Multilingual Characterization and Extraction of Narratives from Online News - SemEval Task 10

Muhammad Khubaib Mukaddam
School of Science and Engineering
Habib University
Karachi Pakistan
mk07218@st.habib.edu.pk

Muhammad Shoaib Khursheed
School of Science and Engineering
Habib University
Karachi Pakistan
mk07149@st.habib.edu.pk

Muminah Khurram
School of Science and Engineering
Habib University
Karachi Pakistan
mk07521@st.habib.edu.pk

Abdul Samad
School of Science and Engineering
Habib University
Karachi Pakistan
abdul.samad@sse.habib.edu.pk

Sandesh Kumar
School of Science and Engineering
Habib University
Karachi Pakistan
sandesh.kumar@sse.habib.edu.pk

***Abstract*—In this advancing world, we face a very big issue of spread of harmful disinformation and propaganda, and with the advancements in modern technology, the rate at which it occurs now is unprecedented. The identification and analysis of narratives within online news are crucial for combating disinformation and manipulation. This research aims to cater to the SemEval 2025 Task 10 on ”Multilingual Characterization and Extraction of Narratives from Online News.” We hope to use different Deep Learning techniques to solve for all three of the subtasks; classify the roles and fine-grained roles of the different entities given in each article, classify the narratives and fine-grained subnarratives of each article, and finally based on the narratives of the article, create a summary that highlights the main context of the article. We aim to conduct comprehensive experiments on multilingual datasets, hoping to provide insights into the dynamics of online narratives and play our part in mitigating the impact of harmful disinformation.**

I. Introduction

In the current digital landscape, the widespread dissemination of deceptive content, misinformation, and propaganda, particularly during times of global crisis, has become a significant concern. Such content can profoundly influence public opinion and exacerbate political and social issues. Addressing these challenges requires advanced techniques to automatically analyze and classify news articles, allowing for a better understanding of manipulation attempts and disinformation strategies.

This paper presents the findings of our team, ”Narrative Miners,” as we participated in SemEval Task 10, titled Multilingual Characterization and Extraction of Narratives from Online News. This task focuses on the automatic identification and analysis of narratives in online news articles, specifically addressing propaganda related to the Ukraine-Russia war and climate change. The goal is to develop systems capable of understanding the nuances of narrative-driven disinformation across multiple languages.

The task comprises three primary objectives: narrative classification, narrative farming, and entity extraction. For narrative classification, the aim is to categorize articles into predefined narratives and subnarratives, uncovering the underlying strategies used in disinformation campaigns. Narrative farming seeks to generate concise explanations, grounded in textual evidence, that justify the assignment of a dominant narrative to an article. Lastly, entity extraction focuses on classifying entities mentioned in news articles based on their roles (e.g., protagonist, antagonist) within the context of manipulative narratives. Through these tasks, our project contributes to the development of automated tools to combat the spread of misleading information by providing a deeper understanding of how narratives are constructed and used to shape public discourse.

II. Research Question

”How can we develop multilingual systems capable of automatically identifying, classifying, and extracting narratives in online news articles to detect and characterize disinformation and manipulation attempts?”

A. Subtask 1 - Entity Framings

“How can we automatically classify entities mentioned in news articles based on their roles (e.g., protagonist, antagonist) within the context of manipulative narratives?”

B. Subtask 2 - Narrative Classification

“How can we accurately classify articles into predefined narratives and subnarratives to uncover the underlying disinformation strategies?”

C. Subtask 3 - Narrative Extraction

“How can we automatically generate concise explanations grounded in textual evidence that justify the assignment of a dominant narrative to an article?”

III. Dataset

The dataset for these tasks consists of news articles available in five different languages: Russian, English, Hindi, Bulgarian, and Portuguese. As of the October 16 2024 release, SemEval has released the data for English, Bulgarian, Hindi, and Portugese. The total number of articles available to us are around 700, with the our focus being on the 200+ English articles provided to us.

