Irapuarani at SemEval-2025 Task 10: Evaluating Strategies Combining Small and Large Language Models for Multilingual Narrative Detection

Gabriel Assis, Lívia de Azevedo
João Vitor de Moraes, Laura Alvarenga and **Aline Paes**
Institute of Computing, Universidade Federal Fluminense, Niterói, RJ, Brazil
{*assisgabriel, liviaazevedosilva, joaovitormoraes, l_alvarenga*}*@id.uff.br*, *alinepaes@ic.uff.br*

Abstract

This paper presents the Irapuarani team's participation in SemEval-2025 Task 10, Subtask 2, which focuses on hierarchical multi-label classification of narratives from online news articles. We explored three distinct strategies: (1) a direct classification approach using a multilingual Small Language Model (SLM), disregarding the hierarchical structure; (2) a translation-based strategy where texts from multiple languages were translated into a single language using a Large Language Model (LLM), followed by classification with a monolingual SLM; and (3) a hybrid strategy leveraging an SLM to filter domains and an LLM to assign labels while accounting for the hierarchy. We conducted experiments on datasets in all available languages, namely Bulgarian, English, Hindi, Portuguese and Russian. Our results show that Strategy 2 is the most generalizable across languages, achieving test set rankings of 22st in English, 8th in Bulgarian, 9th in Portuguese, 10th in Russian, and 11th in Hindi.

1 Introduction

Trusting online content has become increasingly difficult due to the rise of misinformation, disinformation, deceptive content, and deliberate attempts at manipulation (Marwick and Lewis, 2017; Anderson, 2019). Not only is it more challenging to distinguish between credible information and fake news, but the sophisticated techniques used to shape perceptions can intensify conflicts and influence political opinions, potentially swaying voter behavior (Stanley, 2015; Rutheford, 2023). The vast amount of online disinformation highlights the urgent need for automated tools to identify such content (Piskorski et al., 2022).

Our research is centered on SemEval-2025 Task 10 (Piskorski et al., 2025), which addresses the Multilingual Characterization and Extraction of Narratives from Online News. Specifically, we focus on Subtask 2, which involves classifying narratives and sub-narratives within a two-level taxonomy. The primary objective of the task is to foster the development of classification methodologies capable of identifying narratives designed to manipulate readers, instantiated this year in the domains of Climate Change and the Ukraine-Russia War. Additionally, the task provides resources and enables participants to work in at least one of five languages: Bulgarian (BG), English (EN), Hindi (HI), Portuguese (PT), and Russian (RU). Our team, however, has chosen to evaluate our approaches across all available languages, aiming to achieve a multilingual analysis.

We evaluated three distinct methodologies leveraging both Small Language Models (SLMs) and Large Language Models (LLMs) to address this task. Moreover, we intend to assess whether the strategies exhibit generalizability across different languages, pursuing a general framework rather than a language-specific strategy. The approaches are as follows: (1) a direct classification method employing a multilingual SLM; (2) a translation-based approach, where texts in multiple languages were translated into a single target language using an LLM, followed by classification with a monolingual SLM; and (3) a hybrid strategy that integrated the strengths of both model types, utilizing an SLM for domain filtering and an LLM for hierarchical label assignment. Our experiments on the development set show that Strategy 2 is the most generalizable across languages among the approaches we evaluated, ranking 8th in Bulgarian, 9th in Portuguese, 10th in Russian, 11th in Hindi and 22st in English on the test set[1].

2 Related Work

Research on the detection and classification of mis-/disinformation, narratives and propaganda has in-

[1] https://github.com/MeLLL-UFF/irapuarani

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 38–48

creasingly leveraged the advanced capabilities of language models. Encoder-based SLMs have been successfully employed in narrative classification tasks. For instance, Coan et al. (2021) explored combining the RoBERTa model (Liu et al., 2019) with the traditional machine learning algorithm logistic regression, utilizing a taxonomy within the climate change domain. Similarly, Kotseva et al. (2023), working in the context of COVID-19, reported success with a fine-tuned BERT (Devlin et al., 2019) model for classifying narratives.

