WordWiz at SemEval-2025 Task 10: Optimizing Narrative Extraction in Multilingual News via Fine-Tuned Language Models

Ruhollah Ahmadi
Department of Computer Engineering
Amirkabir University of Technology
ruhollah@aut.ac.ir

Hossein Zeinali
Department of Computer Engineering
Amirkabir University of Technology
hzeinali@aut.ac.ir

Abstract

This paper presents the WordWiz team's submissions for Task 10 of SemEval 2025: Multilingual Characterization and Extraction of Narratives from Online News. The Narrative Extraction subtask focuses on generating concise explanations that support dominant narratives identified in multilingual news articles. We employ two complementary approaches: supervised fine-tuning and direct preference optimization of large language models. To enhance training data quality, we develop a preprocessing pipeline. Additionally, we implement a multi-temperature inference strategy, which generates three candidate explanations using varying temperature parameters and selects the most relevant one through semantic similarity scoring. Our final system[1] secured first place in Portuguese and second place in English, Russian, Bulgarian, and Hindi, consistently outperforming baseline systems across all languages.

1 Introduction

The democratization of information on the internet has posed significant challenges in discerning and interpreting manipulative content. News ecosystems, particularly during crises, often serve as contested spaces where disinformation and propaganda narratives vie for public credibility and attention. Consequently, the automated identification and characterization of narratives in multilingual news sources constitute a vital endeavor for content moderators, fact-checkers, and media literacy initiatives. The SemEval 2025 Task on Multilingual Characterization and Extraction of Narratives from Online News addresses this issue through three subtasks spanning five languages English, Bulgarian, Hindi, Portuguese, and Russian Piskorski et al. (2025) Stefanovitch et al. (2025).

Our research centers on Subtask 3: Narrative Extraction, which entails producing a concise explanation (maximum 80 words) that substantiates a given dominant narrative using textual evidence from a news article. This text-to-text generation task necessitates both the precise detection of pertinent evidence and a coherent summary that aligns with the narrative classification. Our methodology leverages recent advancements in large language models, employing two complementary approaches. First, we utilized Supervised Fine-Tuning (SFT) Lee (2024) on pre-trained models, including Phi-3.5, Llama-3, and Llama-3.1-8B. Second, we applied Direct Preference Optimization (DPO) Rafailov et al. (2023) to enhance model outputs by incorporating human preference data. To optimize performance across diverse linguistic contexts, we developed a robust preprocessing pipeline that eliminates duplicate sentences, standardizes text, and augments training examples with narrative-specific details. During inference, our system generates multiple candidate explanations by varying temperature parameters and selects the most apt one via semantic similarity scoring.

Experimental findings reveal that, although larger models frequently exhibit robust performance, the meticulously fine-tuned Phi-3.5 model consistently yielded superior results across most languages. We delineate critical factors for effective narrative extraction, including the identification of relevant evidential sentences, the preservation of cross-lingual consistency in explanation structure, and the equilibrium between textual fidelity and succinct summarization. Our system markedly outperforms baseline models across all evaluated languages.

2 Background

SemEval 2025 Task 10 posed a complex information extraction challenge across five languages: English, Bulgarian, Hindi, Portuguese, and Russian. The dataset comprised news articles primarily covering two domains climate change (CC) and the

[1] https://github.com/roohix/WordWiz

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1276–1281

Ukraine-Russia war (UA) with each article annotated for its dominant narrative and, when applicable, sub-narrative classifications. The training set contained a substantial number of multilingual examples, while the development set was comparatively smaller. Both adhered to the same data structure: article text, narrative label, sub-narrative label, and gold-standard explanations. The task presented significant challenges in identifying evidence supporting the assigned narratives, particularly given the substantial variation in article length. The gold-standard explanations highlighted key textual evidence aligning with the assigned narrative without explicitly referencing its classification. Our dataset analysis indicated that effective explanations required capturing the rhetorical strategies, entities, and argumentation patterns characteristic of each narrative type while maintaining linguistic and structural appropriateness for each language.

