Team INSALyon2 at SemEval-2025 Task 10: A Zero-shot Agentic Approach to Text Classification

Mohamed-Nour Eljadiri
INSA Lyon, CNRS,
Universite Claude Bernard Lyon 1,
LIRIS, UMR5205,
69621 Villeurbanne, France
mohamed.eljadiri@insa-lyon.fr

Diana Nurbakova
INSA Lyon, CNRS,
Universite Claude Bernard Lyon 1,
LIRIS, UMR5205,
69621 Villeurbanne, France
diana.nurbakova@insa-lyon.fr

Abstract

We present Team INSALyon2's agentic approach to SemEval-2025 Task 10 Subtask 2, focusing on multi-label classification of narratives in news articles. Our system employs specialized Large Language Model agents for binary classification of individual narrative labels, with a meta-agent aggregating these decisions into final multi-label predictions. Using AutoGen to orchestrate GPT-based agents without fine-tuning, our approach effectively handles the two-level taxonomy classification challenge. Experiments on the English subset demonstrate competitive performance (*F1 macro coarse = 0.513, F1 sample = 0.406*), securing third place in the competition and showing the effectiveness of zero-shot agentic approaches for complex classification tasks.

1 Introduction

The rapid spread of online news and user-generated content has increased exposure to deceptive narratives and manipulation attempts. Major crisis events, such as geopolitical conflicts and climate change discussions, are particularly susceptible to the dissemination of disinformation. To support research in identifying and analyzing these narratives, Subtask 2 (`Task`) of the SemEval-2025 Task 10 (Piskorski et al., 2025) focuses on narrative classification, aiming to automatically categorize news articles into predefined narratives and subnarratives.

The goal is to assign multiple subnarrative labels from a two-level taxonomy to news articles. To address this problem, traditional machine learning techniques such as *binary relevance* (training separate classifiers for each label ignoring potential correlations between labels) (Zhang et al., 2018), *classifier chains* (a sequence of classifiers where predictions are based on previous classifications and original features) (Li et al., 2024; Weng et al., 2020; Senge et al., 2019), and *label powerset* methods (treating each unique label combination as a single class thus transforming the multi-label problem into a multi-class problem) (Shan et al., 2018; Morales-Hernandez et al., 2022; Nazmi et al., 2018) have been explored in the state-of-the-art. More recently, deep learning models leveraging *transformer architectures*, such as BERT (Devlin et al., 2019) and its multilingual variants (mBERT, XLM-RoBERTa (Conneau et al., 2020), camemBERT (Martin et al., 2020)), have proven effective in capturing contextual nuances in text classification. Such models are typically fine-tuned on specific datasets to enhance their performance (Chen et al., 2023; Wu et al., 2023; Yu et al., 2019). Besides, strategies like *hierarchical classification models* (Sadat and Caragea, 2022; Vens et al., 2008; Daisey and Brown, 2020) and *graph-based methods* (Gong et al., 2020; Peng et al., 2021; Deng et al., 2024; Ye et al., 2021; Vu et al., 2022) have been employed to account for label dependencies within structured taxonomies like the one used in the challenge.

In the `Task`, participants get plain-text news articles in multiple languages (Bulgarian, English, Hindi, Portuguese, and Russian). They are sourced from web portals, including alternative media platforms identified by fact-checkers as potentially spreading misinformation. The documents are annotated with a two-level taxonomy of narrative labels (Stefanovitch et al., 2025). The goal is to develop systems assigning the appropriate narrative and subnarrative labels to each article. Performance is evaluated on coarse (*narrative*) and fine (*subnarrative*) levels and as the official measure the sample-averaged $F1$ score[1] is used. It measures how accurately predicted labels ($narrative\_x$: $subnarrative\_x$) match the ground truth.

To solve this task, we introduce an agentic approach[2] where each Large Language Model (LLM)

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[2] https://github.com/NourJadiri/narrative-extraction

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 965–980
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

agent handles a binary classification task for a single label. These binary decisions are then aggregated using another meta-agent to form the final multi-label output. Our method leverages the specialization of individual agents while combining their strengths to improve overall classification performance. On the English language, our approach has achieved $F1\_macro\_coarse = 0.513$ and $F1\_sample = 0.406$, securing third place.