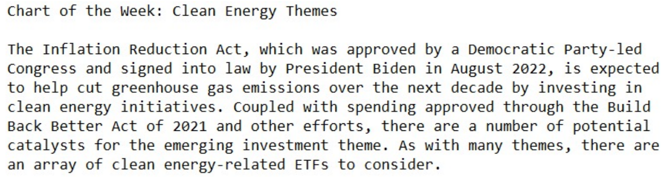

Chart of the Week: Clean Energy Themes

The Inflation Reduction Act, which was approved by a Democratic Party-led Congress and signed into law by President Biden in August 2022, is expected to help cut greenhouse gas emissions over the next decade by investing in clean energy initiatives. Coupled with spending approved through the Build Back Better Act of 2021 and other efforts, there are a number of potential catalysts for the emerging investment theme. As with many themes, there are an array of clean energy-related ETFs to consider.

Fig. 1. Overview of the dataset. Example English Article: EN_CC_10030

A. Subtask 1 - Entity Framings

TABLE I
Training Data for Task 1

Art. ID	Entity. Mention	Start	End	Main. Role	Fine-Grained. Roles

The article ID is the numeric id of input article file, the mentioned entity refers to the entity that we have to classify based on the context provided in the article. Each article may have multiple entities that may need to be classified. The data also provides the relevant starting index and stopping index of the entity's mention in the article. The main role and the fine-grained roles are basically what the model will be predicting based on all the previous content given to it during training.

B. Subtask 2 - Narrative Classification

The dataset for this sub-task consists of news articles. The training dataset however includes news articles annotated with dominant narratives and subnarratives. Each article is assigned an article id, a dominant narrative (e.g., ”Blaming Others”) and a dominant sub-narrative (e.g., ”Ukraine is the aggressor”) structured as follows:

TABLE II
Training Data for Narrative Classification

Art. ID	Dom. Narr.	Dom. Subnarr.

Some taxonomies for Ukraine-Russia War are as follows:

1) Blaming the war on others rather than the invader
   - Ukraine is the aggressor,
   - The West are the aggressors,
2) Discrediting Ukraine
   - Rewriting Ukraine's history,
   - Discrediting Ukrainian nation and society,
   - Discrediting Ukrainian military,

Similarly some taxonomies for Climate Change are as follows:

1) Climate change is beneficial
   - CO2 is beneficial,
   - Temperature increase is beneficial
2) Downplaying climate change
   - Climate cycles are natural,
   - Weather suggests the trend is global cooling

C. Subtask 3 - Narrative Extraction

Subtask 3, follows Subtask 2's pipeline as the input for this task utilizes the results from the output generated in Subtask 2. This setup ensures that the narrative extraction process is informed by the refined textual evidence identified in the previous subtask, enabling more accurate and context-aware extraction of dominant narratives. The dataset for this sub-task consists of news articles annotated with dominant narratives and explanations that justify the assignment of these narratives. Each article is assigned an article id, a dominant narrative (e.g., ”Blaming Others”), a dominant sub-narrative (e.g., ”Ukraine is the aggressor”), and an explanation that connects the text of the article to the assigned narrative. The dataset's format is structured as follows:

TABLE III
Training Data for Narrative Extraction

Art. ID	Dom. Narr.	Dom. Subnarr.	Expl.

IV. Literature Review

A. Subtask 1 - Entity Framings

- **ACCEPT at SemEval-2023:** The team used a unique approach for frame detection in multilingual news articles. Different framings basically represented different perspectives on one and the same reported event. They used an ensemble approach; integrating Large Language Models, Static Word Embeddings, and Commonsense Knowledge Graphs; Graph neural networks that utilize contextualized subgraphs. The system was evaluated on six languages with the team achieving a micro F1-score of 50.69% and a macro F1-score of 50.20% in English, ranking quite high in the task. Something that was quite interesting from their research paper was the use of the of the commonsense knowledge graphs; something our team had never encountered, and how it contributed to the most significant performance boost. The team Fine-tuned pretrained models like RoBERTa for multilingual framing detection by breaking down articles into sentences. Commonsense Knowledge Graphs were extracted from ConceptNet to incorporate commonsense knowledge to improve frame predictions. [1]
- **MarsEclipse at SemEval-2023:** The team achieved first place in framing detection for five out of six languages where training data was available. The unique standout

point in the approach of this team was their use of the contrastive learning in a multi-label text classification task allowing it to identify positive and negative example pairs, pulling similar frames closer together and pushing different frames further apart. They used a two-input architecture; using both the title and body of the article. The team got F1-scores: German (0.711), Italian (0.617), Polish (0.673), and Georgian (0.645). The team made use of XLM-Roberta, applying a brute-force metthod to determine the optimal thresholds for classifying frames. Using the the binary cross-entropy as well, they were able to convert the multi-label classification problem into 14 binary classification problems. [2]

