IRNLP at SemEval-2025 Task 10: Multilingual Narrative Characterization and Classification

Panagiotis Kiousis
Athens University of Economics and Business
and Archimedes, Athena Research Center, Greece
pckiousis@gmail.com

Abstract

This paper presents the IRNLP system for Subtask 2 of SemEval-2025 Task 10, which addresses multilingual narrative classification. The approach utilizes datasets in Hindi, English, and Russian, applying transformer-based models fine-tuned through repeated stratified k-fold validation. The system performs joint detection of narratives and subnarratives using multi-label classification techniques. Extensive ablation studies, in-depth error analysis, and a detailed discussion of model architecture and training procedures are included. The implementation is publicly available [1] to support reproducibility and future research.

1 Introduction

Narratives play a pivotal role in shaping public opinion and framing news reporting, often embedding persuasive messaging or ideological intent. Automatically detecting such narratives is a complex task, particularly in multilingual settings, due to semantic ambiguity, class imbalance, and cross-linguistic variability. Subtask 2 of SemEval-2025 Task 10 addresses this challenge by focusing on the classification of narratives and subnarratives in news articles across five languages.

This paper presents the IRNLP system, developed to address this challenge using multilingual transformer-based models. The system was trained on Hindi, English, and Russian datasets, leveraging repeated stratified k-fold validation to ensure robust evaluation. Unlike standard approaches that rely on single-split validation, the use of repeated k-fold increases generalizability and minimizes overfitting. This work also contributes insights through error analysis and controlled ablation studies.

2 Background and Task Overview

SemEval-2025 Task 10 includes three subtasks; Subtask 2, addressed in this paper, requires identifying the presence of narrative and subnarrative categories in online news articles in five languages: English, Hindi, Russian, Portuguese, and Bulgarian. The task is structured as a multi-label classification problem. Each article may belong to multiple coarse- or fine-grained narrative categories. Models are evaluated using both macro F1-score and F1 samples to capture performance across both label and instance levels.

3 Related Work

Previous work on fine-grained propaganda detection by Da San Martino et al. (2019) introduced structured annotation strategies for identifying persuasive techniques. This laid the groundwork for related tasks such as narrative extraction and classification. Multilingual transformer models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLM-RoBERTa (Conneau et al., 2019) have demonstrated strong performance across tasks such as sentiment analysis, named entity recognition, and text classification. Hugging Face's Transformers library (Wolf et al., 2020) provides scalable implementations of these models and facilitates multilingual fine-tuning.

Few studies have focused explicitly on narrative modeling in multilingual contexts. Work in stance detection and argument mining has highlighted the importance of modeling discursive structures, but the integration of coarse and fine narrative labels in low-resource settings remains under-explored. The IRNLP system aims to fill this gap using repeated validation and tailored preprocessing for each language.

[1] https://github.com/ipanos7/Semeval-Task10-English.git

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 54–57
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

4 System Description

4.1 Preprocessing and Data Preparation

Each dataset underwent language-specific preprocessing. Raw articles were parsed, and narrative annotations were mapped to binary multi-label vectors. Tokenization was handled using the pretrained tokenizer of each transformer backbone. Language-specific augmentations were applied when necessary, including sentence shuffling and synonym replacement. Heuristics were used to correct missing or uncertain labels when metadata provided indirect signals.

4.2 Model Architecture

The IRNLP system used XLM-RoBERTa-base and XLM-RoBERTa-large as the primary backbones. In language-specific experiments, NeuralMind BERT was used for Portuguese and DeepPavlov BERT for Bulgarian. A dense output layer with sigmoid activation computed logits for each narrative label. Loss was calculated using binary cross-entropy.

4.3 Training Strategy

To increase generalization, we adopted a repeated stratified k-fold validation strategy (5 folds, 2 repetitions). This approach allowed each data sample to appear in multiple training and validation splits. Training was conducted on NVIDIA A100 GPU with FP16 precision, using AdamW optimizer (learning rate 5e-5, weight decay 0.01). Gradient accumulation was used to simulate larger batch sizes.

5 Ablation Study

The impact of major design choices was quantified in ablation experiments. Table 1 presents comparative F1 samples for English and Hindi.

Variant	English (F1)	Hindi (F1)
XLM-R Large (baseline)	0.287	0.515
No k-fold validation	0.238	0.472
Unbalanced batches	0.245	0.489

Table 1: Ablation results: F1 samples for key variants.

The ablation results highlight the importance of repeated k-fold validation in preventing overfitting. Removing this step led to a drop in F1 samples (from 0.515 to 0.472 in Hindi). Similarly, unbalanced mini-batches had a negative impact, likely due to dominance of majority labels during optimization. These findings confirm that both evaluation strategy and training stability significantly influence model effectiveness in low-resource settings.

