DUTtask10 at SemEval-2025 Task 10: ThoughtFlow: Hierarchical Narrative Classification via Stepwise Prompting

Pengyuan Du[†], **Huayang Li**[†], **Liang Yang**[*], **Shaowu Zhang**[*]
Department of Computer Science Dalian University of Technology, LiaoNing, China
[†]{2042997144,3170511502}@mail.dlut.edu.cn
[*]{zhangsw, liang}@dlut.edu.cn

Abstract

This paper describes our system for Subtask 2 of SemEval-2025 Task 10: Hierarchical Narrative Classification. We propose a two-step hierarchical approach that combines generative reasoning and fine-tuning for sub-narrative classification. The main techniques of our system are: 1) leveraging a large pre-trained model to generate a reasoning process for better context understanding, 2) fine-tuning the model for precise sub-narrative categorization, 3) using a multi-label classification strategy for more accurate sub-narrative identification, and 4) incorporating data augmentation to increase the diversity and robustness of the training data. Our system ranked 1st in Subtask 2 for Hindi, achieving an F1 macro coarse score of 0.56900 and an F1 samples score of 0.53500. The results demonstrate the effectiveness of our approach in classifying narratives and sub-narratives in a multilingual setting, with the additional benefit of enhanced model performance through data augmentation.

1 Introduction

The rapid growth of online news content and the ongoing spread of disinformation have made it increasingly important to develop tools that can identify and categorize the underlying narratives shaping public discourse (Xu et al., 2022). To address this, we participated in Subtask 2 of SemEval-2025 Task 10, which focuses on hierarchical narrative classification of multilingual news articles. The goal is to assign both high-level (coarse) and fine-grained (sub-narrative) labels to each document based on a predefined two-level taxonomy.

To achieve this, we developed a novel hybrid approach that strategically combines the reasoning capabilities (Yu et al., 2024) of large language models (LLMs) with the fine-grained classification power of fine-tuned models. First, we leveraged GPT-4o's (Schnabel et al., 2025)API to generate a structured reasoning process (or thought process) for each article. This involved prompting the LLM to interpret the text, and identify potential narrative themes before any classification attempt. Then, rather than directly feeding the GPT-4o outputs into a classifier, we used the generated reasoning as input for further fine-tuning. We fine-tuned the smaller and more efficient GEMMA2-9B (Team et al., 2024) model on the reasoning-enriched data, this allowed GEMMA2-9B to focus on the nuanced task of categorizing sub-narratives more efficiently. Moreover, to address the inherent challenges of multi-lingual analysis and to ensure the robustness of our model across the task's five languages (Bulgarian, English, Hindi, (European) Portuguese, and Russian), we implemented a comprehensive data augmentation strategy (Bayer et al., 2022). This involved techniques designed to increase the diversity of training data, including back-translation and synonym replacement (Madukwe et al., 2022). Various multi-label classification approaches and fine-tuning paradigms were explored to optimize the performance for sub-narrative classification.

2 Background

2.1 Dataset Description

The dataset used in this task spans two major domains: the Ukraine-Russia War and Climate Change. It is designed to evaluate hierarchical narrative classification across five languages: Bulgarian, English, Hindi, Portuguese, and Russian. Both the training and test datasets are provided as plain text files, each article accompanied by hierarchical labels. The training set contains a total of 1699 articles, although a small number were omitted due to inaccessible source URLs. The test set comprises 460 articles, and evaluation is conducted exclusively on sub-narrative (fine-grained) predictions.

[†] Both authors contributed equally to this work.
[*] Corresponding author.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 424–430

A summary of dataset statistics is presented in Table 1.

Language	Training Articles	Test Articles
Bulgarian	401	100
English	399	101
Hindi	366	99
Portuguese	400	100
Russian	133	60
Total	**1699**	**460**

Table 1: Dataset Distribution Across Languages

2.2 Task Description

This work focuses on **Subtask 2: Narrative Classification**, which involves automatically assigning news articles to a two-level taxonomy of narrative labels (Piskorski et al., 2025). The taxonomy is domain-specific, covering the Ukraine-Russia War and Climate Change.

The first level (Level 1) contains broad **main narratives**, while the second level (Level 2) captures more fine-grained **sub-narratives** that elaborate on and support the main ones.

Subtask 2 is framed as a multi-label, multi-class classification task. Each article may be associated with multiple main narratives and multiple sub-narratives, requiring systems to output a set of labels at both levels for each instance. This setup tests a model's ability to recognize layered narrative structures and conceptual relationships across levels of abstraction.