- **A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing** [3] performs a comprehensive survey that systematically categorizes methods to address class imbalance in NLP. It provides a detailed analysis on various sampling strategies, data augmentation techniques, staged learning, and the different types of weights that can be used to resolve class imbalances. This paper is relevant to our work, particularly in the context of dealing with class imbalance in the sub task 1, where oversampling techniques were employed to handle the class imbalance.

B. Subtask 2 - Narrative Classification

Narrative classification techniques have advanced significantly with the integration of state-of-the-art machine learning models, particularly transformer-based architectures like BERT, RoBERTa, and GPT. These models have driven progress in shared tasks, such as SemEval-2020, SemEval-2023, and CheckThat! 2024, focusing on fine-grained narrative classifications like propaganda detection, framing, and persuasion techniques.

For instance, SemEval-2020 Task 11 addressed the detection of propaganda in news articles, involving span identification and technique classification for 14 propaganda techniques (e.g., ”Loaded Language” and ”Appeal to Authority”). Top systems like Hitachi and ApplicaAI employed transformers with LSTMs and CRFs, achieving F1-scores of 51.55 for span identification and 62.07 for technique classification. These systems also utilized data augmentation to boost performance, especially in low-resource tasks like detecting subtle techniques such as Whataboutism, which had precision scores of 20-30 percent, compared to 60 percent+ for common techniques like Loaded Language. [4]

SemEval-2023 Task 3 extended narrative classification to multilingual datasets across nine languages. In genre categorization, macro F1 scores reached 0.78-0.85 for English and 0.76-0.82 for other major languages. For framing detection, leading systems achieved 0.65-0.71 for English, but lower for languages like Russian and Italian, due to data limitations. Similarly, persuasion technique detection proved challenging, with micro F1 scores ranging from 0.45 to 0.58, highlighting difficulties in identifying less frequent rhetorical devices like Appeal to Popularity. [5]

In CheckThat! 2024, span-level annotation across five languages showed even more variability, with average F1 scores around 0.50-0.55 for English and Portuguese, and lower for resource-scarce languages like Bulgarian. Challenges in cross-lingual consistency and inter-annotator agreement (IAA) were common, particularly in non-English datasets, where IAA scores ranged from 0.20 to 0.30, well below the recommended 0.667 threshold. [5]

Overall, narrative classification techniques have seen substantial improvements, particularly for broader tasks like genre categorization. However, fine-grained tasks like persuasion detection still face obstacles, with performance variability due to data scarcity, annotation challenges, and cross-lingual inconsistencies. Transformer models, multi-task learning, and data augmentation have been instrumental in addressing these challenges, but more work is needed to ensure accuracy across languages and narrative nuances.

C. Subtask 3 - Narrative Extraction

This subtask focuses on automatically generating concise explanations grounded in textual evidence to justify dominant narratives, involving both text generation and linguistic simplification. Recent efforts in narrative classification, propaganda detection, and persuasion technique identification have leveraged state-of-the-art transformer models and various augmentation techniques.

- Overview of the CLEF 2024 SimpleText Track: The paper highlights several models used for text generation, including LLaMA , GPT-3.5, and Mistral, applied for text simplification and explanation tasks. Techniques like prompt engineering and reinforcement learning were employed to enhance the performance of these models in generating coherent and contextually relevant outputs (LSTM models were also utilized for sentence-level generation). Performance metrics included SARI (for evaluating text simplification), BLEU (for text overlap with references), and readability scores like FKGL. The top teams (e.g., AIIRLab, Sharigans) achieved high precision in some tasks, but ran into issues like spurious content (hallucinations) across tasks.

