From Words to Stories: Narrative Framing, Characterization and Extraction in Online News

1st Kushal Chandani
Computer Science
Habib University
Karachi, Pakistan
kc07535@st.habib.edu.pk

2nd Hareem Siraj
Computer Science
Habib University
Karachi, Pakistan
hs07488@st.habib.edu.pk

3rd Dua e Sameen
Computer Science
Habib University
Karachi, Pakistan
ds07138@st.habib.edu.pk

4th Ayesha Enayat
Computer Science
Habib University
Karachi, Pakistan
ayesha.enayat@sse.habib.edu.pk

***Abstract*—This study addresses multilingual propaganda detection and narrative analysis by tackling three subtasks: entity framing, narrative classification, and narrative extraction, as outlined in SemEval 2025 Task 10. Using transformer-based models such as BERT, GPT-2, T5-Small, BART, and XLM-RoBERTa, the project fine-tunes these architectures for nuanced and hierarchical text processing tasks. GPT-2 excelled in entity framing, BERT in narrative classification, and T5-Small in generating coherent narrative explanations. The results demonstrate the potential of transformer models for multilingual and context-rich applications while highlighting challenges like class imbalance, long-context dependencies, and dataset limitations. This work advances NLP methodologies, offering insights for misinformation detection, ethical journalism, and narrative analysis. Future work includes enhancing model explainability, expanding cross-lingual datasets, and refining narrative comprehension systems.**

***Index Terms*—Natural Language Processing (NLP), Multilingual NLP, Entity Framing, Narrative Classification, Narrative Extraction, Transformer Models, GPT-2, BERT, T5-Small, BART, XLM-RoBERTa, Text Generation, Hierarchical Classification, Misinformation Detection.**

I. Introduction

In an era where information flows ceaselessly across digital platforms, propaganda has emerged as a potent tool for shaping narratives, influencing public opinion, and manipulating societal beliefs. From political campaigns to targeted misinformation, the pervasive nature of propaganda poses critical challenges to democratic systems, public trust, and informed decision-making. The ability to detect and analyze propaganda has become a pressing need to combat its harmful effects on global discourse.

This project tackles the intricate challenges posed by SemEval 2025 Task 10, a groundbreaking initiative designed to advance Natural Language Processing (NLP) techniques for propaganda analysis. The task is divided into three distinct yet interconnected subtasks:

1) **Entity Framing**
   Assign roles (e.g., protagonist, antagonist, innocent) to named entities in news articles based on how they are presented. This is a text classification task.
2) **Narrative Characterization**
   Identify and assign relevant narrative labels to news articles from a predefined taxonomy. This is a document classification task.
3) **Narrative Extraction**
   Generate a short explanation, based on the text, that supports the dominant narrative of a news article. This is a text generation task.

To achieve these objectives, the project leverages state-of-the-art transformer-based models, including **BERT**, **GPT-2**, **BART**, **T5-Small**, and **XLM-RoBERTa**. Each model was selected and fine-tuned based on its strengths for specific tasks: BERT for entity role classification, GPT-2 for entity framing and narrative extraction, T5-Small for generating fluent and semantically aligned explanations, BART for narrative classification, and XLM-RoBERTa for multilingual context understanding. An ensemble approach combining BERT and XLM-RoBERTa was also explored for Subtask 2 to enhance performance by leveraging complementary model capabilities.

By addressing these subtasks, the project not only advances academic research in propaganda detection but also contributes to practical applications, such as combating misinformation in media, fostering critical thinking in education, and enhancing transparency in communication. As the digital landscape continues to evolve, the tools and insights developed here promise to play a vital role in safeguarding the integrity of public discourse.

II. Related Work

The growing complexity of multilingual narratives, framing, and propaganda detection has necessitated innovative approaches in natural language processing (NLP). This section examines existing methodologies and their contributions to multilingual narrative analysis, framing detection, and coherence modeling.

A. Multilingual Framing and Narrative Analysis

Piskorski *et al.* introduced a multilingual dataset annotated with genre, framing, and persuasion techniques across six European languages [1]. This dataset includes over 37,000 persuasion spans from 1,612 articles, covering genre (e.g., objective reporting, satire), framing dimensions (e.g., economic, health), and persuasion techniques (e.g., "Appeal to Fear," "Casting Doubt"). They used XLM-RoBERTa for classification tasks, achieving a macro F1 score of 0.592 for genre classification and demonstrating cross-lingual transfer in framing detection. These findings align with Subtask 2, as the identification of sub-narratives relies on similar hierarchical structures and cross-lingual capabilities.