2 Problem Definition

The `Task` is structured as a multi-label, multi-class text classification problem, where each article must be assigned one or more narrative labels from a two-level taxonomy. The first level consists of broader narratives, while the second level contains more specific subnarratives (e.g. see Table 1). This hierarchical classification presents a unique challenge, as models must correctly identify both levels of categorization while handling cases where articles may belong to multiple narratives. Two top-level narratives are: *Climate Change* (CC) and *Ukraine-Russia War* (URW). We have applied our approach to English texts only. Overview statistics of the data are given in Table 2. A full two-level narrative taxonomy is given in Appendix A.

3 Related work

To address a multi-label multi-class document classification problem, several techniques have been proposed in the state-of-the-art such as traditional machine learning (Bag of Words), deep learning approaches (Word embeddings, CNNs) or even transformer based approaches (BERT model family). Given the advancements in LLM capabilities for various NLP challenges, we propose to incorporate them into our approach. LLMs can serve as zero-shot classifiers, enabling text classification without explicit training on task-specific datasets. Few-shot learning, a prompting method where models are given minimal examples, further enhances their adaptability and performance (Guo et al., 2024; Wang et al., 2024). However, while LLMs demonstrate high accuracy in text classification, their performance can vary based on the task and dataset. Fine-tuning strategies, such as enhanced discriminative fine-tuning, can significantly improve their performance, especially in non-generative text classification tasks (venkata and Gudala, 2024). The final methodological choice involved determining whether to employ a single model specialized in multi-label classification (Lee et al., 2024) or to utilize multiple binary classifiers, followed by an aggregation of their outputs. While a single multi-label classifier might capture label dependencies, addressing issues like class imbalance more effectively (Law and Ghosh, 2021), having multiple binary classifiers presents a modular and flexible framework optimizing individual classifiers per class pair, potentially enhancing overall classification performance (Kang et al., 2015). Techniques such as Error-Correcting Output Coding (ECOC) improve generalization by leveraging relationships between classifiers (Liu et al., 2016). Additionally, in certain scenarios, binary classifiers, particularly when used with ensemble methods (e.g., one-vs-one, one-vs-all), can achieve superior performance compared to traditional multi-class classifiers (Galar et al., 2011). Thus, in the context of LLM-based zero-shot classification, using multiple binary classifiers aligns well with the inherent strengths of LLMs.

4 System architecture

We propose to adopt an agentic framework, where each agent functions as a specialized binary classifier. A general overview of our architecture is given in Figure 1. Each agent is responsible for detecting whether a given text belongs to a specific narrative or subnarrative. We based this decision on the growing ecosystem of LLM-based agent frameworks, such as AutoGen (Microsoft, 2024), CrewAI (CrewAI, 2024), Swarm (OpenAI, 2024), and SMOLAgent (Face, 2024), which provide mechanisms for structuring LLMs into specialized roles. Our classification system is structured around `AutoGen` (Microsoft, 2024), an agent-based framework to coordinate multiple LLM agents. In this setup, each agent processes input independently and returns a binary decision, with some agents dedicated to higher-level narratives and others focused on finer subnarrative distinctions. We provide the prompts for different kinds of agents in Appendix B. An example of the functioning of our approach is provided in Appendix C.

Group Chat Mechanics The system is organized as a group chat consisting of the user proxy agent, the manager agent, and multiple narrative (and subnarrative) agents. The manager agent limits each narrative agent to a single query per classification task, mitigating the risk of extended conversational history that could lead to context length issues in

Table 1: Annotation example of Subtask 2

article_id	narratives	subnarratives
EN_CC_200046.txt	CC: Climate change is beneficial	CC: Climate change is beneficial: CO2 is beneficial