In addition to encoder-based SLMs, LLMs have also been applied in analyzing narratives and propaganda (Liu et al., 2025). Hasanain et al. (2024) experimented with GPT-4 for annotating spans of propaganda in Arabic news articles, highlighting the model's potential when provided with additional contextual information. Jones (2024) evaluated GPT-3.5-turbo's performance in identifying up to 18 possible persuasion techniques in news articles, reporting promising results while noting that the model's ability to detect these techniques varied across some categories. Furthermore, Sprenkamp et al. (2023) compared the performance of RoBERTa with GPT-3 and GPT-4 for propaganda detection, emphasizing that GPT-4 ranked among the best-performing models, alongside RoBERTa.

Our work aims to evaluate strategies that integrate both SLMs and LLMs for classifying narratives in news articles within a multilingual context. Moreover, we aim to determine whether the strategies exhibit generalizability across different languages, pursuing a general approach over a narrowly specialized one. More details are in the sections below.

3 Background

The data utilized in this work was provided by the SemEval-2025 Task 10, which comprises a multilingual corpus of news articles. The corpus spans articles collected between 2022 and mid-2024, focusing on two primary topics: the Ukraine-Russia War and Climate Change. In the context of the addressed subtask, namely Subtask 2, the data also includes labels associated with Narrative Classification, structured into a two-level hierarchy dataset based on the provided annotations (Stefanovitch et al., 2025) . The Ukraine-Russia War (URW) domain includes 11 narrative labels and 38 sub-narratives. In contrast, the Climate Change (CC) domain has 10 narrative labels and 36 sub-narratives. The label *Other* can be used at the narrative level to indicate that a narrative does not match any available labels. It can also be paired with a narrative label to indicate that the corresponding sub-narrative does not fit the predefined categories. Finally, the task organizers pre-divided the set as outlined in Table 1.

Table 1: Distribution of the dataset across languages and its partitioning into train, dev and test sets.

Set	BG	EN	HI	PT	RU	Total
Train	401	400	366	400	348	1915
Dev	35	41	35	35	32	178
Test	100	101	99	100	60	460

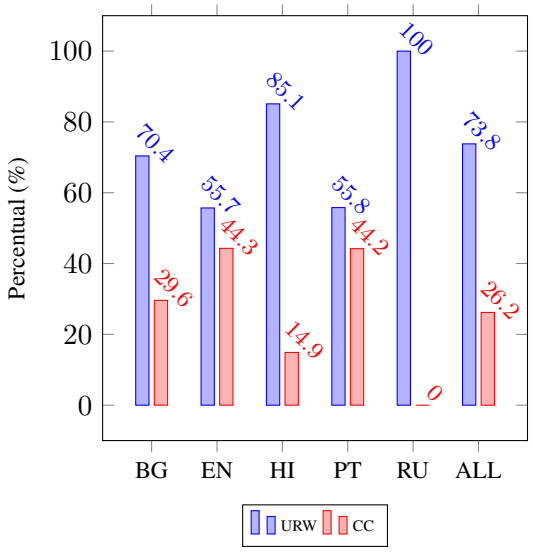

Figure 1: Distribution of the train set across domains.

The dataset exhibits an imbalance across domains, as illustrated in the Figure 1, which highlights the predominance of the URW class over the CC class both overall and across languages. EN and PT are the closest to achieving balance among the analyzed languages. However, even within these relatively balanced languages, a significant observation arises when examining the label taxonomy more closely: not all labels are represented across all languages. For instance, the label associated with the pair {narrative: *Amplifying Climate Fears*, subnarrative: *Whatever we do, it is already too late*} is entirely absent from the training instances in English. Despite this, our data analysis confirms that each label is present in the training set for at least one language. Consequently,

any approach aiming to comprehensively cover all labels across all languages, must address the label underrepresentation. Our approach is detailed in the next section.

4 System Overview

This section outlines the methodology employed for the multilabel classification of narratives within the task's two-level hierarchy. Based on the data characteristics previously described, we evaluated three strategies to ensure comprehensive label and language coverage. The **(a) Single-Model Strategy**, following Vasconcelos et al. (2024), uses a multilingual SLM to classify narratives without considering their hierarchical structure. The **(b) Translation Strategy** involves translating texts into a single target language with an LLM, followed by classification using a monolingual SLM. Lastly, the **(c) Hierarchical Strategy** applies a hybrid approach: an SLM first classifies texts into URW or CC domains, guiding an LLM to assign the final label based on the hierarchy.