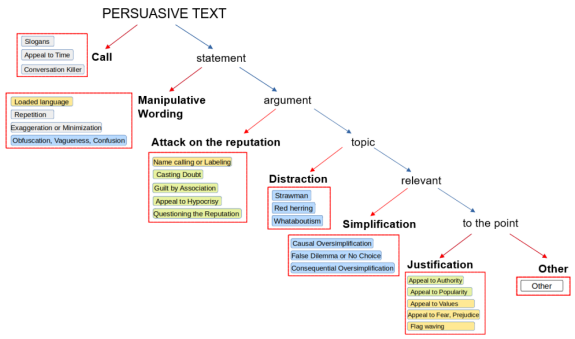

Fig. 1. Decision diagram

Akter and Anastasopoulos proposed the Student-Sourced Noisy Frames Corpus (SNFC), which expands framing analysis to 12 typologically diverse languages, including Bengali, Portuguese, and Chinese [3]. Their work highlights task-specific fine-tuning and cross-lingual transfer, addressing translation noise and semantic drift. These techniques are relevant to both Subtask 1 and Subtask 2, especially in adapting models to multilingual datasets.

B. Graph-Based and Temporal Methods for Coherence

Xu *et al.* introduced the Narrative Cognition (NARCO) framework, which uses graph-based methods to model causal and temporal coherence in narratives [2]. NARCO represents text as a graph, where nodes are text chunks, and edges encode coherence relations through questions like "What motivated this action?" or "What are the consequences?" These graphs enhance downstream tasks like plot retrieval and long-document QA.

This method directly informs Subtask 3: Narrative Extraction, as it requires generating coherent explanations for dominant sub-narratives. The retrospective question generation aligns with the need to maintain causal and temporal coherence in text generation tasks.

C. Transformer-Based Architectures

Transformer-based models have significantly advanced NLP applications in framing and narrative analysis. Piskorski *et al.* used XLM-RoBERTa to preprocess and classify long documents, demonstrating its cross-lingual effectiveness [1]. This is critical for Subtask 1 and Subtask 2, where transformer models are used for multilingual role classification and sub-narrative detection.

Sinelnik and Hovy explored multilingual framing in disinformation campaigns, achieving a macro F1 score of 32.9 across multiple languages [4]. Their work tackled challenges like semantic drift and translation artifacts, directly relevant to improving the multilingual robustness of Subtasks 1 and 2.

D. Dynamic and Memory-Augmented Models

Papalampidi *et al.* developed a dynamic entity memory mechanism integrated into Transformer-XL to maintain logical consistency in narrative generation [5]. By dynamically tracking entity attributes and actions, their approach improved coherence and fluency by over 18%. This aligns with Subtask 1 where assigning consistent roles to entities is essential. Additionally, the memory mechanism offers insights into enhancing coherence in Subtask 3, particularly for grounding explanations in relevant context.

E. Topic Modeling in Multilingual Propaganda

Schäfer *et al.* applied BERTopic to multilingual datasets, extracting dominant themes like vaccine safety and economic impacts during the COVID-19 pandemic [6]. Using c-TF-IDF and hierarchical clustering, BERTopic identified granular subthemes and aligned topics across languages. This methodology informs Subtask 2, where hierarchical clustering parallels the classification of sub-narratives within a two-level taxonomy. For Subtask 3, thematic alignment offers a framework for generating concise and contextually accurate explanations.

F. Applications in Disinformation Campaigns

Sinelnik and Hovy analyzed framing in multilingual disinformation campaigns, identifying variations in narrative strategies across Russian, French, Spanish, and Italian articles [4]. They employed lexicon-based and transformer-based approaches to classify frames like health, crime, and morality. These techniques are directly applicable to Subtask 1 and Subtask 2, where understanding framing variations informs role classification and narrative identification.

G. Advancements in Narrative Generation

Papalampidi *et al.*'s work on dynamic memory models and the use of BART for text generation highlight advancements in maintaining coherence and fluency in narratives [5], [7]. These approaches directly support Subtask 3, where generating coherent explanations requires advanced sequence-to-sequence models and robust grounding in the input text.