6 Experiments and Results

6.1 Evaluation Metrics

The task uses two main metrics: macro F1-score and F1 samples. While macro F1 considers label-level performance, F1 samples captures instance-level accuracy and is prioritized for Subtask 2.

6.2 Performance Comparison

Language	F1 macro	Macro SD	F1 samples	Sample SD
English	0.516	0.402	0.287	0.452
Hindi	0.375	0.467	0.515	0.500
Russian	0.537	0.351	0.116	0.252

Table 2: Performance on test sets.

6.3 Overall Observations

The IRNLP system consistently outperformed the baseline across all three evaluated languages—Hindi, English, and Russian—in both macro F1 (coarse) and F1 samples metrics. The standard deviations further revealed the variability in performance, offering insights into the model's stability. The largest gains were observed in Hindi (F1 samples: 0.515), while the Russian dataset, despite achieving the highest macro F1 (0.537), exhibited comparatively lower instance-level accuracy.

6.3.1 Hindi Test Set

- **F1 macro (coarse):** The system achieved a score of 0.375, significantly higher than the baseline's 0.081. This indicates better generalization across coarse-grained narrative categories.

- **F1 samples:** A notable score of 0.515 was obtained, whereas the baseline failed entirely (0.000). This gap highlights the model's strong predictive capacity on the instance level.

- **Standard Deviations:** The model showed higher variability (0.467 vs. 0.260 for macro F1) compared to the baseline, suggesting fluctuations in predictions across samples.

- **Implications:** These results indicate that cross-linguistic patterns in Hindi are effectively captured. However, the relatively high standard deviation suggests that model consistency could benefit from further fine-tuning or ensembling techniques.

6.3.2 English Test Set

- **F1 macro (coarse):** The model achieved 0.516, substantially outperforming the baseline's 0.030.
- **F1 samples:** A score of 0.287 was recorded, compared to the baseline's 0.013, reflecting the system's improved ability to identify subnarratives.
- **Standard Deviations:** The IRNLP system showed a higher standard deviation (0.402) than the baseline (0.127), indicating greater variance possibly due to the complexity of English samples.
- **Implications:** While performance is notably higher than the baseline, the variability suggests potential benefits from additional regularization or calibration.

6.3.3 Russian Test Set

- **F1 macro (coarse):** The model achieved 0.537, a substantial increase from the baseline's 0.065.
- **F1 samples:** A lower score of 0.116 was observed, though still notably above the baseline's 0.008.
- **Standard Deviations:** The system demonstrated the lowest variance in macro F1 (0.351) compared to Hindi and English, suggesting more consistent predictions.
- **Implications:** These results point to strong macro-level performance in Russian. However, lower F1 samples performance indicates that fine-grained instance classification remains an area for improvement.

7 Error Analysis

The most common source of error was label imbalance, which led the model to favor dominant narrative types while underpredicting rare ones. This was particularly evident in the Hindi and Russian datasets, where certain subnarratives appeared infrequently. Additionally, semantic overlap between similar categories—such as *foreign conspiracy* and *global threat*—confused the model, often resulting in misclassification between conceptually adjacent classes.

Another recurring issue stemmed from the multi-label nature of the task. In some cases, the model correctly identified a coarse-grained narrative but failed to capture accompanying subnarratives, reducing F1 samples scores. This was especially noticeable in Russian, where instance-level prediction was more challenging despite strong macro-level performance.

These findings suggest that future iterations of the system may benefit from more balanced sampling strategies, label smoothing, and architectures that better capture inter-label dependencies.

8 Conclusion

This paper presented the IRNLP system for Subtask 2 of SemEval-2025 Task 10. The system combined transformer models with repeated k-fold validation and language-sensitive preprocessing. Results demonstrated robust generalization in multilingual narrative classification. Future directions include incorporating contrastive loss, data augmentation for low-resource languages, and exploring semi-supervised training.

9 Limitations and Future Work

One limitation of the current system is its reliance on supervised data, which restricts performance in languages with fewer labeled examples. The model also assumes static label definitions, which may not generalize to evolving narrative framings in future news content. Additionally, extensive ensembling or hyperparameter search hadn't been performed due to time constraints.

Future work will explore semi-supervised learning techniques such as pseudo-labeling and contrastive learning. It is also planned to investigate cross-lingual transfer methods to improve performance in low-resource settings by leveraging multilingual embeddings and aligned fine-tuning. Finally, interpretability remains an open challenge in narrative classification, and future iterations will incorporate attention visualization to better understand model behavior.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. Special acknoeledgements also to the (Piskorski et al., 2025) SemEval-2025 Task 10 organizers for their support.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.