UNEDTeam at SemEval-2025 Task 10: Zero-Shot Narrative Classification

Jesús M. Fraile-Hernández
UNED NLP & IR Group
Universidad Nacional de
Educación a Distancia
28040 Madrid, Spain.
jfraile@lsi.uned.es

Anselmo Peñas
UNED NLP & IR Group
Universidad Nacional de
Educación a Distancia
28040 Madrid, Spain.
anselmo@lsi.uned.es

Abstract

In this paper we present our participation in Subtask 2 of SemEval-2025 Task 10, focusing on the identification and classification of narratives in news of multiple languages, on climate change and the Ukraine-Russia war. To address this task, we employed a Zero-Shot approach using a generative Large Language Model (LLM) without prior training on the dataset. Our classification strategy is based on two steps: first, the system classifies the topic of each news item; subsequently, it identifies the sub-narratives directly at the finer granularity. We present a detailed analysis of the performance of our system compared to the best ranked systems on the leaderboard, highlighting the strengths and limitations of our approach.

1 Introduction

The characterisation and extraction of narratives from news texts is an area of growing interest in natural language processing (NLP), with applications in discourse analysis, bias detection and the understanding of social and political dynamics. In this context, SemEval-2025 Task 10, entitled 'Multilingual Characterization and Extraction of Narratives from Online News' (Jakub Piskorski et al., 2025), seeks to advance the identification and classification of narratives in news stories in multiple languages.

This task is structured in three main subtasks: Subtask 1: Entity Framing, which consists of assigning one or more roles to each named entity mention within a news article, using a predefined taxonomy of roles; Subtask 2: Narrative Classification, where each news article must be assigned all relevant sub-narrative labels within a hierarchical taxonomy of narratives in a specific domain; and Subtask 3: Narrative Extraction, which requires generating a free-text explanation justifying the selection of an article's dominant narrative, based on fragments of text supporting that choice.

Recent advancements in large language models (LLMs) have significantly improved the ability to detect and analyze narratives and propaganda techniques in textual data. Leveraging their contextual understanding and capacity to generalize from large-scale datasets, LLMs have been applied to identify persuasive strategies, ideological framing, and coordinated messaging in political discourse and media (Jones, 2024; Liu et al., 2025).

Our work focuses exclusively on Subtask 2. The task presents multiple challenges, including linguistic diversity, subjectivity in categorising narratives and the limited availability of labelled data in multiple languages. Addressing these problems requires robust approaches that can generalise well across languages and domains. To this end, we opt for a Zero-Shot approach, in which we leverage the knowledge of an LLM without performing specific training on the task data. Since some LLMs have been trained with varying amounts of data in the five languages of the task and SOTA models often performs better in English, we performed a machine translation into English using OPUS-MT models prior to inference.

This paper describes in detail our methodology, the results obtained in comparison with the best competing models and a critical analysis of the performance. Finally, we discuss the limitations of our approach and propose future directions for improving automatic narrative classification in multilingual contexts.

2 Subtask description

SemEval-2025 Task 10 Subtask 2 focuses on the identification and classification of narratives in news articles covering multiple languages and subject domains.

The task is multilingual in five different languages: English, Bulgarian, Portuguese, Hindi

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 165–173
July 31 - August 1, 2025

and Russian. Having a large number of languages avoids linguistic and cultural biases in the classification models, allowing the systems to be more robust and adaptable to different contexts. Moreover, the presence of languages rarely used in LLM training, such as Bulgarian or Hindi, introduces additional challenges in identifying and structuring narratives and enriches the evaluation of the performance of multilingual and multicultural models.

Also, the subtask focuses on two thematic domains of great current relevance such as the war between Ukraine and Russia and climate change. Both topics generate a large amount of content on social networks and in the media, which allows us to analyse the propagation of narratives in contexts of high political, economic and social impact. Moreover, by covering both a geopolitical conflict and a global environmental crisis, different types of narratives are covered: some focused on politics and war, and others on science, economics and sustainability. Similarly, including news in Russian is particularly relevant for the analysis of the Ukraine-Russia conflict. Russian media often offer a different perspective than Western media, which makes it possible to study how narratives are constructed and disseminated within Russia and internationally.