4.1 Single-Model Strategy

This approach aims to evaluate the performance of a simplified solution to the problem, deliberately disregarding hierarchical structures (Vasconcelos et al., 2024). To achieve this, a label engineering process is applied, combining each narrative with its respective sub-narratives to create a single, flattened level of possible labels. Formally, let N represent a narrative and $S = \{S_1, S_2, \ldots, S_k\}$ represent its associated sub-narratives. For each pair (N, S_i), a new label is generated in the form N-S_i, where $i \in \{1, 2, \ldots, k\}$. Additionally, for each narrative N, a corresponding "Other" label N-$Other$ is created to represent cases where no specific sub-narrative is identified. Finally, a global label Other-Other is included to handle instances where neither the narrative nor its sub-narratives are recognized.

For the classification process, we employed the multilingual version of the DeBERTa (He et al., 2021a,b) model[2], with linear layers appended to the top of the model's language representation stack. For this approach, we leverage supervised fine-tuning, allowing the weights of both the language model and the newly added linear layers to be jointly optimized. The selection of this model was motivated by its effectiveness as a robust alternative for classification tasks, owing to its advanced ability to encode and represent contextual information (He et al., 2021a).

4.2 Translation Strategy

Given that the previously presented approach assigned the classifier the dual responsibility of handling both multilingual representation and multilabel classification within the hierarchy, the present approach seeks to decouple these two tasks. The goal is to determine whether such a separation leads to any observable improvement in performance.

To achieve this, non-English texts were first translated into English. The decision to translate into English was based on the extensive availability of state-of-the-art models and resources for this language (Joshi et al., 2020; Üstün et al., 2024), also enabling an evaluation of whether a monolingual model could outperform the multilingual approach used in the Strategy 4.1. For a more aligned comparison, a monolingual DeBERTa (He et al., 2021a,b) model[3] was selected, configured similarly to the previous strategy, also with labels presented in a flat structure.

We evaluated two models for the translation stage, namely the Aya Expanse 8B model (Dang et al., 2024) and GPT-4o-mini (Hurst et al., 2024), both with recognized multilingual capabilities. This selection aims to assess the potential impact of translation differences throughout the process by comparing a leading open-source, smaller-scale model with a top-tier proprietary model. Such a comparison enables informed implementation decisions based on the available resources. Lastly, the prompt used can be found in the Appendix A.

4.3 Hierarchical Strategy

In this strategy, we evaluate the performance of a larger, general-purpose LLM by directly assigning multilabel narrative labels. However, similar to Strategy 4.2, we also divide the task into two distinct stages, forming a two-level classification hierarchy (Zangari et al., 2024). In the first stage, we utilize an SLM — specifically, the same multilingual DeBERTa model employed in Strategy 4.1 —— as the basis for a classifier responsible for determining the domain of each article. This classifier predicts a label from the set {URW, CC, Other}, a ternary classification scheme derived through a label engineering process applied to the training

[2]https://huggingface.co/microsoft/mdeberta-v3-base

[3]https://huggingface.co/microsoft/deberta-v3-base

dataset. Importantly, when the classifier assigns the label "Other" to a given text, our framework automatically designates the corresponding sub-narrative label as "Other".

Next, for texts classified as either URW or CC, an LLM is employed with an appropriately designed prompt, guiding the model to provide both the narrative and the sub-narrative. In this context, the selected model was the state-of-the-art GPT-4o (Hurst et al., 2024), specifically its mini version, to address cost-related constraints. This prior domain classification allows for a prompt that is not excessively long, focusing on each specific domain and enhancing the model's performance, as LLMs often struggle with overly extended instructions (Levy et al., 2024). Conversely, we are aware that a hierarchical approach that deepens the hierarchy levels — for instance, with separate stages for classifying the narrative and the sub-narrative — may yield more accurate results. However, we focus on a broader level, as specializing in each narrative could result in a plethora of over-specialized models, which, in real-world scenarios, may reduce generalizability and require retraining with each new label added. The algorithm in Appendix B gives an overview of the hierarchical implementation.

