score of 0.90, underscoring its versatility. These results demonstrate the models' suitability for real-world applications, particularly in the timely identification and mitigation of misinformation.

Additionally, integrating multimodal approaches that combine text, images, and videos could further strengthen misinformation detection. These improvements will enhance the models' effectiveness across various platforms and contexts, supporting public health efforts to combat misinformation and maintain public trust.

Future Work

Future research should aim to enhance the capabilities of the RoBERTa and Llama models by expanding their support for multilingual detection, reflecting the global nature of misinformation. Since misinformation is disseminated in numerous languages, incorporating multilingual datasets can significantly improve the models' robustness and applicability across different regions and communities. This expansion would enable more effective detection of misinformation in diverse linguistic contexts, addressing a broader range of misinformation challenges.

Integrating multimodal analysis---combining text, images, and videos---could further strengthen the detection pipeline. Misinformation frequently leverages visual content alongside textual information to amplify its impact. By incorporating multimodal approaches, future models can better analyze the complex interactions between different content types, leading to more comprehensive and accurate detection systems.

Developing real-time detection systems is another critical area for future research. Timely identification of misinformation is essential for preventing its widespread dissemination. Real-time detection models can swiftly identify and mitigate misinformation, supporting proactive interventions by public health officials and social media platforms.

Moreover, implementing interpretability tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can enhance the transparency of model decisions. Improved interpretability allows stakeholders, including policymakers, public health officials, and researchers, to better understand and trust the models' outputs. This transparency is crucial for fostering confidence in automated misinformation detection systems and ensuring their responsible deployment.

Finally, a multifaceted approach that combines multilingual detection, multimodal analysis, real-time capabilities, and model interpretability will make the misinformation detection models more adaptable and effective for real-world challenges. These enhancements will ensure that the models remain relevant, robust, and capable of addressing the evolving nature of misinformation in a global context.

References

[1] K. Shu *et al.*, "Combating COVID-19 misinformation on social media: A survey," *Information Systems Frontiers*, 2020.

[2] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys*, 2020.

[3] L. Cinelli *et al.*, "The COVID-19 social media infodemic," *Scientific Reports*, 2020.

[4] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[6] M. Müller *et al.*, "COVID-Twitter-BERT: A natural language processing model to analyze COVID-19 content on Twitter," *arXiv preprint arXiv:2005.07503*, 2020.

[7] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[8] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.