The task organizers provided a baseline system utilizing zero-shot prompting with the Microsoft Phi-3-small-8k-instruct model. This approach relied exclusively on the model's pre-trained capabilities to interpret prompts and generate explanations, without any fine-tuning.

Research on information and event extraction from news articles remains in its early stages. Sentence-level methods, frequently falling under the umbrella of Timeline Summarization (TLS), typically generate linear sequences representing a single story's progression. Recent variations on TLS aim to capture more complexity, such as comparative timelines highlighting contrasting events between datasets Duan et al. (2020) or Multi-Timeline Summarization (MTLS) which extracts distinct parallel storylines from a corpus Yu et al. (2021). Alternative paradigms like Summarize Dates First reverse the typical TLS pipeline by summarizing individual dates before selecting relevant ones Quatra et al. (2021). Document and cluster-level methods often employ more complex graph structures to represent interactions between multiple storylines and events, moving beyond simple linearity to capture convergences and divergences.

Recent advancements have incorporated diverse techniques and representations. For sentence-level extraction, methods like WILSON utilize PageRank and BERT embeddings within a divide-and-conquer framework for date selection and summarization Liao et al. (2021), while specialized approaches like TexSL construct spatio-temporal storylines for disaster events using neural embeddings and integer linear programming Yuan et al. (2019). At the document level, optimization techniques continue to be refined, building upon earlier 'Connect the Dots' concepts to maximize coherence and other criteria, sometimes incorporating entity information and temporal decay factors Barranco et al. (2019). Cluster-level approaches represent events as groups of documents, using techniques like Temporal Event Maps (TEMs) based on LDA and mutual information metrics Cai et al. (2019), Event Phase Oriented News Summarization (EPONS) using structural clustering and random walks Wang et al. (2018), or the iterative construction of Story Forests using classifiers and tree-based operations Liu et al. (2020).

3 Methodology

Our approach to narrative extraction integrates two complementary methodologies. These methodologies leverage the strengths of large language models (LLMs) while addressing their limitations in generating concise, evidence-based explanations.

3.1 Supervised Fine-Tuning

For our SFT approach, we fine-tuned multiple LLMs on the training dataset using the Unsloth framework (Daniel Han and team, 2023), which facilitates efficient fine-tuning. The models were trained to generate explanations aligned with the provided dataset and corresponding narrative labels. To ensure clarity and consistency, we structured our prompts to emphasize task requirements, incorporating explicit instructions on output format and content constraints. The prompt template comprised sections for instructions, input documents, narrative information, task specifications, and response areas. This structured format guided the model toward producing focused explanations that directly addressed the assigned narrative within the specified character limits. Additionally, adding language-specific tokens enabled the model to adapt its responses to the target language, a critical feature for multilingual applications.

3.2 Direct Preference Optimization

We employed DPO, a technique that fine-tunes models using preference data. DPO allows the model to learn from human preferences between output pairs, aligning its generations more closely with human judgments of quality and relevance. Unlike traditional reinforcement learning methods, DPO optimizes the model without explicit reward

modeling, instead training it to produce outputs preferred by human evaluators. This approach is particularly valuable for our task, where explanation quality is inherently subjective and difficult to quantify using standard metrics. By leveraging preference data derived from human annotations, we guided the model toward generating explanations that were factually accurate and stylistically aligned with human expectations for narrative justification. Specifically, the model was trained using annotated explanations as preferred responses, while its zero-shot outputs served as rejected explanations, facilitating more effective preference-based learning.

3.3 Inference Strategy

We implemented a multi-candidate generation strategy during inference to address variability in language model outputs. For each article-narrative pair, we generated three candidate explanations using temperature values of 0.5, 0.7, and 0.9, exploring a spectrum from conservative to more diverse outputs.