2.1 Dataset description

The classification structure used in this task follows a three-level hierarchy (Stefanovitch et al., 2025), in which the top level is defined by the overall news topic, which represents the general thematic domain to which the content belongs, such as ‘Russia-Ukraine War’ or ‘Climate Change’. At the second level are the main narratives, which include general interpretative frameworks within each topic, providing an overall perspective on the issue. Finally, the third level is composed of sub-narratives, which detail in greater granularity specific aspects within each main narrative. In addition, the category *Other* is included at the topic level, for those news items that do not fit clearly into any of the topics, and at the subnarrative level, for those news items that do not fit with the defined subnarratives or could support other subnarratives within the main narrative.

The dataset used has a total of 10 main narratives and 46 sub-narratives (36 specific and 10 labelled Other) related to Climate Change, as well as 11 main narratives and 49 sub-narratives (38 specific and 11 labelled Other) on the Ukraine-Russia war. In total, the training + development set contains 576 news items, distributed across languages and categories.

	Train	**Dev**	**Test**	**Total**
EN	399	41	101	541
BG	401	35	100	536
PT	400	35	100	535
RU	348	32	60	440
HI	366	35	99	500
Total	1914	178	460	

Table 1: Number of news items by dataset.

Regarding the number of news items available in the datasets, Table 1, lists the number of instances per language in each available dataset.

To illustrate the distribution of narratives in the training + development corpus, a bar chart showing the frequency of the 20 most common sub-narratives is presented in Figure 1, which allows us to observe the variability in the representation of each category within the dataset.

2.2 Evaluation

The official evaluation measure for this sub-task is the F1 of samples averaged over documents. This metric assesses the precision and recall of the labels of the narratives and sub-narratives assigned to each news item. In addition, the standard deviations of both F1 values are indicated.

3 Methodology

In this research, we have decided to use a Zero-Shot approach to news classification, without using pre-trained examples. This decision responds to the difficulty of finding sufficiently large datasets labelled by human annotators to train a model when we move to real world scenarios. By employing a Zero-Shot approach, we take advantage of the power of LLMs, which have been trained on a large amount of data and are able to generalise to new tasks without the need for prior examples.

To overcome the linguistic limitations that LLMs may have, we have translated all news items into English using machine translation models based on Opus-MT (Tiedemann), an initiative of the University of Helsinki that provides multilingual machine translation models, facilitating translation across languages. This decision is justified by the fact that not all available LLMs have been trained with a large amount of data in the five languages of interest. By translating the articles into English, we