  For our Subtask 3 on Narrative Extraction, the overlap with text generation tasks is significant, especially regarding the generation of grounded explanations. Techniques that ensure content alignment with source text, such as fine-tuned models and prompt engineering, will be critical in generating narrative explanations without hallucinations [6].
- Team Sharingans at SimpleText: Text Simplification via Fine-Tuned Large Language Models: The SimpleText task focused on simplifying scientific texts for accessibility using large language models (LLMs). The approach by Team Sharingans involved fine-tuning GPT-3.5 Turbo for text simplification, including prompt-engineering techniques to ensure clarity and fidelity to original content. By applying zero-shot and few-shot learning, their model handled both sentence

and document-level simplifications. The use of GPT-3.5 showcased the capacity of LLMs to generate simplified yet coherent summaries while maintaining a high level of factual accuracy [7].

V. Approaches

A. Subtask 1 - Entity Framings

This subtask focused on developing a model capable of classifying mentioned entities in news articles related to topics such as the Ukraine-Russia War, climate change, and other domains. The goal was to assign each entity a main role (e.g., protagonist, antagonist, innocent) and a fine-grained role. Notably, the fine-grained role formed the primary basis for evaluation, making it the most critical aspect of the task. Below, we outline our methodology, including data preparation, model selection, training strategies, and evaluation techniques.

Data Preparation: The dataset for this subtask comprised news articles annotated with the start and stop indices of each mentioned entity, the entity's name, its main role, and its fine-grained roles. Articles often contained multiple entities, and in some cases, the same entity appeared multiple times within an article, assigned different fine-grained roles based on the context. Considering that the data has not been fully released, our experiments rely on the currently available training and development sets. To prepare the available data, the start and stop indices were used to extract relevant sentences, comprising 300 characters before and 300 characters after each entity mention. This ensured the creation of contextually relevant input samples for model training. However, the dataset exhibited significant class imbalance among the fine-grained roles.

TABLE IV
Most and Least Occurring Fine-grained Labels

Most Occurring	Count	Least Occurring	Count
Instigator	47	Forgotten	1
Guardian	39	Spy	3
Incompetent	35	Exploited	6
Foreign Adversary	32	Traitor	7
Victim	32	Scapegoat	8

To address this, we performed data augmentation using the **GEMINI API**. The API was provided with the original sentence, the entity, its main role, and its fine-grained role, and it generated contextually similar sentences that embodied the specified main role and fine grained roles. This approach successfully expanded the dataset and mitigated class imbalance, enabling more robust training.

Model Selection and Training: Initially, we experimented with transformer-based models like BERT and DeBERTa, but these models failed to deliver satisfactory results. Subsequently, we fine-tuned BART models, which showed significant improvement. Among them, **BART-CNN** emerged as the best-performing model, achieving the highest evaluation scores.

We also experimented with **Contrastive Loss** to enhance the model's ability to differentiate between similar contexts associated with different roles. However, the inclusion of Contrastive Loss did not yield any notable improvement in performance for this task.

Evaluation: The primary evaluation metric for the task was **Exact Match Ratio (EMR)**, which measured the proportion of instances where both the main role and fine-grained role predictions exactly matched the ground truth. Additionally, precision, recall, and F1-score were computed to assess the overall model performance.

By leveraging data augmentation techniques and fine-tuning BART-CNN, we achieved the best results for the subtask. However, the task remains challenging due to the intricate nature of context-dependent role classification and class imbalance. The team is still waiting for the complete dataset so that we can further explore the complexity of the dataset and look into different techniques and task-specific pretraining to further enhance performance.

B. Subtask 2 - Narrative Classification

Our approach for Subtask 2 focused on developing a model that could classify news articles related to topics such as the Ukraine-Russia War, climate change, or other domains, into their corresponding narratives and subnarratives. This hierarchical multi-label classification task was achieved by fine-tuning pre-trained language models on a task-specific dataset and evaluating performance using a variety of metrics. Below, we detail our methodology, including data preparation, model selection, training strategies, and evaluation techniques.

Data Preparation: The dataset for this subtask comprised news articles annotated with dominant narratives, subnarratives, and explanations. As of December 4, 2024, the dataset is being released incrementally; hence, our experiments were conducted using the available training and development sets. To ensure consistency and accuracy, the gold-standard explanations in the training data were preprocessed to standardize formats and tokenize content effectively. This preprocessing aligned the data with the input requirements of transformer-based language models used in our experiments. However, due to the limited size of the dataset and the extensive range of narratives and subnarratives, it was imperative to employ data augmentation techniques to enhance model training and improve generalization.