To select the optimal explanation, our selection algorithm focused on two key components:

1. **Keyword Matching**: We extracted all unique words from the narrative text (after lowercasing) to create a set of narrative-specific keywords. These keywords represented the core concepts that should appear in a relevant explanation. For example, for climate-related narratives, keywords might include `temperature`, `measurement`, `uncertainty`, or `scientific`. Our algorithm measured the overlap between these narrative keywords and the words in each candidate explanation, giving preference to explanations incorporating more narrative-specific terminology.

2. **Length Scoring**: We implemented a normalized length scoring mechanism that awarded maximum points (1.0) to explanations containing approximately 80 words (the task's upper limit), with diminishing returns for shorter explanations, ensuring that explanations were substantial enough to convey necessary information without being overly verbose or truncated.

These scores were combined using a weighted formula: $(\text{keyword\_overlap} \times 2) + \text{length\_score}$, with keyword overlap weighted more heavily to prioritize content relevance over length. This prioritization ensured that even slightly shorter explanations with strong narrative alignment would be selected over longer explanations with weaker relevance.

3.4 Preprocessing Strategies

Our preprocessing pipeline incorporated advanced techniques to improve data quality and model performance. We removed extraneous elements such as emojis, URLs, hashtags, and email addresses to reduce noise and standardized punctuation and whitespace across all five languages. Case normalization minimized variability, and narrative enhancement expanded abbreviated codes into full expressions, enriching semantic context. Additionally, content filtering eliminated the redundant sentences commonly found in news articles, sharpening the model's focus on relevant information. Collectively, these preprocessing techniques produced cleaner, more consistent training examples, enhancing the model's ability to discern narrative patterns and maintain cross-lingual consistency across different approaches and outputs.

4 Experimental Setup

Our experimental framework was designed to evaluate the effectiveness of our approach across multiple languages and model architectures.

4.1 Dataset and Evaluation Metrics

Our system exclusively utilizes the official dataset provided for the task, adhering to the default training-development split. We use the development set solely to assess various experimental configurations during the development phase. The language model is fine-tuned on the training and development sets for the final submission. Performance is evaluated using the F1 macro score, the standard metric for this subtask, despite its tendency to favor majority classes over minority ones.

4.2 Training Strategy

We selected the pre-trained language models employed in our system from those available on Hugging Face. We utilized the PyTorch deep learning framework (version 2.5). For the SFT approach, we employed a batch size of 2 and a learning rate of 2e-4, with five warm-up steps and training for 60 steps. To optimize memory usage on consumer-grade hardware, we applied 4-bit quantization, which reduced memory requirements significantly while

maintaining model performance. For DPO, we further refined our SFT models using preference data derived from the development set, aligning the model outputs more closely with human judgments of explanation quality and relevance.

4.3 Models

We experimented with several pre-trained models from different architectural families and parameter scales. Our primary models included Phi-3.5-mini, Meta-Llama-3.1-8B, and Mistral-7B. All models were fine-tuned using Low-Rank Adaptation (LoRA) with a rank of 16, targeting key parameter matrices (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj) with a LoRA alpha of 16. This parameter-efficient fine-tuning approach enabled us to adapt large models to our specific task while minimizing computational costs and mitigating the risk of catastrophic forgetting.

5 Results

The experimental results demonstrate the effectiveness of our approach across multiple languages and model architectures, consistently outperforming baseline systems.

5.1 Overall Performance

We evaluated multiple language models to identify the most effective architecture for narrative extraction, with Phi-3.5 emerging as the top-performing model. Table 1 provides a comparative analysis of different models across five languages, reporting F1 macro scores alongside baseline performance on the validation set.

Based on our findings, we highlight the following key observations:

- **Model Size vs. Performance**: The smaller Phi-3.5 model consistently outperformed the larger Llama-3.1-8B across most languages. This suggests that model architecture and fine-tuning strategies may play a more critical role than parameter count in optimizing performance for this task.

- **Language-Specific Performance**: Llama-3.1-8B demonstrated superior performance in Hindi, significantly surpassing Phi-3.5 (0.7375 vs. 0.6801). This discrepancy may stem from differences in pre-training corpora, indicating that certain models are inherently better suited for specific languages.