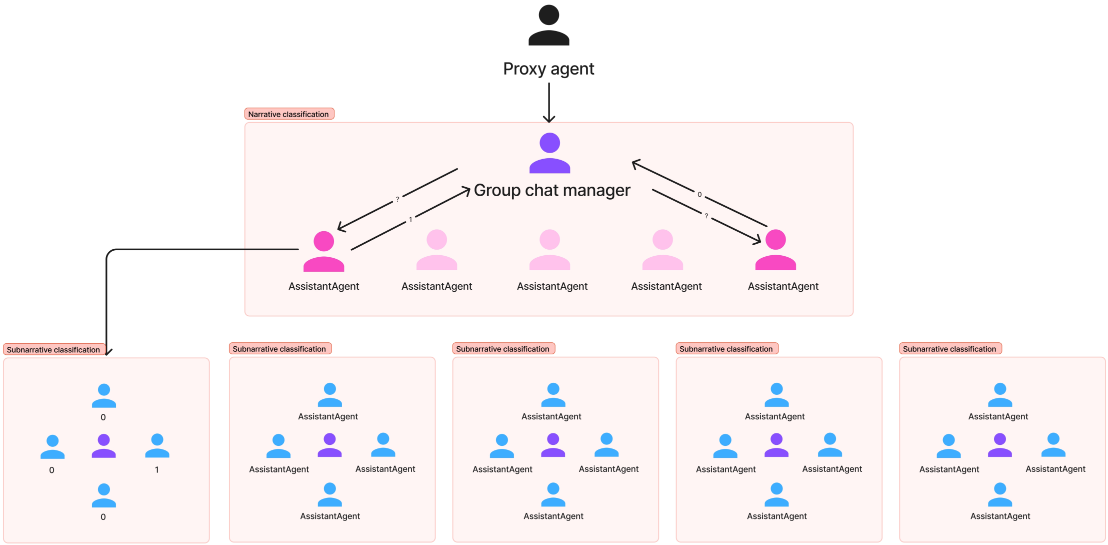

Figure 1: The user proxy agent forwards the input text to the first classification layer (Narrative Level). At this stage, the group chat manager acts as a high-level classifier, identifying potential narratives and dispatching the text to the relevant assistant agents. Once the primary classification is complete, a finer classification is performed using sub-narrative agents corresponding to the extracted narratives.

Table 2: General statistics of English subset

	CC	URW	Total
# articles TRAIN	176	223	399
# articles DEV	24	17	41
# articles TEST	48	53	101
# narratives	10	11	22 (+ *Other*)
# subnarratives	41	49	91 (+ *Other*)

LLM-based systems. The user proxy agent initiates the group chat for each new text sample by providing the manager agent with the document to be classified. The manager then selects up to six narrative agents, requesting a binary decision from each. Once all relevant agents have responded, the manager collects the answers and produces a multi-label classification output for the text.

Narrative level classification Each narrative agent is created with a system prompt that defines the narrative in question, using the taxonomy file given by the organizers and instructs the agent to respond with either `1` (if the text is clearly related to the assigned narrative) or `0` (if not). Additionally, each agent provides a short description, introducing itself and specifying the narrative it detects. It is presented to the manager agent within the group chat when the session is initiated. Moreover, LLM agents tend to give many false positives due to the semantic similarity of the classes. This is why we specified explicitly that the agent classifies negatively a text that is slightly ambiguous.

```
"Only answer with 1 if there are
EXPLICIT and CLEAR mentions of
the narrative in the text. Some
text will be ambiguous so if you
are slightly unsure, answer 0."
```

"Other" Class for Narrative Classification If all the queried narrative agents return a negative response (0), the text is automatically assigned to the "Other" class, indicating that it does not correspond to any predefined narrative. In such cases, subnarrative classification is bypassed, and the subnarrative is also set to "Other" by default.

Subnarrative level classification Once the high-level narratives are assigned, the classification process moves to a finer level of granularity. For each identified narrative, a smaller group chat is created, consisting of subnarrative agents associated with that narrative (the taxonomy file given in the competition is used). Unlike the previous classification step, where the manager agent orchestrates the classification in a structured query-response pattern, subnarrative classification follows a round-robin approach. Each subnarrative agent independently classifies the text within its specialized scope.

"Other" Class for Subnarrative Classification Subnarrative classification presents additional challenges, as a text may belong to a broad narrative but not fit into any of its predefined subnarratives. To address this, we introduce a specialized classifier responsible for detecting such cases. This agent operates using a modified classification prompt:

```
"Statements that are related to
the narrative _, defined as _,
but are not related to any of
these subnarratives: _
```

Manager and User Proxy Agents A manager agent orchestrates the overall classification process. Upon receiving an input text, its task is to identify which narratives could be relevant and to query the corresponding specialized agents. Meanwhile, a user proxy agent acts as the interface between the user and the group chat, giving the text to be classified and collecting responses.