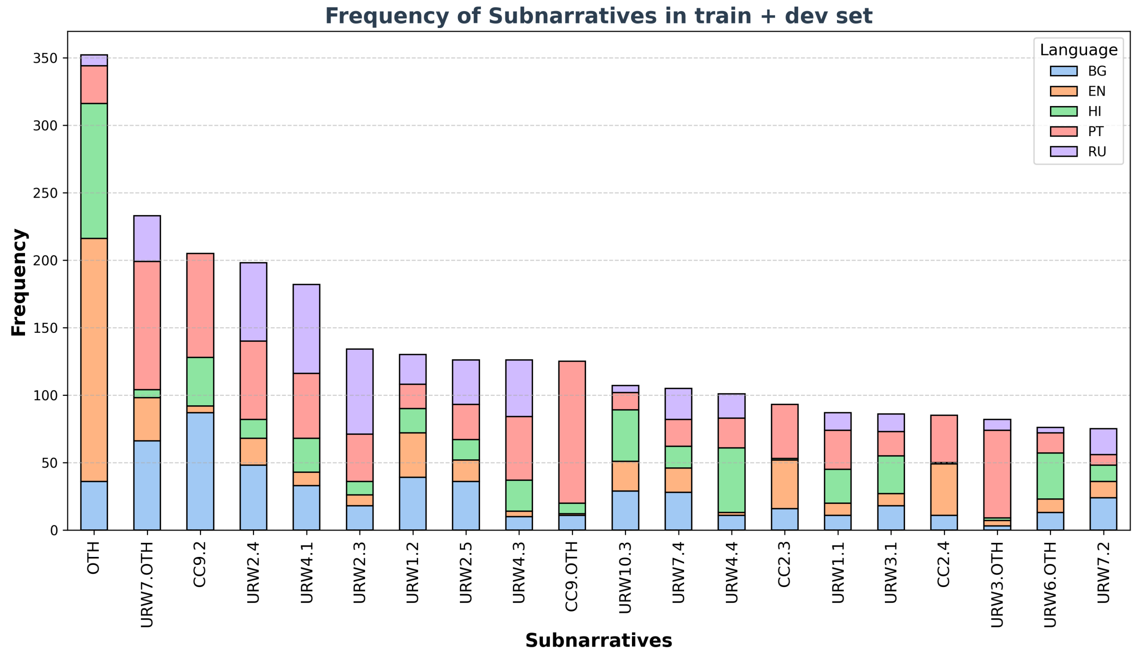

Figure 1: Frecuency of subnarratives in train+dev set grouped by country.

seek to maximise the accuracy of classification and narrative detection, taking advantage of the knowledge of the models in this language. In addition, this methodology allows for a more homogeneous comparison of the news items, regardless of the original language but adding possible errors in the translator models.

3.1 Hierarchical classification

The classification process has been divided into two stages, both carried out by formulating prompts, as described in Appendix A.

1. The main theme of the news item has been classified, assigning it to one of two possible topics: URW (Russia-Ukraine War) or CC (Climate Change).

2. Within each assigned topic, we have classified the corresponding subnarratives. The label *Other* has been incorporated at both topic and subnarrative level.

Subsequently, labels for the narratives were extracted from the model responses.

3.2 LLM configuration

For each of the ranking tasks, we selected the calme-2.4-rys-78b (Panahi, 2023) model, a fine-tuned model based on Qwen 2 78B (Qwen et al., 2025), which has demonstrated high performance with a score of 0. 669 in MMLU-Pro Leaderboard (Wang et al., 2024), placing it as the fourth best model in the Open LLM Leaderboard (Ope).

This model was quantized to 4 bits to optimise memory usage and speed up inference. Calculations were performed using the *bfloat16* format, which ensures greater efficiency without compromising accuracy. For inference, temperature was set to 0.75, top_k to 5, and max_new_tokens was set to 200.

3.2.1 Prompts

To carry out the hierarchical classification of the news, two specific prompts have been built, one for each level as described in 3.1. The first prompt is destined to the classification of the main topic of each news item, assigning it to one of the two possible themes: URW (Russia-Ukraine War) or CC (Climate Change). The second prompt is used for the classification of the sub-narratives within each of the assigned topics, allowing the model to identify the specific sub-narratives that the news item supports. These prompts only include the title of the topic or subnarratives to be classified.

In addition, the model's response has been requested to be delivered in JSON format, including a detailed reasoning for the classification made. This structure allows the model to provide clear explanations of its decisions, which helps to improve the transparency and interpretability of the classification process and facilitates the evaluation of the consistency and validity of the assigned labels.

The prompts used can be found in Appendix A.

4 Results

To evaluate the performance of our Zero-Shot approach, we performed an initial evaluation on the training and development sets, which allowed us to analyse the model's ability to correctly classify both the topic and the subnarratives.

Table 2 presents the results for each language and the overall macro result of the system on the development set. This includes the total number of news items in each set, the F1 metric for the topic, the F1 for the Other category (indicating the model's ability to identify news items that do not fit into any predefined narrative or topic), as well as the metrics used in the overall evaluation: F1 coarse (narrative classification), standard deviation coarse, F1 samples (subnarrative classification) and standard deviation samples.

We then applied the model to the test set, generating the final inferences. Table 3 includes the same metrics by language with the results of our system on the test set together with the best model recorded on the task. In addition, the difference between our approach and the leading model in each language is shown, allowing us to quantify the difference in performance between the two.

5 Discussion

The obtained results show a variable performance of the model depending on the language, with significant differences between the dev set and the test set. In the evaluation on dev, it is observed that the model achieves a high F1 Topic in all languages, with values above 87%, indicating that topic classification (URW or CC) is relatively straightforward for the model. However, F1 Other is much lower, especially in Portuguese (PT), where the model did not correctly identify any news items in the Other category, suggesting that the model's ability to detect news items that do not fit the predefined narratives varies by language. As for the classification of sub-narratives, F1 Samples values show moderate performance, with Bulgarian (BG) as the best performing language (0.4248), while Hindi (HI) and Portuguese (PT) show the lowest values (0.2305 and 0.2257, respectively).