- **Baseline Comparison**: Our top-performing model, Phi-3.5, exceeded baseline performance across all languages, with the most notable gains observed in English (+8.05%) and Russian (+6.98%).

- **Mistral Performance**: The Mistral-7B model consistently underperformed relative to other models and even the baseline, particularly in Portuguese, where it achieved an F1 macro score of only 0.4252.

5.2 Final Submission Results

Our final submission to the SemEval 2025 Narrative Extraction task employed the Phi-3.5 model, incorporating our enhanced preprocessing pipeline and multi-temperature inference strategy. Table 2 presents the official evaluation results across all five languages, ranked by F1 macro score.

Our system consistently outperformed the baseline across all five languages, demonstrating strong performance in Portuguese and English. The lowest improvement was observed in Bulgarian (+4.96%), which nonetheless represented a significant advancement.

The model's robust performance across typologically diverse languages highlights the effectiveness of our approach. We attribute these improvements to several key factors:

- **Effective Preprocessing**: Our preprocessing pipeline which included duplicate sentence removal, text normalization across languages, and targeted evidence filtering ensured cleaner, more relevant inputs for model training and inference.

- **Multi-Temperature Inference**: By generating multiple candidate explanations with varying temperature settings and selecting the most relevant one based on narrative alignment, we achieved significant improvements over single-temperature inference strategies.

- **Model Selection**: Despite having fewer parameters than some alternative architectures, Phi-3.5 exhibited strong instruction-following capabilities and demonstrated superior performance when fine-tuned for narrative extraction.

These results validate our approach, illustrating that carefully designed preprocessing and inference strategies can yield significant performance gains, even when leveraging smaller language models.

Table 1: F1 macro scores for each model across different languages using the SFT approach. Bold values indicate the highest performance per language. The baseline results presented in this table correspond to those provided by the shared task organizers.

Language	Baseline	Microsoft PHI-3.5	Meta-Llama-3.1-8B	Mistral-7B	Improve (%)
Portuguese	0.6804	**0.7487**	0.6627	0.4252	10.04%
English	0.6671	**0.7477**	0.6924	0.5497	12.09%
Russian	0.6442	**0.7141**	0.6447	0.5095	10.84%
Hindi	0.6697	0.6801	**0.7374**	0.6731	10.11%
Bulgaria	0.6343	**0.6853**	0.6613	0.5010	8.04%

Table 2: Evaluation of the WordWiz system across five languages using the F1 macro score.

Language	F1 macro	Rank	Improve(%)
Portuguese	0.7486	1	+10.0%
English	0.7455	2	+11.8%
Russian	0.7040	2	+9.3%
Hindi	0.7336	2	+9.5%
Bulgarian	0.6839	2	+7.8%

Table 3: Comparison of Phi-3.5 SFT and DPO on the English Dataset

Model	Precision	Recall	F1 macro
Phi-3.5 SFT	0.7683	**0.7287**	**0.7477**
Phi-3.5 DPO	**0.7756**	0.6798	0.7243
Baseline	0.6554	0.6796	0.6672

5.3 Direct Preference Optimization

To further enhance our narrative extraction capabilities, we implemented DPO as an additional fine-tuning approach for the English language track. Compared to standard SFT, DPO provides a more sophisticated alignment technique by directly optimizing model outputs based on human preferences without requiring explicit reward modeling.

Our DPO implementation utilized the Phi-3.5 model, previously fine-tuned with SFT, as the reference model. To construct preference pairs, we designated human-annotated explanations from the training set as the chosen responses, while outputs from the non-fine-tuned model served as the rejected explanations. This approach enabled the model to distinguish high-quality narrative justifications from suboptimal ones.

As shown in Table 3, the DPO-tuned model exhibited higher precision (+0.73%) compared to SFT but at the expense of lower recall (-4.89%), resulting in a slightly lower F1 macro score. This experiment suggests that DPO encourages selectivity, prioritizing high-confidence explanations while potentially omitting some valid ones.