If an article does not match any of the predefined narrative or sub-narrative categories, an "Other" pseudo-label is assigned.

Participants are required to submit two separate lists per article: one for predicted main narratives and another for predicted sub-narratives. Notably, the task does not enforce consistency between the predicted narratives and sub-narratives (i.e., sub-narratives may not need to align hierarchically with predicted main narratives).

3 System Overview

In this section, we detail the methodology developed for SemEval-2025 Task 10, Subtask 2 — a multi-label, multi-class classification task centered on the hierarchical narrative classification of news articles.

Our system adopts a hybrid framework that integrates **generative reasoning** with **fine-tuned classification**. Specifically, we leverage **GPT-4o** to generate structured reasoning chains, which are then used to enrich the input for a smaller, efficient model — **GEMMA2-9B** — responsible for sub-narrative prediction.

The overall architecture of our system is illustrated in Figure 1. To further enhance performance and robustness, we incorporate comprehensive data preprocessing, augmentation, and multilingual adaptation techniques, addressing the diverse challenges posed by the dataset.

3.1 Data Augmentation

To enhance the diversity and robustness of the training data, we applied two data augmentation techniques:

Back-Translation:

- We used the Google Translate API for back-translation. Each article in the training set, across all five languages, was translated into English and then back into its original language.
- Semantic similarity was maintained using cosine similarity scores from pre-trained language-specific embeddings (e.g., fastText for Hindi), with a threshold of 0.85 to ensure content fidelity.

Synonym Replacement (Hindi-Specific):

- Hindi WordNet (Yadav et al., 2024) was used to replace key terms with synonyms, focusing on narrative-relevant words.
- The replacement was constrained to preserve grammatical coherence, and manual verification was performed on a small subset.

We generated two augmented examples per original article (across all languages for back-translation, and for Hindi only for synonym replacement), resulting in approximately 1500 augmented Hindi articles added to the training set.

3.2 Reasoning Generation

In this paper, we introduce the GPT-4o model to generate reasoning chains to improve the accuracy and interpretability of label prediction in text categorization tasks.

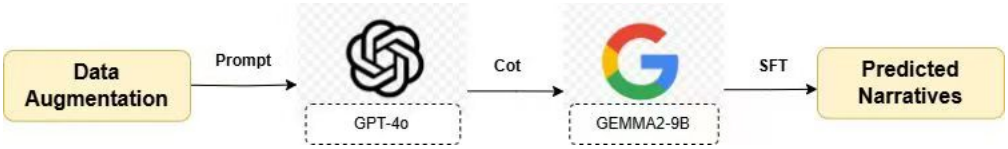

Figure 1: The overall framework of our system proposed for SemEval-2025 Task 10-Subtask2.

During the reasoning chain generation process, based on the prompt we set, GPT-4o derives the logical relationships between the relevant labels and the text. It analyzes the content of the input text to help the model understand how to infer the correct primary label and its associated secondary labels. In brief, the process for generating the text T_R with the reasoning chain explanation is:

$$T_R = \text{GPT-4o}(T_D, y, \text{prompt})$$

where T_D is the result of the original data T after data augmentation, y represents the label of the data, and the prompt is specifically designed to guide GPT-4o's thought process.

The final data T_F, used in supervised fine-tuning, is obtained as:

$$T_F = T_D \oplus T_R$$

3.3 Fine-tuning and Classification

3.3.1 Multi-language embedding

We use the GEMMA2-9B model for fine-tuning in Subtask 2. GEMMA2-9B is a multilingual model based on the Transformer architecture (Vaswani et al., 2017), with strong capabilities for cross-lingual understanding. For the k-th input multilingual text T_F^k, we first apply a dynamic disambiguation process using the SentencePiece tokenizer (Kudo and Richardson, 2018), which segments the text into subword units. These tokens are then mapped into vector representations using a shared embedding table $E \in \mathbb{R}^{|V| \times d}$, where $|V| = 256\text{k}$ is the vocabulary size and $d = 3072$ is the embedding dimension:

$$e_i = E(t_i^k)$$

where t_i^k is the i-th token of the k-th text.

Since GEMMA2-9B has been pre-trained using both Masked Language Modeling (MLM) and Translation Language Modeling (TLM), it is capable of capturing general cross-lingual representations and can effectively process input texts across multiple languages.