The prompt used for classification with the LLM was refined through empirical testing, incorporating two key instructions: *"In the following text, identify the core narrative that aligns with the author's perspective"* and *"If multiple narratives are equally significant, include them all."* The first instruction was placed at the beginning of the prompt, as experimental results indicated that, in its absence, the model frequently assigned indirect labels to texts, failing to distinguish between internal quotations and the overarching narrative. For example, a text might cite statements from an activist with the intent to discredit them, in which case the appropriate label would be *"Ad hominem attacks on key activists."* However, error inspections revealed that the model occasionally misclassified such texts, interpreting the activist's statements as representative of the main narrative, thereby diverging from the intended annotations. Appendix C presents an example of the incorrect classifications observed. Furthermore, the second directive was appended at the end of the prompt to mitigate the overly restrictive effect of the first instruction. Preliminary experiments demonstrated that relying solely on the first instruction led the model to apply excessively narrow labels. The complete prompts for the URW and CC domains are provided in Appendices D and E, respectively.

5 Experimental Setup

Implementation Details The proposed solution was developed utilizing the Hugging Face Transformers (Wolf et al., 2020) and scikit-learn (Pedregosa et al., 2011) libraries, using the *MultiLabelBinarizer* approach. All experiments were conducted on two Nvidia RTX 4090 GPUs, each featuring 24GB of VRAM.

Models Hyperparameters For the classification with the DeBERTa models, the following hyperparameters were employed: *batch size* = 16, *number of epochs* = 10, *maximum sequence length* = 512, *learning rate* = 2×10^{-5}, and *weight decay* = 0.01. Predictions were made using a threshold of 0.8 for logits, considering the number of labels and the multi-label classification setup. For classification with GPT4o-mini, the configuration included a *temperature* of 0.7, *top-p* = 0.95, and *maximum completion tokens* = 200. Lastly, for translation, the parameters were set as follows: *max new tokens* = 4000, *do sample* = True, *temperature* = 0.8, and *top-p* = 0.95, Lastly, all random seeds were set to 42 wherever applicable. The training and inference parameters were selected based on a 5-fold cross-validation performed on the training set.

Evaluation Metrics The evaluation metrics employed are based on the sample-level F1 score, with an emphasis on $F1_{Samples}$, which focuses on sub-narratives, rather than $F1_{Coarse}$, which targets the narratives-level, in alignment with the official task directives.

6 Results

This section aims to analyze the proposed strategies. Table 2 presents, for each language, the results of each proposed configuration on the development set, alongside the baseline provided by the Task organizers. Additionally, for the test set, the table displays the results of the final submitted strategy, the baseline, and the best overall performance achieved by any team for each language. Notably, our analysis focuses on the $F1_{Samples}$ score, as it is the main metric adopted for the Subtask.

First, all proposed strategies outperform the baseline. Notably, the most basic approach – the Single

Table 2: $F1_{Coarse}$ and $F1_{Samples}$ values for each language on the dev. and test sets. Best results in **bold**.

Strategy	Bulgarian		English		Hindi		Portuguese		Russian	
	F1 Coarse	F1 Samples	F1 Coarse	F1 Samples	F1 Coarse	F1 Samples	F1 Coarse	F1 Samples	F1 Coarse	F1 Samples
Validation										
Single-model	0.206	0.183	0.284	0.266	0.124	0.075	0.275	0.168	0.279	0.146
Translation *(Aya-8B)*	0.319	0.178	0.328	0.179	0.321	0.161	0.458	**0.289**	0.328	0.149
Translation *(GPT-4o-mini)*	0.347	**0.186**	0.304	0.176	0.320	**0.173**	0.371	0.228	0.354	0.174
Hierarchical	0.553	0.162	0.268	**0.268**	0.313	0.101	0.466	0.032	0.375	**0.219**
Baseline	0.038	0.014	0.106	0.000	0.100	0.051	0.067	0.010	0.041	0.013
Test										
Best Team	0.631	**0.460**	0.590	**0.438**	0.569	**0.535**	0.664	**0.480**	0.709	**0.518**
Baseline	0.056	0.022	0.030	0.013	0.081	0.000	0.037	0.014	0.065	0.008
Translation *(GPT-4o-mini)*	0.366	0.183	0.335	0.188	0.234	0.110	0.435	0.225	0.359	0.191

Model Strategy – achieves its best results in English and Bulgarian, while ranking among the least effective solutions for the other languages. This pattern may indicate that the need for a single classifier to adapt simultaneously to language representation and classification itself during training may hinder its ability to generalize performance across languages.