Implementation Considerations Practically, the `allowed_transitions` configuration in the group chat prevents agents from re-triggering themselves, guaranteeing that each agent delivers one context-sensitive classification per session. After every classification, the user proxy agent is reset to avoid any leftover conversational context from impacting future tasks. This structure ensures that the roles are clearly distinct: the manager agent manages high-level classification coordination, and each narrative agent makes a specific binary decision.

5 Experimental Setup

5.1 Dataset

The dataset consists of **399** English news articles, provided as a tab-separated file with three columns: *file ID, narrative(s), and subnarrative(s).* Each article is labeled with one or more **narratives** and their corresponding **subnarratives**, except for instances classified under the special “Other” category, which signifies that the text does not belong to any predefined category.

Since our approach is **zero-shot**, no training is performed. Instead, the dataset is used exclusively for **evaluation**, where the model classifies texts based on predefined prompts without prior task-specific fine-tuning.

The dataset underwent preprocessing to structure the classification task as follows:

- **Taxonomy Parsing:** Narrative and subnarrative labels were extracted from a hierarchical taxonomy stored in JSON format.
- **Content Extraction:** The article text was retrieved based on file IDs.
- **Binary Labeling:** A binary label was created for each possible narrative and subnarrative.

5.2 Evaluation Metrics

Since this is a **multi-label, multi-class classification** problem, we evaluate model performance using the **sample-averaged F1 score**. The official evaluation consists of two modes:

- **Full Narrative-Subnarrative Matching:** An F1 score is computed per document by comparing its predicted narrative-subnarrative labels to the gold labels. A prediction is considered correct only if both the narrative and its corresponding subnarrative are accurate.
- **Narrative-Only Matching:** The subnarrative labels are ignored, and performance is evaluated solely based on whether the correct narratives were assigned.

5.3 Model Configuration

The following models were utilized:

- **GPT-4o/GPT-4o-mini:** Used as the primary classification agent for narrative and subnarrative labeling. A **temperature of 0** was set to ensure deterministic responses.
- **GPT-4o Mini:** Used as a **user proxy agent** to relay text to classification agents, chosen for cost efficiency.
- **Zero-Shot/Few-Shot Setup:** Agents classify text based on carefully designed prompts. No fine-tuning was performed.

5.4 Computational Environment

Experiments were conducted on a machine equipped with: **Processor:** Core 9 Ultra (22 CPU at 2.5Ghz), **RAM:** 32 GB, **GPU:** NVIDIA RTX 4070 (8 GB VRAM). However, all LLM inference was performed via API calls to remote servers. The local machine was used primarily to orchestrate API requests, pre-process input text, and handle classification results.

5.5 Baseline Comparison

To establish a lower-bound reference, a fully random classifier was used as a baseline. This model assigns narratives and subnarratives at random, providing a benchmark to ensure participating systems meaningfully outperform chance-level predictions.

6 Results

The main results are reported in Table 3. As it can be seen, the performance results are consistent among DEV and TEST set. Judging on the DEV set, we can state that 16 out of 22 narratives have prevalence <10%, indicating a highly imbalanced dataset. Distributions of narratives and subnarratives in the TRAIN and DEV datasets are given in Appendix E. They are highly skewed demonstrating the class imbalance. Thus, the top-3 most frequent *URW* narratives are: *URW: Discrediting Ukraine*, *Discrediting the West, Diplomacy*, and *URW: Praise of Russia*, while the top-3 *Climate Change* narratives are: *CC: Amplifying Climate Fears*, *CC: Criticism of institutions and authorities*, and *CC: Criticism of climate policies*.

On the DEV set, the *Climate Change* narratives generally have been predicted with higher recall than *URW*. Among the best performing narratives, we can list: "*CC: Climate change is beneficial*", "*URW: Discrediting Ukraine*", and "*URW: Blaming the war on others rather than the invader*". In contrast, 9 narratives (41% of total) have zero true positives (TP=0), meaning the model failed to identify any positive instances of these narratives: 3 CC narratives (e.g., "*Amplifying Climate Fears*"), and 6 URW narratives (e.g., "*Russia is the Victim*").