To address the data scarcity challenge, we utilized back-translation as a primary augmentation method. Articles in the dataset were translated from English to other languages, such as Hindi and Belgian, and then translated back to English. This process effectively modified sentence structures, altered word order, and introduced synonyms, thereby diversifying the textual representations without changing the original meaning.

For instance: An article was translated from English to Hindi and then back to English. Similarly, another article was translated from English to Belgian and then back to English. These transformations introduced subtle variations in phrasing and vocabulary while preserving the underlying narratives and subnarratives, enriching the dataset with linguistically diverse examples.

Model Selection and Training: We experimented with several state-of-the-art transformer-based models, including BERT, GPT-2, Flan-T5, and BART. The models were fine-tuned on the annotated dataset with the specific objective of generating explanations that connected the article text to the assigned narratives and subnarratives. Among the models tested, BERT consistently delivered the best performance in terms of both quantitative and qualitative metrics. Consequently, we adopted a hierarchical approach leveraging multiple BERT models for this task.

The hierarchical pipeline consisted of the following components:

1) **Group Prediction Model:** The first model predicted the group of the article, which could be either Ukraine-Russia War, Climate Change, or Other
2) **Narrative Prediction Models:** Separate models were trained for narrative prediction within each group. One model specialized in the narratives for Ukraine-Russia War articles, while another focused on Climate Change.
3) **Subnarrative Prediction Models:** For each group, another layer of models predicted subnarratives based on the narrative classification. Two predefined taxonomies, structured as dictionaries, guided this stage. Each narrative served as a key, with its associated subnarratives as items.

In this multi-step process, an article first passed through the group prediction model. Based on the assigned group, it was routed to the appropriate narrative prediction model. Subsequently, the predicted narrative determined which subnarrative model would be applied, ensuring a streamlined and context-specific classification.

Evaluation: Model performance was evaluated using a combination of quantitative and qualitative metrics to ensure accuracy. The primary evaluation metrics were **Precision, Recall, and Macro F1 Score.** These metrics provided insights into the models' correctness, completeness, and overall balance in generating coherent and accurate classifications.

By leveraging a filtered hierarchical approach, our methodology ensured efficient handling of domain-specific taxonomies while maintaining high accuracy in narrative and subnarrative predictions. This modular framework is adaptable to the evolving dataset and can incorporate new narratives and subnarratives with minimal overhead.

C. Subtask 3 - Narrative Extraction

For Subtask 3, our approach focused on developing a models that was capable of generating concise explanations grounded in textual evidence to justify the classification of a dominant narratives within news articles. The process involved fine-tuning pre-trained language models on the task-specific dataset and evaluating their performance using a variety of metrics. Below, we detail the steps taken, including data handling, model selection, training strategies, and evaluation methodologies.

Data Preparation: The dataset for this subtask consisted of news articles annotated with dominant narratives, subnarratives, and explanations. As the dataset is still being released incrementally, our experiments were conducted using the available training and development sets as of 4th December, 2024. The training data included gold-standard explanations, which to ensure consistency and accuracy, we preprocessed the text to standardize formats, and tokenize content effectively, structured to align with the input requirements of the transformer-based language models used in our experiments.

Model Selection and Training: We experimented with several modern language models, including variants of BART, GPT-2, and Flan-T5. Each model was fine-tuned using the annotated training dataset, focusing on the task of generating explanations that connect the article text to the assigned narratives.

- BART-CNN Large and BART Large: Both versions of BART were trained on the summarization task and proved that they are well-suited for generating coherent and contextually rich text. BART-CNN Large achieved the best overall performance in our experiments, excelling in capturing narrative context and producing concise, accurate explanations while BART Large showed slightly different behavior, with a slight tradeoff between precision and recall but a larger tradeoff in coherence and semantic structure of sentences.
- GPT-2: Although GPT-2 is a versatile model for text generation, it struggled with maintaining relevance in the generated explanation, and this limitation was evident in its lower F1 scores compared to BART-based models.
- Flan-T5: Flan-T5 exhibited moderate performance. While it showed potential in certain cases, its explanations were less detailed and contextually grounded than those generated by BART models and GPT-2.
- LLaMA 3.2 1b: Due to hardware limitations, our experiments with LLaMA models encountered significant challenges, including frequent CUDA out-of-memory errors which prompted us to switch from working on Google Colab to Kaggle to the universities GPU but to no avail. As a result, comprehensive evaluation of LLaMA was not feasible.