On the other hand, the Translation-Based Strategy yields the best results among the proposed approaches for Portuguese (with translations generated by the Aya model), as well as Bulgarian and Hindi (with translations generated by the GPT-4o-mini model). Additionally, the Aya-translated approach secures second place in Hindi, while the GPT-based variant achieves this same ranking for Portuguese and Russian. Notably, translation-based strategies do not exhibit abysmal performance in any language — unlike, for example, the Single Model Strategy in Hindi. This observation suggests that translation-based approaches may enhance generalization by allowing the final classifier to focus solely on the classification, rather than concurrently handling multilingual representation. Regarding the performance differences between the Aya and GPT-4o-mini translations, a more detailed analysis of their pre-training corpora could offer valuable insights. However, such resources are not publicly available for the GPT model.

Despite achieving the best performance in English and Russian, the Hierarchical Strategy demonstrated poor Portuguese results and was outperformed by the translation-based strategies in other languages. An analysis of the first stage of the Hierarchical Strategy on the validation set (as test labels are not available) suggests that the results may be influenced by a specific characteristic in Portuguese data. Table 3 reveals a high concentration of documents from the *Climate Change* domain in Portuguese, a pattern not shared by any other language (Appendix F provides the corresponding matrices for the remaining languages).

(a) Russian

True / Pred.	URW	CC	Other
URW	16	1	11
CC	0	0	0
Other	0	0	4

(b) Portuguese

True / Pred.	URW	CC	Other
URW	8	0	1
CC	0	25	0
Other	1	0	0

Table 3: Confusion matrices for domain classification in Russian and Portuguese on the validation set.

The table also shows performance for Russian, which, despite exhibiting a higher number of absolute classification errors compared to Portuguese, yielded better results in the final classification stage, as shown in Table 2. Though seemingly counter-intuitive, this observation indicates that the LLM-based final classification struggled specifically with the *Climate Change* domain in Portuguese. In contrast, Russian domain predictions related to the *Ukraine–Russia War* were more frequently classified correctly in the second stage of the Hierarchical Strategy. Additionally, we attribute the poor domain classification performance for Russian to the extreme class imbalance in the data for that language, as previously shown in Figure 1. This indicates the potential of future work to examine

intra-domain classification behavior in multilingual scenarios more closely.

Consequently, aiming to evaluate the most generalizable approach, our final submission was based on the Translation-Based Strategy utilizing the GPT-4o-mini model. In the test set, the submitted approach once again outperformed the baseline across all languages, consistently avoiding any notably poor results in any of them. This further reinforces its potential for generalization. Future research may also consider a qualitative evaluation of the translations, which we regard as beyond the scope of the current work.

7 Conclusion

This work addresses SemEval-2025 Task 10 and evaluates three distinct strategies for multilabel classification within a two-level taxonomy of narratives and sub-narratives in online news. Our results indicate that the approach relying solely on a multilingual SLM to classify texts in multiple languages failed to generalize its strong performance in languages such as English to other linguistic contexts. Similarly, the strategy that employed a multilingual SLM as a domain filter and an LLM to assign the final labels achieved the best result on the development set for English but performed poorly in languages such as Portuguese, also highlighting a generalization gap. Conversely, the approach that translated all texts into English and utilized a monolingual SLM for classification demonstrated more consistent and generalizable results across languages, frequently ranking as the best or second-best performing strategy among those we analyzed.

Future work may explore the evaluation of additional combinations of SLMs and LLMs to identify the most effective pairings for this task. Furthermore, the Translation-Based Approach, which demonstrated robust generalization, could be extended by translating texts into target languages other than English. This would enable a more comprehensive analysis of the impact of the translation step on classification performance across diverse linguistic contexts.

Limitations

The translations were conducted exclusively with English as the target language. While this decision was made to ensure the feasibility of the experiments, it may have hindered the evaluation of culturally specific and critical nuances inherent to each language. Another notable limitation of this study is the lack of in-depth qualitative analyses of the predictions and translations generated by the proposed approaches. While potentially complex due to the large number of possible labels and the high degree of subjectivity involved — and, in the specific context of this work, also combined with the time limit of the task — such analyses may be important, as they could offer valuable insights into narrative detection and potentially reveal manipulation strategies.

Ethics Statement

Language Bias The Translation-based strategy, while necessary for multilingual analysis, may introduce biases due to potential discrepancies between translated and naturally occurring language. Additionally, the underrepresentation of labels in non-English languages and the inherent bias towards English in terms of available models and resources could compromise the fairness and effectiveness of our methodologies.