3.3.2 Loss Function

The set of tags corresponding to each text is: $\{(L_1, \{S_1^{(1)}, S_1^{(2)}, \dots\}), (L_2, \{S_2^{(1)}, \dots\}), \dots\}$, where L_i denotes the i-th main label, and $\{S_i^{(1)}, S_i^{(2)}, \dots\}$ are the corresponding sub-labels associated with L_i. Given a dataset of N training samples, we define the binary cross-entropy loss for the main label predictions as:

$$L_{main} = -\sum_{i=1}^{m}\sum_{c=1}^{C}\left[L_{BCE}(L_i^c, \hat{L}_i^c)\right]$$

where C is the total number of main label classes, $\hat{L}_i^c$ is the predicted probability for class c in the i-th sample, and $L_i^c \in \{0, 1\}$ is the corresponding ground-truth indicator.

For each sub-label of the text we then have:

$$L_{sub} = -\sum_{i=1}^{m}\sum_{q=1}^{Q}\left[L_{BCE}(S_i^q, \hat{S}_i^q)\right]$$

where Q is the number of possible sub-label classes, and $\hat{S}_i^q$ is the predicted probability for sub-label q of the i-th sample.

The total loss combines the two components with a trade-off parameter $\alpha \in [0, 1]$:

$$L_{\text{total}} = \frac{1}{N}\sum_{i=1}^{N}(\alpha L_{\text{main}}(i) + (1-\alpha)L_{\text{sub}}(i))$$

3.4 Post-processing

To mitigate the issue of over-confidence in certain predictions—which may lead the model to under-predict relevant category labels—we apply *temperature scaling* (Kull et al., 2019) to calibrate the output probabilities. This technique softens the probability distribution, making it smoother and less prone to sharp peaks.

Given the unnormalized logits z_i for each category, the calibrated probability distribution P_i' is computed as:

$$P_i' = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)}$$

where T is the temperature coefficient, z_i is the unnormalized output logits of the model, and P_i' is the probability after temperature smoothing. By increasing the T, the probabilities of the categories can be made more balanced, avoiding over-biasing of the model towards a particular category.

4 Experimental Setup

4.1 Dataset Split

In the experiments of this paper, both the original training set and the training set that has been processed by data augmentation (DA) and incorporated into the thinking process were randomly sampled and divided into 10 non-overlapping subsets. In each cross-validation process, a different subset of the dataset was rotated as the validation set and the remaining subset as the training set. A 10-fold cross-validation was used in all experiments to ensure that the adopted strategy would show good generalization ability on the final test set. The experimental results presented are the average of the 10-fold cross-validation (Browne, 2000), thus maximizing the ability to assess the stability and performance of the model.

4.2 Pre-processing

In the experiments in this paper, we used our own Python script to process the news texts provided by the task organizer, which were initially stored in separate txt files according to their respective languages, and also contained their corresponding English classification labels, and were eventually processed and stored in CSV format. Subsequently, we performed data cleaning to remove some unreasonable data. Finally, for data enhancement, we expanded the dataset with low-resource languages mainly through translation in order to increase the diversity and quantity of data.

4.3 Evaluation Metrics

The evaluation metric for Task 10 is the F1 score, specifically focusing on the **F1 macro coarse** score of the sub-labels. The **F1 macro coarse** score is an effective measure of the model's performance, averaging the F1 scores of each label without considering label frequency. It ranges from 0 to 1, where higher values indicate better classification performance. **In the following table, all F1 scores refer to the F1 macro coarse metric.**

System &F1 score	En	Po	Ru	Bu	Hi
w/ data augmentation					
Baseline	27.2	9.7	17.1	14.9	16.7
+ DA	29.0	9.9	17.0	15.2	52.1
w/o data augmentation					
Baseline	29.0	9.9	17.0	15.2	52.1
+ RG	**30.4**	**10.4**	**17.6**	**15.9**	53.4
+ ML	30.0	10.1	18.1	15.1	**53.7**
+ TS	29.2	10.0	17.3	15.2	52.9
+ All	31.0	10.7	18.5	17.2	56.9

Table 2: Average results with training methods we used. And RG is Reasoning Generation, ML is Multi-label Loss, TS is Temperature Scaling, DA is Data Augmentation

5 Results

5.1 Overall Performance

Finally, according to the official scoring system, our system achieved the first place in the evaluation set for Hindi, and 23, 13, 13, and 11 for English, Russian, Portuguese, and Bulgarian, respectively, and for the sake of presentation, all experimental results are multiplied by 100.

