NarrativeMiners at SemEval-2025 Task 10: Combating Manipulative Narratives in Online News

Muhammad Khubaib[†*], Muhammad Shoaib Khursheed[†], Muminah Khurram[†]

Sandesh Kumar[†], Abdul Samad[†]
[†]Habib University, Karachi, Pakistan
mk07218@st.habib.edu.pk, mk07149@st.habib.edu.pk, mk07521@st.habib.edu.pk,
sandesh.kumar@sse.habib.edu.pk, abdul.samad@sse.habib.edu.pk

Abstract

Harmful disinformation and propaganda proliferate at unprecedented rates, highlighting the need for effective detection and analysis methods. Identifying and analyzing manipulative narratives in online news is critical to mitigate their impact on public opinion. This paper addresses SemEval-2025 Task 10 on "Multilingual Characterization and Extraction of Narratives from Online News" Piskorski et al. (2025) by focusing on three subtasks that revolve around classifying entities, categorizing news articles into narratives and subnarratives, and generating concise summaries for a given article.

We have employed various deep learning techniques in multilingual settings to tackle these challenges. Our results demonstrate the effectiveness of BART-based models in capturing the framing context of entities and generating narrative-focused summaries, ultimately offering insights into the dynamics of online narratives and contributing to efforts against harmful disinformation.

1 Introduction

Misinformation in online news has become an increasingly urgent concern. Manipulative articles can sway public opinion, exacerbate crises, and compromise the reliability of digital content. Detecting, classifying, and explaining harmful narratives is therefore a vital step toward combating disinformation. Recent advances in machine learning, especially large language models, have made it possible to automate these tasks at scale. Our project contributes to the development of these tools to combat the spread of misleading information by providing a deeper understanding of how narratives are constructed and used to shape public discourse.

However, the complexity and variety of narratives pose substantial challenges. Articles covering major geopolitical events (e.g., the Ukraine-Russia war) or global issues (e.g., climate change) often embed many subtle manipulative cues within lengthy text. Accurately identifying the entity roles, dominant narratives, and subnarratives at play requires a nuanced understanding of context. In this work, we present our approach for SemEval-2025 Task 10, where we focus on three subtasks: entity framing (Subtask 1), narrative classification (Subtask 2), and narrative extraction and summarization (Subtask 3). Our methods leverage recent transformer-based architectures, together with selective data augmentation, to manage real-world complexities such as imbalanced labels and limited high-quality training data.

2 Research Question

1. Entity Framing: How can entities in news articles be accurately classified according to their roles within manipulative narratives?

2. Narrative Classification: What methods effectively categorize articles into dominant narratives and subnarratives, given the variability in topic?

3. Narrative Extraction: How can we generate concise, evidence-based summaries that highlight manipulative narratives within articles?

3 Dataset

We use the official multilingual dataset from SemEval-2025 Task 10, which comprises approximately 700 news articles in five languages: Russian, English, Hindi, Bulgarian, and Portuguese. Annotations for this include entity mentions (with roles), narrative labels, and short textual explanations. Our primary focus is on the ~200 English articles, though we also made partial use of the multilingual data for augmentation and validation.

[*]Corresponding author

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1639–1644
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

4 Literature Review

4.1 Subtask 1 - Entity Framings

At SemEval-2023 Task 3, Heinisch et al. (Heinisch et al., 2023) (Team ACCEPT) combined Large Language Models, static embeddings, and commonsense knowledge from ConceptNet within a Graph Neural Network framework. Their fine-tuned RoBERTa model achieved strong results in English framing detection (F1: 50.69% micro, 50.20% macro), highlighting the value of external knowledge integration. In contrast, Liao et al. (Liao et al., 2023) (Team MarsEclipse) applied contrastive learning using a dual-input XLM-RoBERTa architecture (title + body), clustering similar frames while separating dissimilar ones. Treating the task as 14 binary problems, they optimized thresholds to achieve top F1 scores across multiple languages, including German (0.711) and Polish (0.673).