Analysing the results on the test set, a generalised decrease in ranking metrics is observed with respect to the dev set, indicating that the distribution of labels in the test set might be slightly different from the dev set. Comparison with the best model reveals noticeable performance differences. For example, in Russian (RU), our model obtains an F1 Coarse of 0.513, while the best model achieves 0.709, which represents a difference of 0.196 points. Similarly, in Portuguese (PT), the difference is 0.127 points.

One aspect to note is that the difference in F1 Samples (classification of subnarratives) is larger than in F1 Coarse, indicating that the identification of subnarratives remains a greater challenge than the classification of narratives. Furthermore, the standard deviation of our model and the best ranked model across all languages remains relatively high, suggesting considerable variability in the quality of predictions. This behaviour is especially visible in English (EN) and Hindi (HI), where the standard deviation values are the highest, suggesting that the model is less consistent in these languages.

To better interpret these results, it would have been useful to have a measure of the Inter-Annotator Agreement when constructing the dataset. Knowing the Inter-annotator agreement would allow contextualising the F1 values and standard deviations, providing a reference on the intrinsic difficulty of the task. If inter-annotator agreement were low, this would indicate that even for humans the classification of certain news items into specific narratives is ambiguous, which would help to establish a reasonable threshold for evaluating the model's performance. On the other hand, if the agreement were high, the observed variability in the model's predictions could be attributed mainly to limitations in its generalisability. This analysis would be particularly relevant in languages with higher variability in F1 and high standard deviations, as it would allow distinguishing between problems arising from annotation ambiguity and model-inherent errors.

6 Conclusion and Future Work

In this research, we have presented a LLM-based system using a Zero-Shot approach for the SemEval 2025 Task 10 Subtask 2, together with a machine translation into English, focused on the classification of narratives in news stories from five different languages. Our system approached this task without using training data, using only the ability of the pre-trained model to identify the gen-

	n_news	F1 Topic	F1 Other	F1 Coarse	Std Coarse	F1 Samples	Std Samples
EN	41	0.8645	0.5333	0.5125	0.3451	0.3836	0.3412
BG	35	0.9106	0.5455	0.6078	0.3536	0.4248	0.3794
PT	35	0.8165	0.0000	0.4771	0.3993	0.2257	0.3106
RU	32	0.9455	0.6667	0.6691	0.3348	0.4312	0.3288
HI	35	0.8116	0.2500	0.3329	0.3761	0.2305	0.3446
Global	178	0.8697	0.3991	0.5199	0.3392	0.3067	0.3409

Table 2: System results on the dev set

Lang	Rank	F1 Coarse	Std Coarse	F1 Samples	Std Samples
EN (our)	11	0.512	0.364	0.313	0.294
EN (best)		0.590	0.353	0.438	0.333
EN delta		0.078	-0.011	0.125	0.039
BG (our)	4	0.574	0.353	0.363	0.312
BG (best)		0.631	0.338	0.460	0.333
BG delta		0.057	-0.015	0.097	0.021
PT (our)	7	0.537	0.324	0.270	0.262
PT (best)		0.664	0.260	0.480	0.254
PT delta		0.127	-0.064	0.210	-0.008
RU (our)	7	0.513	0.325	0.330	0.270
RU (best)		0.709	0.274	0.518	0.282
RU delta		0.196	-0.051	0.188	0.012
HI (our)	4	0.449	0.460	0.376	0.456
HI (best)		0.569	0.484	0.535	0.494
HI delta		0.120	0.024	0.159	0.038

Table 3: System results on the test set

eral theme of the news and the sub-narratives they support.

The results obtained show that subnarrative classification remains a challenge with low performance and high variability in predictions. Comparison with the best model of the SemEval Task 10 Subtask 2 shared task shows that our system performs worse with noticeable differences, especially in languages such as Russian and Portuguese.