5.2 Data Augmentation

In order to evaluate the effect of data augmentation (DA), we conducted experiments on the original training set and the training set augmented by DA, respectively. We can clearly see the improvement of the model performance by data enhancement in Table 2. The experimental results show that the performance of the system improves significantly after using the augmented dataset. Especially with the effect on Hindi, it can be said that our use of translation expansion as well as synonym substitution using hindwordnet greatly improves the training efficiency for low-resource languages, and therefore, we can conclude that richer training sets definitely help in building more robust models.

5.3 Reasoning Generation

In this paper, we make GPT generate text-to-label thinking reasoning text through the form of prompt. In the Table 2, it can be clearly seen that the F1 score improves after adding the reasoning generation. The results show that besides data augmentation, reasoning generation is the biggest way to improve model performance. Especially in com-

Overall Weight	Hindi F1
0%	51.4
30%	**56.9**
50%	50.9
75%	47.8
100%	37.9

Table 3: Results on training with multi-label loss.

plex multi-label secondary classification tasks, this approach allows the model to analyze the relationship between data and labels at a deeper level.

5.4 Multi-label Loss

As mentioned earlier, we adopted the standard binary cross-entropy loss for both main and sub-labels during training. We experimented with different values of the weighting parameter α to balance their contributions. As shown in Table 3, the best performance was achieved when α was set to 0.3. This indicates that sub-narratives play a more significant role in our system's classification performance.

5.5 Temperature Scaling

For the model trained with data augmentation and reasoning generation, we conducted ablation experiments to assess the effectiveness of the temperature scaling technique. As shown in Table 2, applying temperature scaling led to slight performance improvements, indicating that this method has a positive impact on the classification task by mitigating overconfident predictions.

5.6 Negative Results

In addition to the aforementioned strategies, we also experimented with multi-task learning (Zhang and Yang, 2021) on low-resource languages. For example, in an attempt to reduce the semantic distance between Hindi and English, we incorporated a translation task into a multi-task learning framework. However, contrary to our expectations, this approach resulted in a performance drop. We speculate that this may be due to the significant difference between the output format of the narrative classification task and that of the translation task, which could have caused a conflict in the learning objectives.

6 Conclusion

By leveraging a series of optimization techniques—including data augmentation, reasoning generation, multi-label loss design, and post-processing—we developed a robust framework capable of performing multi-label level-2 classification of news texts in a multilingual and cross-linguistic setting. Our system achieved first place on the low-resource language Hindi, along with competitive results across other languages.

In future work, beyond further enriching the training data, we aim to explore language-specific structural features inherent to different language families (e.g., Hindi), and to incorporate novel methods and architectures to further enhance model performance, particularly for low-resource languages.

References

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Michael W Browne. 2000. Cross-validation methods. *Journal of mathematical psychology*, 44(1):108–132.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Kosisochukwu Judith Madukwe, Xiaoying Gao, and Bing Xue. 2022. Token replacement-based data augmentation methods for hate speech detection. *World Wide Web*, 25(3):1129–1150.

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.

Julian A Schnabel, Johanne R Trippas, Falk Scholer, and Danula Hettiachchi. 2025. Multi-stage large language model pipelines can outperform gpt-4o in relevance assessment. *arXiv preprint arXiv:2501.14296*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zihang Xu, Ziqing Yang, Yiming Cui, and Zhigang Chen. 2022. Hfl at semeval-2022 task 8: A linguistics-inspired regression model with data augmentation for multilingual news similarity. *arXiv preprint arXiv:2204.04844*.

Preeti Yadav, Sandeep Vishwakarma, and Sunil Kumar. 2024. Deep learning-based word sense disambiguation for hindi language using hindi wordnet dataset. In *Federated learning for Internet of Vehicles: IoV Image Processing, Vision and Intelligent Systems*, pages 140–159. Bentham Science Publishers.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39.

Yu Zhang and Qiang Yang. 2021. A survey on multitask learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609.

A Appendix

Table 4 shows the prompt we used to generate the reasoning chain.

Prompt for Reasoning Chain Generation
Instruction: You are given a news article in any language, along with its corresponding main narrative label(s) and sub-narrative label(s). Your task is to write a concise and clear explanation in English that logically connects the article's content with the given labels. Please identify key themes, arguments, or facts in the article that support each main narrative and its related sub-narratives. Please keep your explanation within 100 words.
Input - Article (original language): ARTICLE_TEXT
Input - Narrative Structure: - Main_Label_1: [Sub_Label_1a, Sub_Label_1b, ...] - Main_Label_2: [Sub_Label_2a, Sub_Label_2b, ...]
Output - Reasoning Chain (in English): Explanation_Text

Table 4: Prompt format for generating reasoning chains with GPT-4o