III. Methodology & Experiments

A. Task 1: Entity Framing:

Initially, we trained the **BERT Transformer Model** from Hugging Face as the baseline for our task. BERT models are well-suited for classification tasks due to their encoder-based architecture and bi-directional context understanding. Given

the dataset's small size (400 rows), we focused on classifying main roles rather than fine-grained categories. Custom prompting techniques were used to improve performance, and the baseline model was fine-tuned using the standard hyperparameters mentioned later.

To enhance the efficiency and accuracy of our approach, we employed an ensembling strategy that combined two models to improve predictive performance. The models chosen were:

- **BART-base:** This model excels at tasks requiring sequence-to-sequence modeling, making it effective for capturing contextual relationships.
- **BERT-base:** Provides encoder-based contextual understanding for classification tasks.
- **GPT-2:** A decoder-based model known for its generative capabilities.
- **XLM-RoBERTa-base:** A large model trained on multilingual corpora, enhancing the model's performance on diverse textual data.

We initially thought the ensembling approach would leverage the predictions of these individual models to produce a more accurate outcome, effectively balancing their strengths.

The dataset consisted of 400 English-language news articles focused on disinformation narratives surrounding the Ukraine-Russia war and Climate Change. Each article was labeled with roles: Protagonist, Antagonist, and Innocent and then further divided into fine-grained roles.

We also applied the **Dataset Augmentation technique** using the **Gemini API**, but limitations like its 60 requests per minute cap and Google Colab session constraints hindered scalability. Despite these challenges, we were able to increase the size of the dataset from the original 400 rows to over 800 rows using the technique. During Augmentation we also manually checked the rows and made sure that custom-prompting is indicating existing sentences in the dataset provide us with the active to passive or more synonymous sentences that were concatenated with the original dataset. The total dataset was of around **1200 rows** (800 Augmented rows and 400 Original rows)

All the models were fine-tuned as follows:

- **Data Split**: 80 percent training, 10 percent validation, 10 percent testing.
- **Preprocessing**: Text cleaning with **NLTK stopwords**, followed by tokenization using the **tokenizers of each model**.
- **Training Details**: The model was trained using the hyperparameters mentioned in Table I.

B. Task 2: Narrative Characterization

Initially, to establish a baseline for narrative characterization, we fine-tuned the **BERT (bert-base-uncased)** model. Its tokenization capabilities preserved the relationships between narratives effectively, making it a suitable choice for this task.

TABLE I
TASK 1: HYPERPARAMETERS

Model(s)	Learning Rate	Epochs	Batch Size
All Models	$5e^{-5}$	5	16

The dataset consisted of **200 labeled documents**, with each document associated with narratives and multiple sub-narratives. The labels were multi-class, representing the diverse sub-narrative structure of the dataset.

To improve the performance and efficiency of our approach, we employed an **augmentation strategy** and an **ensembling technique**:

- **Dataset Augmentation**: The dataset size was increased from the original 200 rows to over 400 rows using the **Gemini API**, despite its limitations (as mentioned in Task 1). The total dataset comprised approximately **530 rows** (330 augmented rows and 200 original rows).
- **Loss Function**: **BCEWithLogitsLoss** was used to handle multi-label classification.

  The models chosen included:
  - **BART-base**: Effective for sequence-to-sequence contextual relationships.
  - **BERT-base**: Provides encoder-based narrative understanding.
  - **XLM-RoBERTa-base**: Excels with multilingual and diverse textual data.
  - **Ensemble of BERT-base and XLM-RoBERTa-base**: Combines predictions to leverage their strengths.

  All models were fine-tuned using the following setup:
- **Data Split**: 80 percent training, 10 percent validation, 10 percent testing.
- **Preprocessing**: Text cleaning with **NLTK stopwords** and tokenization using the models' respective tokenizers.
- **Training Details**: The hyperparameters used are detailed in Table II.

TABLE II
TASK 2: HYPERPARAMETERS

Dataset	Learning Rate	Epochs	Batch Size
Narratives	$5e^{-5}$	5	16
Sub-narratives	$1e^{-5}$	5	16

C. Task 3: Narrative Extraction

Initially, to establish a baseline for narrative Extraction, we fine-tuned the **BART** model. Its sequence-to-sequence capabilities preserved the contextual relationships between narratives effectively, making it a suitable choice for the text

generation task.