4.2 Subtask 2 - Narrative Classification

Prior shared tasks have laid important groundwork for narrative and persuasion analysis. SemEval-2020 Task 11 tackled propaganda detection by identifying spans and classifying 14 techniques, achieving F1-scores of 51.55 for span identification and 62.07 for classification (Martino et al., 2020). Data augmentation proved especially helpful for rare techniques like *Whataboutism*. SemEval-2023 Task 3 expanded narrative classification to nine languages, with genre categorization yielding strong macro F1 scores (0.78–0.85 for English) and framing detection peaking at 0.71 (Piskorski et al., 2023).

However, detecting persuasion techniques remained challenging, particularly in low-resource settings. Similarly, CheckThat! 2024 introduced span-level annotations across five languages, but performance varied: while English and Portuguese achieved F1 scores around 0.50–0.55, inter-annotator agreement (IAA) remained low for under-resourced languages (IAA: 0.20–0.30), far below the recommended 0.667 threshold (Ermakova et al., 2024).

4.3 Subtask 3 - Narrative Extraction

In the CLEF 2024 SimpleText track, several teams explored the use of advanced language models such as LLaMA, GPT-3.5, and Mistral for scientific text simplification and explanation tasks (Ermakova et al., 2024). Common strategies included prompt engineering and reinforcement learning, with BLEU emerging as a key evaluation metric. Despite high precision, challenges like hallucinations persisted, even among top-performing teams such as AIIRLab and Sharingans. The latter, as detailed by Ali et al. (Ali et al., 2024), fine-tuned GPT-3.5 Turbo using zero-shot and few-shot learning to enhance clarity while preserving factual integrity. Their use of carefully crafted prompts demonstrated the potential of LLMs to produce coherent, faithful simplifications—insights that directly inform our approach to generating grounded narrative explanations.

5 Approaches

5.1 Data Augmentation and Preparation

To ensure consistency and enable a unified training process, we translated all non-English narrative datasets into English. Specifically, 211 Bulgarian, 115 Hindi, and 200 Portuguese files were translated using the Google Translate API. This step consolidated multilingual resources into a single English-language dataset of 726 files (including 200 original English files). While automated translation may introduce minor semantic drift, it allowed for scalable integration of multilingual resources into our English-centric narrative classification model (Table 1).

Original Language	**Files**	**Method**
Bulgarian	211	Google Translate
Hindi	115	Google Translate
Portuguese	200	Google Translate
English (original)	200	-
Total Files	**726**	**Unified**

Table 1: Translation of non-English narrative datasets into English using Google Translate API

To increase dataset diversity and improve generalization, we applied back translation to 399 English samples—translating into Bulgarian, Portuguese, Hindi, and Russian before converting back to English—resulting in paraphrased variants that preserved meaning while introducing linguistic variability. Combining translation, back-translation, and augmentation resulted in a final dataset of 1125 English-language samples used across all tasks (Table 2).

We also applied label-aware augmentation to tackle class imbalance in Subtask 1. As shown in

Source	Files
Translated	726
Back-Translated Augmented	399
Total Files in Final Dataset	**1125**

Table 2: Final composition of the English dataset combining translated and augmented files

Table 3, entity framing labels were highly skewed. To mitigate this, we used the **GEMINI API** to generate semantically similar sentences for underrepresented roles, and **Mistral** to generate contextual variations, enhancing model exposure to diverse textual patterns.

Table 3: Variance of Fine-grained Labels in Initial Training Data

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

5.2 Subtask 1 - Entity Framings

This subtask required classification of entities into fine-grained framing roles, such as *Instigator*, *Victim*, and *Guardian*, within complex narrative contexts. Although initial translations also included Bulgarian samples in sub task 1, they were excluded after empirical evaluations showed reduced performance. During data preprocessing, we standardized input structure by extracting 200 characters of surrounding context, then generated a **Prompt** column by concatenating the context and the entity. Using the augmented dataset (**8,900 samples**), we had a 90% - 10% training-test split.

We begun experimenting 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.

The following **Training Configuration** was set up with the Key Hyperparameters. We trained for **6 epochs** to balance learning and overfitting, using a **batch size of 16** (on A100/T4 GPUs). **Mixed precision (fp16=True)** was enabled for efficiency, and models were evaluated per epoch to track accuracy.

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.