Evaluation: Model performance was evaluated using BERTScore. This metric directly aligned with the task's requirements and leaderboard evaluation criteria. BERTScore computes similarity between machine-generated explanations and gold-standard references using contextual embeddings derived from pre-trained BERT models, evaluating text generation quality based on token-level cosine similarity while considering both semantic and contextual relevance. This metric was also used in training and fine-tuning the models instead of seeking lowest loss only.

VI. Results and Discussion

A. Subtask 1 - Entity Framings

TABLE V
EMR Scores of Tested Models

Model	Exact Match Ratio
Baseline	0.12090
BART-CNN	**0.24180**
DistilBERT-base-uncased	0.13190
BERT-base-uncased	0.12090
BART-Large	0.21980

Among the tested models, BART (CNN) achieved the highest EMR of 0.24180, showcasing its superior ability to capture the context of the sentences and handle the framing of entities in manipulative narratives. This could be because the model was pretrained on a large dataset of CNN articles that do reflect on the Ukraine-Russia War and Climate Change. BART-Large also performed well with an EMR score of 0.21980, slightly below BART (CNN). In contrast, BERT-base-uncased significantly underperformed with an EMR score of 0.12090, likely due to its inability to capture multilingual or narrative-specific nuances effectively. DistilBERT-base-uncased, while marginally better than BERT-base-uncased with an EMR score of 0.13190, still demonstrated inadequate performance for this task. Moreover, experiments with Contrastive Loss failed to yield any prominent improvements, suggesting that the loss function may not have been optimally configured or that the task-specific nuances were insufficiently addressed.

B. Subtask 2 - Narrative Classification

TABLE VI
F1 Scores of Tested Models

Model	F1 macro fine	F1 st. dev. fine
BERT Base	0.17100	0.37600
Baseline	0.00700	0.04500

The baseline performance for the narrative classification task was exceptionally low, with a Macro F1 score of **0.007**. This highlighted the significant challenge posed by the task, which involved a small training dataset and a wide range of narratives and subnarratives. Despite these difficulties, our data augmentation strategies and hierarchical modeling approach substantially improved the performance, achieving a final Macro F1 score of **0.171**.

Improvement Over Baseline: The substantial improvement from the baseline score underscores the effectiveness of our techniques, particularly data augmentation via back-translation and the hierarchical classification framework. These methods allowed the model to better handle the diversity and granularity of the task.

Performance Variations Across Groups:

- **Ukraine-Russia War Articles:** The model performed noticeably better on *Ukraine-Russia War* articles compared to *Climate Change*.
- **Reason for Variation:** This disparity can be attributed to the composition of the training dataset, which contained a significantly larger proportion of articles about the Ukraine-Russia War. Consequently, the model had more examples to learn from, resulting in improved predictions for this category.
- **Climate Change Articles:** In contrast, the smaller representation of *Climate Change* articles limited the model's ability to generalize effectively, leading to relatively lower performance in this domain.

The results highlight the importance of data quantity and diversity in training robust classification models for complex tasks involving extensive taxonomies. While our techniques mitigated some of the challenges posed by data scarcity, they also revealed the limitations of imbalanced datasets. These findings emphasize the need for further dataset expansion and targeted data augmentation, particularly for underrepresented categories like *Climate Change*.

This study demonstrates the potential for substantial improvement in low-resource narrative classification tasks through thoughtful preprocessing, augmentation, and modeling strategies. Future work could focus on incorporating domain-specific pretraining or transfer learning to further enhance performance, especially for less-represented domains.