The dataset consisted of **80 labeled documents**, with each document associated with dominant narratives and sub-narratives. Preprocessing included tokenization mainly.

- **Loss Function**: **Cross-Entropy Loss** was used for multi-class classification tasks. The models chosen included:
  - **BART**: Effective for sequence-to-sequence contextual relationships.
  - **GPT-2**: Provides strong generative capabilities for narrative elaboration.
  - **T5-small**: Excels in text-to-text transfer tasks for flexible narrative modeling.

All models were fine-tuned using the following setup:

- **Data Split**: 80 percent training, 10 percent validation, 10 percent testing.
- **Training Details**: Following table III includes the training hyperparameters.

TABLE III
TASK 3: HYPERPARAMETERS

Model(s)	Learning Rate	Epochs	Batch Size
All Models	$2e^{-5}$	5	6

Note For Task 3: We did try the **data augmentation technique** and we had combined data of over 200 rows however due to the limitations of Colab limits we were not able to get the scores. The same limitation we encountered during the **ensembling technique** as well as Colab does not have enough storage to train 2 large models like BART and GPT2 at the same time.

Lastly, we did not clean the data in this task as cleaning resulted in low scores. All the characters such as colons, inverted commas played a major role in making sure the generated summary is aligned with the original text.

IV. RESULTS AND FINDINGS

A. Task 1: Entity Framing

Initially, we fine-tuned the **BERT-base-uncased** model to establish a baseline for performance. During training, we recorded key metrics such as validation loss, accuracy, and F1 score for each epoch, using a consistent 80-10-10 training-validation-test split. The results of our fine-tuning, detailed in Table IV, indicate that the BERT model achieved a validation loss of **0.162**, an accuracy of **75.24%**, and an F1 score of **0.652** over 5 epochs.

To further improve performance, we evaluated additional models, including **BART**, **GPT2**, and **XLM-Roberta**. Among these, **GPT2** demonstrated the best overall performance, achieving a validation loss of **0.113**, an accuracy of **82.51%**, and an F1 score of **0.756** over 5 epochs. **BART** also outperformed BERT, with an F1 score of **0.743** and a validation loss of **0.131**. In contrast, **XLM-Roberta** underperformed relative to other models, with an F1 score of **0.535** and an accuracy of **69.46%**.

TABLE IV
PERFORMANCE COMPARISON OF MODELS

Model(s)	Validation Loss	Accuracy	F1 Score	Epoch
BART	0.131210	0.817176	0.743243	5
BERT	0.162397	0.752423	0.651741	5
GPT2	**0.113133**	**0.825133**	**0.756152**	**5**
XLM-Roberta	0.187561	0.694608	0.535433	5

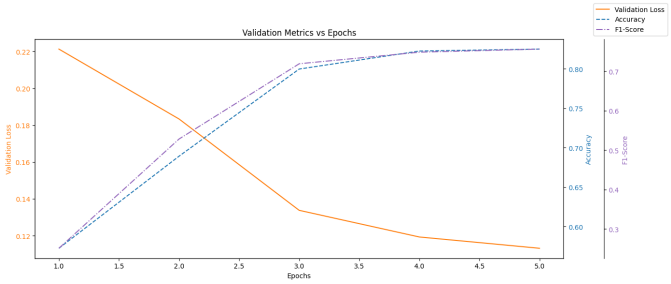

Fig. 2. GPT2's performance over 5 epochs

These findings underscore the competitive performance of **GPT2** as the best candidate for future applications, with BART as a strong alternative. The following table shows Weighted-average F1 scores on the test set, which also supports the results:

Model	Weighted-average F1
BART	0.6723
BERT	0.5401
Ensemble (GPT-2 + BART)	0.7074
GPT-2 ★	0.7166
XLM-RoBERTa	0.3820

Fig. 3. Visual representation of F1 scores across different models

B. Task 2: Narrative Characterization

To evaluate the performance of the models for Subtask 2, we fine-tuned and tested **BART**, **BERT**, **XLM-RoBERTa**, and an **Ensemble** approach combining **BERT** and **XLM-RoBERTa**. The objective was to classify narratives and sub-narratives while ensuring hierarchical consistency. Performance was measured using accuracy and F1 Score across five epochs, with the best-performing epoch for each model selected for comparison.