These results suggest some directions for future work. First, an analysis of the importance of machine translation would allow us to quantify the degree of error introduced and its effect on classification. It would be interesting to explore direct classification without translation in those languages with sufficient coverage in the base models. In addition, to obtain a more reliable estimate of model performance, it would be necessary to perform multiple runs on the test set and employ ensemble techniques, combining predictions from several model runs to reduce variability and improve the robustness of the system. Finally, the data provided in the training and development set could be used to test supervised approaches such as Few-Shot or Fine-Tuning.

Limitations

Our approach has several limitations that must be considered. First, due to hardware constraints, it was necessary to quantise the model to 4 bits. This may have affected the accuracy of the model by losing information in the weights. In addition, the inference time was considerably high, which hinders the scalability of the system in large data volume applications.

Another important limitation is the use of machine translation, which is likely to introduce noise in the textual representations and may affect the classification of narratives. Also, the unsupervised Zero-Shot approach prevents tuning the model with task-specific examples, which limits its ability to learn finer patterns in classification. Additionally, the large number of sub-narratives and the length of news stories pose problems in managing context within LLMs.

Finally, an additional limitation is the lack of a more robust performance evaluation, as only a single model run was performed. To obtain more reliable results, it would be necessary to perform multiple runs and apply ensemble techniques that reduce the variability of the predictions.

Acknowledgments

This work was supported by the HAMiSoN project grant CHIST-ERA-21-OSNEM-002, AEI PCI2022-135026-2 (MCIN/AEI/10.13039/501100011033 and EU "NextGenerationEU"/PRTR).

References

Open LLM Leaderboard - a Hugging Face Space by open-llm-leaderboard.

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.

Daniel Gordon Jones. 2024. Detecting Propaganda in News Articles Using Large Language Models. *Engineering: Open Access*, 2(1):1–12. Publisher: Opast Publishing Group.

Jiateng Liu, Lin Ai, Zizhou Liu, Payam Karisani, Zheng Hui, May Fung, Preslav Nakov, Julia Hirschberg, and Heng Ji. 2025. PropaInsight: Toward Deeper Understanding of Propaganda in Terms of Techniques, Appeals, and Intent. *arXiv preprint*. ArXiv:2409.18997 [cs].

Maziyar Panahi. 2023. MaziyarPanahi/calme-2.4-rys-78b · Hugging Face.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. *arXiv preprint*. ArXiv:2412.15115 [cs].

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).

Jorg Tiedemann. Parallel Data, Tools and Interfaces in OPUS.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *arXiv preprint*. ArXiv:2406.01574 [cs].

A Used Prompts

Your main function is to analyse a news item and classify it according to the thematic of the text.
Themes to detect:

1: The war between Ukraine and Russia.
2: Climate change.

The text of the news item you have to analyse is:

(Start of news item to be analysed)

`news text`

(End of news item to be analysed)

Instructions for Classification:

1- Read carefully the news.
2- Determine what the main topic of the news item is.
3- You have to generate a .json structure:

```
 {"classification": [1, 2] only one of the two categories (Write it between
[]),
"reasoning": 'reasoning of the answer in maximum 50 words.' }
```

Figure 2: Topic classification prompt.

Your primary role is to analyze a new and categorize them according to predefined narrative and sub-narrative themes that reflect different portrayals and perspectives of the Ukraine-Russia war (URW). Your classification should help in understanding the overarching sentiments and strategic messaging in public discourse.
Narratives and Sub-narratives to Detect:
URW1.: Blaming the war on others rather than the invader.
- URW1.1: Ukraine is the aggressor.
- URW1.2: The West are the aggressors.
URW2.: Discrediting Ukraine.
- URW2.1: Rewriting Ukraine's history.
- URW2.2: Discrediting Ukrainian nation and society.
⋮
- URW10.3: There is a real possibility that nuclear weapons will be employed.
- URW10.4: NATO should/will directly intervene.
URW11.: Hidden plots by secret schemes of powerful groups.