Misclassification The politically sensitive nature of the topics — climate change and the Ukraine-Russia war — increases the risks associated with misclassification. Misidentifying disinformation could inadvertently amplify its spread, while overzealous identification could stifle legitimate discourse and censor genuine activism. Ongoing collaboration with linguists and social scientists could better capture the complexities of human language in social interaction and regular reevaluation of the narratives and labels in the corpus could be essential to ensure that our research remains relevant and ethically sound.

Acknowledgements

This research was financed by CNPq (National Council for Scientific and Technological Development), grant 307088/2023-5, FAPERJ – *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, processes SEI-260003/002930/2024, SEI-260003/000614/2023, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. This work was also supported by compute credits from a Cohere Labs Research Grant. These grants are intended to support academic partners conducting research aimed at releasing scientific artifacts and data for socially beneficial projects.

References

Kyle Anderson. 2019. Truth, lies, and likes: Why human nature makes online misinformation a serious threat (and what we can do about it). *Law & Psychol. Rev.*, 44:209.

Travis G. Coan, Constantine Boussalis, John Cook, and Mirjam O. Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, 11(1):22320.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024. Large Language Models for Propaganda Span Annotation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14522–14532, Miami, Florida, USA. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. DeBERTa: Decoding-Enhanced Bert With Disentangled Attention. In *International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, and Alex Baker-Whitcomb et al. 2024. Gpt-4o system card. *arXiv:2410.21276*.

D.G. Jones. 2024. Detecting Propaganda in News Articles Using Large Language Models. *Eng OA*, 2(1):01–12.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Bonka Kotseva, Irene Vianini, Nikolaos Nikolaidis, Nicolò Faggiani, Kristina Potapova, Caroline Gasparro, Yaniv Steiner, Jessica Scornavacche, Guillaume Jacquet, Vlad Dragu, Leonida della Rocca, Stefano Bucci, Aldo Podavini, Marco Verile, Charles Macmillan, and Jens P. Linge. 2023. Trend analysis of covid-19 mis/disinformation narratives–a 3-year study. *PLOS ONE*, 18(11):1–26.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Jiateng Liu, Lin Ai, Zizhou Liu, Payam Karisani, Zheng Hui, Yi Fung, Preslav Nakov, Julia Hirschberg, and Heng Ji. 2025. PropaInsight: Toward Deeper Understanding of Propaganda in Terms of Techniques, Appeals, and Intent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5607–5628, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *New York: Data & Society Research Institute*, pages 7–19.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12(null):2825–2830.

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.

Jan Piskorski, Nikos Nikolaidis, Nicolas Stefanovitch, Biliana Kotseva, Ilaria Vianini, Shirin Kharazi, and Jens Linge. 2022. Exploring data augmentation for classification of climate change denial: Preliminary

study. In *Proceedings of the Text2Story'22 Workshop, Stavanger (Norway), 10-April-2022*, volume 3117 of *CEUR Workshop Proceedings*, pages 97–109. JRC128613.

Nicholas Rutheford. 2023. About mis-disinformation, its potential impacts, and the challenges to finding effective countermeasures. *Laboratoire sur l'intégrité de l'information*.

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large Language Models for Propaganda Detection. *Preprint*, arXiv:2310.06422.

Jason Stanley. 2015. *How Propaganda Works*. Princeton University Press, Princeton, NJ.

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).

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Arthur Vasconcelos, Luiz Felipe De Melo, Eduardo Goncalves, Eduardo Bezerra, Aline Paes, and Alexandre Plastino. 2024. BAMBAS at SemEval-2024 Task 4: How far can we get without looking at hierarchies? In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 455–462, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Alessandro Zangari, Matteo Marcuzzo, Matteo Rizzo, Lorenzo Giudice, Andrea Albarelli, and Andrea Gasparetto. 2024. Hierarchical text classification and its foundations: A review of current research. *Electronics*, 13(7).

A Translation Prompt

TRANSLATE THE FOLLOWING TEXT INTO ENGLISH. BE AS PRECISE AS POSSIBLE IN RETAINING THE INFORMATION CONVEYED.