Table V summarizes the results, highlighting the accuracy and F1 Score achieved by each model.

TABLE V
PERFORMANCE METRICS FOR SUBTASK 2 MODELS

Model	Epoch	Accuracy	F1 Score
BART	5	0.60	0.34
BERT	**5**	**0.59**	**0.36**
XLM-RoBERTa	5	0.58	0.27
Ensemble	5	0.50	0.05

BERT achieved the highest F1 Score of **0.36**, showcasing its ability to balance precision and recall, making it well-suited for handling imbalanced classes. Despite slightly lower accuracy of **0.59** compared to BART, BERT's performance highlights its strengths in hierarchical narrative classification. The trends in BERT's performance over epochs are visualized in Figure 4.

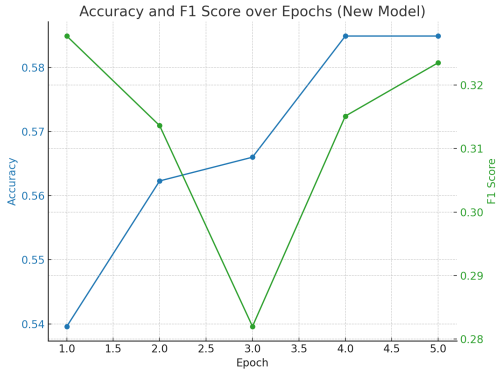

Fig. 4. Accuracy and F1 Score trends for BERT.

BART recorded the highest accuracy of **0.60** with a competitive F1 Score of **0.34**, reflecting its strong overall classification performance. However, it struggled slightly with rare sub-narratives, where BERT performed better.

XLM-RoBERTa exhibited moderate accuracy of **0.58** and a lower F1 Score of **0.27**, indicating limited effectiveness in balancing precision and recall. This suggests a potential mismatch between its capabilities and the requirements of hierarchical classification.

The **Ensemble** approach underperformed, achieving only **0.50** accuracy and **0.05** F1 Score. This indicates a lack of complementary strengths between BERT and XLM-RoBERTa, leading to poor generalization.

C. Task 3: Narrative Extraction

Our Subtask 3 focuses on generating concise explanations for dominant narratives/sub-narratives in news articles. The initial approach involved fine-tuning pre-trained transformer-based language models (BART, GPT-2, and T5-Small) for this task. Each model was trained on a dataset using a similar training pipeline with hyperparameters optimized for performance. The goal was to evaluate the models based on BLEU, and RougeL scores, which reflect their ability to generate coherent and accurate text.

Model	Validation Loss	BLEU	RougeL
BART	3.817860	0.032723	0.283015
GPT-2	3.054182	0.127164	0.366494
T5 (Small)	1.753615	0.097883	0.407796

TABLE VI
COMPARISON OF MODELS

The final evaluation for this task was conducted using the **BertScore F1 metric**, which assesses the semantic similarity between the generated explanations and the ground truth references. Figure 4. summarizes the average F1 scores achieved by each model.

Model	Average F1 Score
T5-small ★	0.8306
GPT-2	0.8179
BART	0.6564

Fig. 5. Average F1 Scores for Subtask 3 Models (BertScore)

The **T5-Small model** achieved the highest average F1 score of **0.8306**, demonstrating its superior ability to generate semantically aligned and fluent narrative explanations. This result highlights the model's strength in effectively capturing the meaning and structure of the reference narratives, making it the most suitable for this task.

The **GPT-2 model** closely followed, with an average F1 score of **0.8179**. While slightly lower than T5-Small, GPT-2 showed good performance in maintaining semantic coherence and fluency, making it a competitive alternative for narrative extraction tasks.

The **BART model**, with an average F1 score of **0.6564**, demonstrated comparatively weaker performance. This indicates challenges in generating semantically accurate explanations and aligning with the ground truth references.

V. DISCUSSION AND IMPLICATIONS

Across the tasks, our models demonstrated strengths in semantic understanding, hierarchical classification, and text generation while revealing areas where we can improve, such as handling ambiguous contexts, imbalanced datasets, and extended input dependencies.

For Subtask 1, we used GPT-2 to capture entity roles within narratives, showcasing its generative modeling capabilities for text classification. However, challenges like role ambiguity and overlapping contexts showed us the need to integrate external knowledge sources, such as symbolic reasoning or knowledge graphs, to enhance contextual understanding. These improvements could benefit tasks like opinion mining and character-driven narrative analysis.