C. Subtask 3 - Narrative Extraction

For the narrative extraction task, we went with fine-tuned the facebook bart-large-cnn model using prompts structured as: "Based on the following narrative [DOMINANT NARRATIVE], [DOMINANT SUB-NARRATIVE], find the summary of the article [ARTICLE TEXT]." With this explicit guidance integrating both the dominant narrative and sub-narrative (when available) with the article text, the pre-trained model and tokenizer were adapted for the task, truncating inputs to fit token limits while retaining essential context. Fine-tuning utilized a learning rate of 2e-5 with a warmup period of 10 steps, training for 7 epochs with a batch size of 4 and the performance was evaluated using BERTScore F1 between the generated outputs and gold-standard references. The model's peak BERTScore F1 determined the best configuration. After training, the fine-tuned model was applied to unseen data, generating narrative explanations through beam search.

TABLE VII
BERTScore of Tested Models

Model	Percesion	Recall	F1 Score
Baseline	0.65540	0.67957	0.66719
BART-CNN	0.7286	0.7488	**0.7385**
BART Large	0.76180	0.69615	0.72723
GPT-2	0.5854	0.6964	0.6360
Flan-T5	0.6727	0.6217	0.6456

Among the tested models, BART (CNN) demonstrated the strongest performance, with an F1 score of 0.7385, precision of 0.7286, and recall of 0.7488. These results highlighted its capability to effectively capture contextual information for text generation. BART Large followed closely, achieving slightly lower but comparable scores. In contrast, GPT-2 and Flan-T5 underperformed, with F1 scores of 0.6360 and 0.6456, respectively. These models exhibited difficulties in generating coherent and contextually grounded explanations, particularly in scenarios requiring nuanced understanding of narratives.

VII. Conclusion

The results demonstrate the effectiveness of BART-based models for subtasks 1 and 3, particularly in capturing the framing context of entities in manipulative narratives. Further exploration of advanced fine-tuning techniques and hyperparameter optimization is necessary to enhance the performance of transformer models.

One major challenge was the limited availability of computational resources, which hindered experimentation with more resource-intensive models like LLaMA. Additionally, the incremental release of data restricted our ability to train models on a fully representative dataset. The complete dataset is expected to be released by the end of the first week of December, and we hope to use the complete dataset and mould our approaches on the basis of that.

References

[1] P. Heinisch, M. Plenz, A. Frank, and P. Cimiano, ”ACCEPT at SemEval-2023 Task 3: An Ensemble-based Approach to Multilingual Framing Detection,” in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023, pp. 1358–1365. [Online]. Available: https://aclanthology.org/2023.semeval-1.187

[2] Q. Liao, M. Lai, and P. Nakov, ”MarsEclipse at SemEval-2023 Task 3: Multi-lingual and Multi-label Framing Detection with Contrastive Learning,” in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023, pp. 84–90. [Online]. Available: https://aclanthology.org/2023.semeval-1.10

[3] S. Henning, W. Beluch, A. Fraser, and A. Friedrich, “A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing,” Jan. 2023, *doi:* https://doi.org/10.18653/v1/2023.eacl-main.38.

[4] Da San Martino, Giovanni, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. ”SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles.” In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, pages 1377–1414, Barcelona, Spain (Online), December 12, 2020.

[5] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, Preslav Nakov, *SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multilingual Setup*, Institute of Computer Science, Polish Academy of Science, Poland, European Commission Joint Research Centre, Italy, Department of Mathematics, University of Padova, Italy, Mohamed bin Zayed University of Artificial Intelligence, UAE, 2023.

[6] Liana Ermakova, Eric SanJuan, Stéphane Huet, Hosein Azarbonyad, Giorgio Maria Di Nunzio, Federica Vezzani, Jennifer D’Souza, Jaap Kamps, *Overview of the CLEF 2024 SimpleText Track: Improving Access to Scientific Texts for Everyone*, Université de Bretagne Occidentale, France; Avignon Université, France; Elsevier, Netherlands; University of Padua, Italy; Leibniz Information Centre for Science and Technology, Germany; University of Amsterdam, Netherlands, 2024.

[7] Syed Muhammad Ali, Hammad Sajid, Owais Aijaz, Owais Waheed, Faisal Alvi, Abdul Samad, *Team Sharingans at SimpleText: Fine-Tuned LLM based approach to Scientific Text Simplification*, Computer Science Program, Dhanani School of Science and Engineering, Habib University, Karachi-75290, Pakistan, 2024.