5.4 Qualitative Analysis

A qualitative assessment of the DPO-generated outputs reveals a tendency toward more nuanced, contextually aligned justifications. For instance, given article `EN_CC_200040.txt`, the DPO model produced:

> "The article critiques the climate movement, highlighting instances of vandalism and disruption by protesters. It questions the effectiveness of their methods and the public sentiment towards the climate change narrative."

In contrast, the SFT output was:

> "The text criticizes the climate movement for being disruptive and for targeting cultural heritage sites."

The DPO-generated response provides a more comprehensive and contextually enriched explanation, capturing both the criticism and its underlying rationale.

6 Conclusion

This paper presents the WordWiz team's solution for extracting narrative explanations from multilingual news articles. By combining advanced preprocessing techniques with two complementary fine-tuning strategies (SFT and DPO) our system demonstrates substantial improvements over the baseline model across all five competition languages.

The success of our approach underscores the effectiveness of targeted text preprocessing, which facilitates cleaner, more focused input for model training. Our structured prompting strategy effectively guided model generation toward task-relevant outputs. The multi-candidate generation with temperature-based sampling enabled the exploration of diverse response possibilities, while

the selection of candidates based on narrative relevance ensured that final explanations were well-aligned with the intended narrative characterization. Furthermore, our results suggest that, for specialized applications, model architecture and pre-training approach can be more critical than model size.

Future research could explore the extraction of evidence from news text to support narrative extraction. Incorporating more advanced evidence extraction techniques could further enhance the grounding of explanations in the source material. Extending our system's multilingual capabilities to encompass low-resource languages and investigating cross-lingual transfer learning may also expand the system's applicability to a broader range of global contexts.

References

Roberto Camacho Barranco, Arnold P. Boedihardjo, and Mahmud Shahriar Hossain. 2019. Analyzing evolving stories in news articles. *Int. J. Data Sci. Anal.*, 8(3):241–256.

Yi Cai, Haoran Xie, Raymond Y. K. Lau, Qing Li, Tak-Lam Wong, and Fu Lee Wang. 2019. Temporal event searches based on event maps and relationships. *Appl. Soft Comput.*, 85.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Yijun Duan, Adam Jatowt, and Masatoshi Yoshikawa. 2020. Comparative timeline summarization via dynamic affinity-preserving random walk. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1778–1785. IOS Press.

Jieh-Sheng Lee. 2024. InstructPatentGPT: Training patent language models to follow instructions with human feedback. *CoRR*, abs/2406.16897.

Yiming Liao, Shuguang Wang, and Dongwon Lee. 2021. WILSON: A divide and conquer approach for fast and effective news timeline summarization. In *Proceedings of the 24th International Conference on Extending Database Technology, EDBT 2021, Nicosia, Cyprus, March 23 - 26, 2021*, pages 635–645. OpenProceedings.org.

Bang Liu, Fred X. Han, Di Niu, Linglong Kong, Kunfeng Lai, and Yu Xu. 2020. Story forest: Extracting events and telling stories from breaking news. *ACM Trans. Knowl. Discov. Data*, 14(3):31:1–31:28.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Ricardo 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.

Moreno La Quatra, Luca Cagliero, Elena Baralis, Alberto Messina, and Maurizio Montagnuolo. 2021. Summarize dates first: A paradigm shift in timeline summarization. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 418–427. ACM.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, Purificação Silvano, Shivam Sharma, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari Gonçalves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, European Commission Joint Research Centre, Ispra (Italy).

Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2018. Event phase oriented news summarization. *World Wide Web*, 21(4):1069–1092.

Yi Yu, Adam Jatowt, Antoine Doucet, Kazunari Sugiyama, and Masatoshi Yoshikawa. 2021. Multi-timeline summarization (MTLS): improving timeline summarization by generating multiple summaries. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 377–387. Association for Computational Linguistics.

Ruifeng Yuan, Qifeng Zhou, and Wubai Zhou. 2019. dtexsl: A dynamic disaster textual storyline generating framework. *World Wide Web*, 22(5):1913–1933.