In Subtask 2, we saw BERT excel in hierarchical classification, achieving the highest F1 score despite class imbalances. This underscores the value of contextual embeddings in addressing uneven class distributions. We believe exploring further approaches, such as contrastive learning in narratives, could further boost performance in complex tasks like discourse analysis and sentiment prediction.

For Subtask 3, our use of T5-Small demonstrated its capability to generate fluent, semantically aligned narrative explanations. However, the challenge of managing long-context dependencies remains, pointing to the need for innovations like memory-augmented models or hierarchical attention mechanisms. These could also improve other tasks requiring extended context management, like multi-document summarization and generative question answering.

VI. Conclusion

Conclusively, we tackled the challenges of propaganda detection and narrative analysis through three tasks: entity framing, narrative classification, and narrative extraction, as outlined in SemEval 2025 Task 10. By employing advanced transformer models like BERT, GPT-2, T5-Small, BART, and XLM-RoBERTa, we demonstrated their effectiveness in addressing nuanced, multilingual tasks. GPT-2 excelled in entity framing, showcasing the strengths of generative models in contextual role assignment. BERT delivered a good performance in hierarchical narrative classification despite class imbalance, achieving the highest F1 score. For narrative extraction, T5-Small stood out by generating fluent and semantically aligned narrative explanations, highlighting the potential of sequence-to-sequence architectures for text generation.

While we achieved promising results, challenges such as dataset imbalance, long-context dependencies, and data limitations highlighted areas for improvement. Addressing these issues will require innovations in model architectures, training strategies, and data augmentation techniques. .

VII. Future Work

Our project delivered promising results but highlighted areas for improvement. To address the problems, like class imbalance in task 2, we plan to implement techniques like upsampling, synthetic data generation, and class-weighted loss functions. For task 3, challenges with long-context dependencies point to the need for advanced architectures like Pegasus or hierarchical attention mechanisms. Domain-specific pretraining and multi-task learning could further enhance performance across all tasks by enabling more precise and context-aware representations however we are always limited by resources such as Google Colab limits.

The upcoming release of additional training data offers an opportunity to fine-tune our models, improving the scalability. Expanding to multilingual datasets could broaden our models' applicability for effective propaganda detection. By integrating explainability techniques, we aim to make our models more transparent, fostering adoption in applications like media monitoring, education, and ethical journalism. These improvements will enhance both the practical relevance and societal impact of our work.

References

[1] J. Piskorski, N. Stefanovitch, N. Nikolaidis, G. Da San Martino, and P. Nakov, "Multilingual Multifaceted Understanding of Online News in Terms of Genre, Framing, and Persuasion Techniques," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 3001–3022.

[2] L. Xu, J. Li, M. Yu, and J. Zhou, "Fine-Grained Modeling of Narrative Context: A Coherence Perspective via Retrospective Questions," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 5822–5838.

[3] S. S. Akter and A. Anastasopoulos, "A Study on Scaling Up Multilingual News Framing Analysis," *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4156–4173, 2024, doi: https://doi.org/10.18653/v1/2024.findings-naacl.260.

[4] A. Sinelnik and D. Hovy, "Narratives at Conflict: Computational Analysis of News Framing in Multilingual Disinformation Campaigns," pp. 225–237, Jan. 2024, doi: https://doi.org/10.18653/v1/2024.acl-srw.21.

[5] P. Papalampidi, K. Cao, and T. Kocisky, "Towards Coherent and Consistent Use of Entities in Narrative Generation," *arXiv.org*, 2022. [Online]. Available: https://arxiv.org/abs/2202.01709. [Accessed: Nov. 25, 2024].

[6] K. Schäfer, J.-E. Choi, I. Vogel, and M. Steinebach, "Unveiling the Potential of BERTopic for Multilingual Fake News Analysis – Use Case: Covid-19," *arXiv.org*, 2024. [Online]. Available: https://arxiv.org/abs/2407.08417. [Accessed: Nov. 27, 2024].

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *arXiv preprint arXiv:1810.04805*, 2018.

[8] Hugging Face. "Transformers Library Documentation." Available: https://huggingface.co/docs/transformers.

[9] Explosion AI. "Spacy NLP Library." Available: https://spacy.io/.