TEXT
{TEXT}

B Hierarchical Strategy Algorithm

Algorithm 1 Hierarchical Classification

Require: Article text T
Ensure: Final label set $\mathcal{L}$
Step 1: Domain Classification with SLM
Load pretrained SLM (Multilingual DeBERTa)
Define label set {URW, CC, Other}
Predict domain $D \leftarrow \text{SLM}(T)$
if $D =$ Other **then**
    Assign $\mathcal{L} \leftarrow$ {"Other-Other"}
else
    Step 2: Sub-Narrative Classification with LLM
    Select LLM (GPT4o-mini)
    Select appropriate prompt P based on D:
    if $D =$ URW **then**
        $P \leftarrow$ Prompt for URW sub-narr. classification
    else if $D =$ CC **then**
        $P \leftarrow$ Prompt for CC sub-narr. classification
    end if
    Predict sub-narrative labels $\mathcal{L}_{sub} \leftarrow \text{LLM}(P, T)$
    Combine domain label with sub-narrative labels:
    $\mathcal{L} \leftarrow \{D\} \cup \mathcal{L}_{sub}$
end if
return $\mathcal{L}$

C Illustrative Case of Incorrect Classification by GPT4o-mini

The excerpt below is taken from one of the documents made available in the task dataset. While the human annotators labeled it as *"Criticism of climate movement: Ad hominem attacks on key activists"*, our early experiments showed that the language model assigned the labels *"Amplifying Climate Fears: Amplifying existing fears of global warming"* and *"Criticism of climate movement: Climate movement is alarmist"*. Although these labels are semantically plausible — particularly

considering the quotation attributed to Greta Thunberg within the text — we conjecture that the language model failed to interpret the pragmatic function of the indirect citation. Specifically, it may not have recognized that the quote was employed not to convey the activist's message, but rather to undermine her credibility, as noted by human annotators. Therefore, enhancing models' capacity for pragmatic understanding may constitute a valuable direction for future research on narrative classification and persuasive discourse identification.

> " [...] '*A top climate scientist is warning that climate change will wipe out all of humanity unless we stop using fossil fuels over the next five years.*'
>
> Thunberg shared a now-deleted Grit Post article by Scott Alden citing a prediction from James Anderson [...]' "

D URW Classification Prompt

IN THE FOLLOWING TEXT, IDENTIFY THE CORE NARRATIVE THAT ALIGNS WITH THE AUTHOR'S PERSPECTIVE. CLASSIFY IT BASED ON THE OPTIONS IN THE LIST BELOW. IF THE NARRATIVE IN THE TEXT FALLS OUTSIDE THE LIST, ANSWER "OTHER".

OPTIONS LIST (NARRATIVES AND SUBNARRATIVES)
BLAMING THE WAR ON OTHERS
- UKRAINE IS THE AGGRESSOR
- THE WEST ARE THE AGGRESSORS
- OTHER

DISCREDITING UKRAINE
- REWRITING UKRAINE'S HISTORY
- DISCREDITING UKRAINIAN NATION AND SOCIETY
- DISCREDITING UKRAINIAN MILITARY
- DISCREDITING UKRAINIAN GOVERNMENT AND OFFICIALS AND POLICIES
- UKRAINE IS A PUPPET OF THE WEST
- UKRAINE IS A HUB FOR CRIMINAL ACTIVITIES
- UKRAINE IS ASSOCIATED WITH NAZISM
- SITUATION IN UKRAINE IS HOPELESS
- OTHER

RUSSIA IS THE VICTIM
- THE WEST IS RUSSOPHOBIC
- RUSSIA ACTIONS IN UKRAINE ARE ONLY SELF-DEFENCE
- UA IS ANTI-RU EXTREMISTS
- OTHER

PRAISE OF RUSSIA
- PRAISE OF RUSSIAN MILITARY MIGHT
- PRAISE OF RUSSIAN PRESIDENT VLADIMIR PUTIN
- RUSSIA IS A GUARANTOR OF PEACE AND PROSPERITY
- RUSSIA HAS INTERNATIONAL SUPPORT FROM A NUMBER OF COUNTRIES AND PEOPLE
- RUSSIAN INVASION HAS STRONG NATIONAL SUPPORT
- OTHER

OVERPRAISING THE WEST
- NATO WILL DESTROY RUSSIA
- THE WEST BELONGS IN THE RIGHT SIDE OF HISTORY
- THE WEST HAS THE STRONGEST INTERNATIONAL SUPPORT
- OTHER