5.3 Subtask 2 - Narrative Classification

Our approach for Subtask 2 focused on hierarchical multi-label classification of news articles into narratives and subnarratives within domains such as the Ukraine-Russia War (URW) and climate change (CC). Given the complexity of the task, we designed a structured classification pipeline that progressively refined predictions across multiple levels. The goal was to enhance classification accuracy while ensuring contextual relevance.

We implemented a structured classification pipeline using five fine-tuned BERT models, each handling a specific stage of the classification process:

1. Topic Classification: The first model categorized articles into three broad groups: URW, CC, or Other. If classified as "Other," the article was assigned generic labels, and no further classification was required.

2. Narrative Classification: Articles labeled as URW or CC were passed to a dedicated narrative classification model trained on domain-specific data. This means that URW was trained on just URW data and CC was trained on just CC data by the specific invidual models. This ensured focused learning and reduced cross-topic interference.

3. Subnarrative Classification: After predicting the narrative, a subnarrative classification model assigned the most relevant label from a predefined set, filtering out unrelated subnarratives.

This multi-tiered approach progressively narrowed down the label space at each stage, improving classification precision while optimizing training efficiency by ensuring models only learned from domain-relevant data.

5.4 Subtask 3 – Narrative Extraction

This subtask requires generating concise, evidence-based explanations that justify the dominant narrative and sub-narrative labels in news articles. Our dataset comprises articles annotated with narrative labels, sub-narratives, and gold-standard explanations.

Early attempts at **data augmentation**—using the Gemini API for synthetic explanations and multilingual back-translation—failed to boost performance (F1 plateaued at 0.71 or declined), so we proceeded without synthetic samples.

We fine-tuned **four transformer variants** on the training split: **BART-CNN Large**, **BART Large**, **GPT-2**, and **Flan-T5** An attempt to fine-tune LLaMA 3.2 1B was infeasible due to persistent memory constraints.

Evaluation. We used BERTSCORE to measure token-level semantic similarity between generated and reference explanations, ensuring both accuracy and contextual relevance. BART-CNN Large consistently outperformed other models, confirming its suitability for narrative extraction tasks.

6 Results and Discussion

6.1 Subtask 1 - Entity Framings

Table 4: Exact Match Ratio (EMR) Scores of Tested Models on Dev Set

Model	EMR
Baseline	0.1209
BART-CNN	**0.3407**
DistilBERT-base-uncased	0.1319
BERT-base-uncased	0.1209
BART-Large	0.2198

BART-CNN outperformed other models with an EMR of 0.3407 on the dev set, likely benefiting from pretraining on CNN articles aligned with task domains. **BART-Large** followed with 0.2198, while **BERT-based models** underperformed, lacking sufficient narrative and framing awareness.

Using **contrastive loss** did not improve results, likely due to suboptimal configuration or the task's nuanced semantics.

Generalization Issues: BART-CNN's test EMR dropped to **0.2128** (baseline: 0.0383), highlighting generalization challenges which may have stemmed from the Augmentation Noise as the synthetic data could have caused label drift, not being able to control the class imbalance effectively. Overfitting on the Dev set patterns may not have allow the model to generalize well for the test set.

Despite strong dev set results as highlighted by Table 9, test-time robustness remains a key challenge. Improvements may come from better class balancing, and cleaner augmentation which may improve our standings. Currently we rank 13/32 teams.

Table 5: EMR Scores vs. Baseline

Dataset	Baseline	BART-CNN	Gain
Dev Set	0.1209	0.3407	**2.82**×
Test Set	0.0383	0.2128	**5.56**×

6.2 Subtask 2 - Narrative Classification

Table 6: F1 Scores of Tested Models

Model	F1 macro fine	F1 st. dev. fine
BERT Base	0.24600	0.4100
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 **0.246**. We scored 17/27 in the final test evaluation conducted by SemEval.

Table 7: F1 scores on dev and test set

Dataset	Baseline	F1 macro fine
Dev Set	0.10	**0.246**
Test Set	0.007	**0.246**

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:

- **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*.