The text of the news item you have to analyse is:

(Start of news item to be analysed)

`news text`

(End of news item to be analysed)

Instructions for Classification:

1- Read carefully the news.
2- Determine which sub-narrative(s) it supports based on the content and sentiment expressed, a news item can align with several sub-narratives if it incorporates elements from more than one category.
If the text supports a narrative, e.g. URW1., but does not support any of the sub-narratives proposed for that narrative you have to write the code of the narrative followed by OTH, e.g. URW1.OTH
If the text does not support any narrative write OTH.OTH

Valid labels are: ['URW1.1', 'URW1.2', 'URW1.OTH', 'URW2.1', 'URW2.2', 'URW2.3', 'URW2.4', 'URW2.5', 'URW2.6', 'URW2.7', 'URW2.8', 'URW2.OTH', 'URW3.1', 'URW3.2', 'URW3.3', 'URW3.OTH', 'URW4.1', 'URW4.2', 'URW4.3', 'URW4.4', 'URW4.5', 'URW4.OTH', 'URW5.1', 'URW5.2', 'URW5.3', 'URW5.OTH', 'URW6.1', 'URW6.2', 'URW6.3', 'URW6.OTH', 'URW7.1', 'URW7.2', 'URW7.3', 'URW7.4', 'URW7.5', 'URW7.6', 'URW7.OTH', 'URW8.1', 'URW8.2', 'URW8.OTH', 'URW9.1', 'URW9.2', 'URW9.OTH', 'URW10.1', 'URW10.2', 'URW10.3', 'URW10.4', 'URW10.OTH', 'URW11.OTH', 'OTH.OTH']
3- You have to generate a .json structure:

```
{"classification": ['URW1.1', 'URW2.1', 'URW7.OTH', ...] The valid labels of
the sub-narratives supported by the text according to instruction 2.,
"reasoning": 'reasoning of the answer in maximum 50 words.'}
```

Figure 3: URW subnarratives classification prompt.

Your primary role is to analyze a new and categorize them according to predefined narrative and sub-narrative themes that reflect different portrayals and perspectives of the Climate Change (CC). Your classification should help in understanding the overarching sentiments and strategic messaging in public discourse.
Narratives and Sub-narratives to Detect:
CC1.: Criticism of climate policies.
-CC1.1: Climate policies are ineffective.
-CC1.2: Climate policies have negative impact on the economy.
-CC1.3: Climate policies are only for profit.
CC2.: Criticism of institutions and authorities.
-CC2.1: Criticism of the EU.
-CC2.2: Criticism of international entities.
⋮
CC10.: Green policies are geopolitical instruments.
-CC10.1: Climate-related international relations are abusive/exploitative.
-CC10.2: Green activities are a form of neo-colonialism.
The text of the news item you have to analyse is:

(Start of news item to be analysed)

`news text`

(End of news item to be analysed)

Instructions for Classification:

1- Read carefully the news.
2- Determine which sub-narrative(s) it supports based on the content and sentiment expressed, a news item can align with several sub-narratives if it incorporates elements from more than one category.
If the text supports a narrative, e.g. CC1., but does not support any of the sub-narratives proposed for that narrative you have to write the code of the narrative followed by OTH, e.g. CC1.OTH
If the text does not support any narrative write OTH.OTH

Valid labels are: ['CC1.1', 'CC1.2', 'CC1.3', 'CC1.OTH', 'CC2.1', 'CC2.2', 'CC2.3', 'CC2.4', 'CC2.OTH', 'CC3.1', 'CC3.2', 'CC3.OTH', 'CC4.1', 'CC4.2', 'CC4.3', 'CC4.4', 'CC4.5', 'CC4.6', 'CC4.7', 'CC4.8', 'CC4.OTH', 'CC5.1', 'CC5.2', 'CC5.3', 'CC5.4', 'CC5.OTH', 'CC6.1', 'CC6.2', 'CC6.3', 'CC6.OTH', 'CC7.1', 'CC7.2', 'CC7.3', 'CC7.4', 'CC7.OTH', 'CC8.1', 'CC8.2', 'CC8.OTH', 'CC9.1', 'CC9.2', 'CC9.3', 'CC9.4', 'CC9.OTH', 'CC10.1', 'CC10.2', 'CC10.OTH', 'OTH.OTH'] 3- You have to generate a .json structure:

```
 {"classification": ['CC1.1', 'CC2.1', 'CC4.OTH', ...] The valid labels of the
sub-narratives supported by the text according to instruction 2.,
"reasoning": 'reasoning of the answer in maximum 50 words.'}
```

Figure 4: CC subnarratives classification prompt.