SPECULATING WAR OUTCOMES
- RUSSIAN ARMY IS COLLAPSING
- RUSSIAN ARMY WILL LOSE ALL THE OCCUPIED TERRITORIES
- UKRAINIAN ARMY IS COLLAPSING
- OTHER

DISCREDITING THE WEST, DIPLOMACY
- THE EU IS DIVIDED
- THE WEST IS WEAK
- THE WEST IS OVERREACTING
- THE WEST DOES NOT CARE ABOUT UKRAINE, ONLY ABOUT ITS INTERESTS
- DIPLOMACY DOES/WILL NOT WORK
- WEST IS TIRED OF UKRAINE
- OTHER

NEGATIVE CONSEQUENCES FOR THE WEST
- SANCTIONS IMPOSED BY WESTERN COUNTRIES WILL BACKFIRE
- THE CONFLICT WILL INCREASE THE UKRAINIAN REFUGEE FLOWS TO EUROPE
- OTHER

DISTRUST TOWARDS MEDIA
- WESTERN MEDIA IS AN INSTRUMENT OF PROPAGANDA
- UKRAINIAN MEDIA CANNOT BE TRUSTED
- OTHER

AMPLIFYING WAR-RELATED FEARS
- BY CONTINUING THE WAR WE RISK WWIII
- RUSSIA WILL ALSO ATTACK OTHER COUNTRIES
- THERE IS A REAL POSSIBILITY THAT NUCLEAR WEAPONS WILL BE EMPLOYED
- NATO SHOULD/WILL DIRECTLY INTERVENE

- Other

Text
{text}

Provide only the **most** relevant narratives that best fit the text's intent.
If multiple narratives are equally significant, include them all.
Answer **only** with the classifications and always include narratives and subnarratives.

E Climate Change Classification Prompt

In the following text, identify the core narrative that aligns with the author's perspective. Classify it based on the options in the list below. If the narrative in the text falls outside the list, answer "Other".

Options List (Narratives and subnarratives)
Criticism of climate policies
- Climate policies are ineffective
- Climate policies have negative impact on the economy
- Climate policies are only for profit
- Other

Criticism of institutions and authorities
- Criticism of the EU
- Criticism of international entities
- Criticism of national governments
- Criticism of political organizations and figures
- Other

Climate change is beneficial
- CO2 is beneficial
- Temperature increase is beneficial
- Other

Downplaying climate change
- Climate cycles are natural
- Weather suggests the trend is global cooling
- Temperature increase does not have significant impact
- CO2 concentrations are too small to have an impact
- Human activities do not impact climate change
- Ice is not melting
- Sea levels are not rising
- Humans and nature will adapt to the changes
- Other

Questioning the measurements and science
- Methodologies/metrics used are unreliable/faulty
- Data shows no temperature increase
- Greenhouse effect/carbon dioxide do not drive climate change
- Scientific community is unreliable
- Other

Criticism of climate movement
- Climate movement is alarmist
- Climate movement is corrupt
- Ad hominem attacks on key activists
- Other

Controversy about green technologies
- Renewable energy is dangerous
- Renewable energy is unreliable
- Renewable energy is costly
- Nuclear energy is not climate friendly
- Other

Hidden plots by secret schemes of powerful groups
- Blaming global elites
- Climate agenda has hidden motives
- Other

Amplifying Climate Fears
- Earth will be uninhabitable soon
- Amplifying existing fears of global warming
- Doomsday scenarios for humans
- Whatever we do it is already too late
- Other

Green policies are geopolitical instruments
- Climate-related international relations are abusive/exploitative
- Green activities are a form of neo-colonialism
- Other

Text
{text}

Provide only the most relevant narratives that best fit the text's intent.
If multiple narratives are equally significant, include them all.

ANSWER **ONLY** WITH THE CLASSIFICATIONS AND ALWAYS INCLUDE NARRATIVES AND SUBNARRATIVES.

F Confusion matrices for domain classification

(a) Bulgarian

True / Pred.	URW	CC	Other
URW	15	0	1
CC	0	13	0
Other	2	4	0

(b) English

True / Pred.	URW	CC	Other
URW	10	0	3
CC	0	14	3
Other	1	2	8

(c) Hindi

True / Pred.	URW	CC	Other
URW	25	0	4
CC	0	4	0
Other	1	0	1

Table 4: Confusion matrices for domain classification in Bulgarian, English, and Hindi on the validation set.