6.3 Subtask 3 - Narrative Extraction

We fine-tuned the Facebook BART-Large-CNN model to extract narrative summaries by providing explicit guidance on the dominant narrative (or sub-narrative, when available) alongside the full article text. Specifically, we used prompts of the form: "Based on the following narrative [DOMINANT NARRATIVE]/[DOMINANT SUB-NARRATIVE], find the summary of the article: [ARTICLE TEXT]."

To adapt the pretrained model and tokenizer for this task, inputs were truncated to respect the token limit while preserving critical context. Fine-tuning was performed with a learning rate of 2×10^{-5}, a linear warmup over 10 steps, and training for 7 epochs with a batch size of 4. Model performance was evaluated using BERTScore F1 against gold-standard summaries, selecting the best checkpoint by peak F1 score. Summary generation on unseen data employed beam search decoding. As shown in Table 8, among the tested models, BART-Large-CNN outperformed all other models. It achieved an F1 score of 0.8134 (precision 0.7949, recall 0.8332), highlighting its capability to effectively capture contextual information for text generation. BART-Large achieved the next best performance with a comparable F1 score of 0.7272, while GPT-2 and Flan-T5 lagged behind with F1 scores of 0.6360 and 0.6456, respectively, indicating challenges in generating coherent, contextually grounded narrative summaries, particularly in scenarios requiring nuanced understanding of said narratives.

Table 8: BertScore of Tested Models

Model	Precision	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

Table 9: Comparison of F1 Scores with Baseline

Dataset	**Baseline**	**BART-CNN**
Dev Set	0.66719	**0.81339**
Test Set	0.6669	**0.7291**

The fine-tuned BART-Large-CNN model consistently outperformed the baseline on both the development and test sets (Table 9). This improvement is primarily attributed to rigorous data cleaning, which yielded syntactically improved inputs and enabled the model to better understand and capture narrative context. Notably, all BART variants surpassed the baseline, highlighting the advantage of leveraging pretrained transformer models even with limited data. This strength was further felt when, despite having even less data, our best model had achieved a 0.7385 F1 score on the initial smaller development set.

Additionally, our findings indicate that the primary driver of further improvement would be more high-quality, real training data, rather than translated or augmented data, as both of these approaches worsened the performance. The addition of well-annotated, real-world data would likely yield even greater improvements in the models accuracy and robustness.

7 Limitations

7.1 Subtask 1 - Entity Framings

Our fine-grained framing labels were heavily skewed (e.g. Instigator vs. Forgotten), which biased the model toward majority classes. We used paraphrasing (Gemini, Mistral) and cross-lingual translations to boost minority classes, but improvements were modest and sometimes introduced label noise.

7.2 Subtask 2 - Narrative Classification

One of the major challenges we faced in this task was due to the lack of the dataset for the two-level taxonomy of the classification problem. There were a total of approximately 74 narrative subclasses, with each top-level narrative category having around three subclasses on average. This high granularity, combined with limited training samples per subclass, made it extremely challenging for the model to learn meaningful patterns across all categories.

We attempted to mitigate this issue through data augmentation techniques, such as back translation, which showed some improvement in subclass classification. However, the effectiveness of augmentation plateaued beyond a certain point, indicating the need for either more labeled data, better hierarchical modeling, or external knowledge sources to truly improve subclass performance.

7.3 Subtask 3 - Narrative Extraction

Due to limited and imbalanced data availability, we experimented with data translation using the Google Translate API without rigorous post-editing or quality control which may have introduced semantic drift or subtle distortions in meaning. Our results with the synthetically generated and translated samples showed no significant performance improvements and, in some cases, even degraded results. Ultimately, synthetic explanations generated via large language models (e.g., Gemini) were excluded due to inconsistent tone and factual inaccuracies.

8 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 small size of the dataset restricted our ability to train models on a fully representative dataset that was large enough to capture all the nuances. Because of the serious under representation of some classes we did try training with augmented data where possible, but having more real data would have significantly improved performances.

References

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

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

P. Heinisch, M. Plenz, A. Frank, and P. Cimiano. 2023. 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)*, pages 1358–1365.

Q. Liao, M. Lai, and P. Nakov. 2023. 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)*, pages 84–90.

Giovanni Da San Martino, 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).

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Guimarães, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

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