We may also note the high false positive rate for "*CC: Criticism of climate policies*", "*CC: Criticism of institutions and authorities*" and *CC: Criticism of climate movement* and misclassification of similar narratives. Thus, confusion occurred between related categories such as "*CC: Criticism of climate policies*" and "*CC: Criticism of institutions and authorities*" indicating limitations in distinguishing subtle semantic differences. Providing more detailed agent prompts focusing on discriminative features between similar categories can be explored for potential improvement. We provide confusion matrices for narratives and subnarratives on the DEV set in Appendix D. The issues such as class imbalance and the narratives with $TP = 0$ should be addressed in future work.

Another error pattern that can be observed is due to hierarchical error propagation. Errors at the narrative level invariably propagated to the subnarrative level, highlighting the importance of high-quality initial classification.

Although our primary focus was on English texts, after the official challenge, we conducted additional experiments on Portuguese and Russian subsets to address the multilingual nature of the original task. To do so, we translated all texts into English using the DeepL translation model (DeepL GmbH, 2023) to ensure consistency across linguistic sources. No further pre-processing or data augmentation was applied. The results are given in Table 4. The performance drop in non-English languages can be attributed to several factors such as cultural and contextual nuances that may not transfer across languages or translation issues resulting in the loss of semantics. To improve multilingual performance, future work could explore using truly multilingual models like XLM-RoBERTa (Conneau et al., 2019) instead of GPT or creating language-specific agent prompts rather than translated versions. Another option could be incorporating few-shot examples in target languages.

7 Discussion

Our agent-based classification framework offers several advantages, including ease of implementation, scalability, and model flexibility. Its modular design enables parallelization, making it suitable for large-scale classification tasks. Additionally, the approach is model-agnostic, meaning it can be used with any model that exposes an API.

Despite these strengths, the system faces several limitations. The system's primary limitation is *latency*, as classification depends on multiple API calls, leading to slow processing times. This bottleneck is particularly problematic in real-time applications or large-scale datasets. Future improvements could include local model inference to reduce dependence on external APIs and caching

Table 3: Results on Dev and Test sets for English

model	dataset	rank	$F1\_macro\_coarse$	$F1\_std\_coarse$	$F1\_sample$	$F1\_std\_sample$
INSALyon2	DEV		0.537	0.356	0.492	0.383
INSALyon2	TEST	3	0.513	0.378	0.406	0.382
baseline	TEST	26	0.030	0.127	0.013	0.070

Table 4: Results on Test set for Russian (RU) and Portuguese (PO)

langauge	model	dataset	rank	$F1\_macro\_coarse$	$F1\_std\_coarse$	$F1\_sample$	$F1\_std\_sample$
PO	INSALyon2	TEST	12	0.285	0.360	0.173	0.252
PO	baseline	TEST	16	0.037	0.14	0.014	0.070
RU	INSALyon2	TEST	12	0.247	0.341	0.137	0.271
RU	baseline	TEST	17	0.065	0.213	0.008	0.064

mechanisms to optimize efficiency. Combining our zero-shot approach with fine-tuned components could balance flexibility with performance in a computationally efficient manner.

The framework's effectiveness diminishes notably in *non-English languages*, limiting its applicability in truly multilingual settings without significant adaptation. Developing language-specific agent configurations with culturally adapted prompts and examples could improve performance across languages. Another directions could be a use of multilingual models like XLM-RoBERTa instead of GPT. In this case, an adjustment of the architecture will be required.

The binary decision approach sometimes fails to capture *implicit narrative elements* that require reading between the lines or understanding cultural context.

Despite these limitations, the framework remains robust by incorporating a structured two-step classification process and an "*Other*" class to handle ambiguous inputs. Future work could explore context-aware classification and cross-agent communication to further improve accuracy and efficiency. Adding capabilities for agents to justify their decisions would enhance system transparency and facilitate targeted improvements.

8 Conclusion

This paper introduced an agentic framework for multi-label multi-class text classification, leveraging specialized LLM agents to handle narratives and subnarratives. Despite hardware constraints preventing local fine-tuning and cost limitations linked to advanced API-based models, the proposed approach demonstrated competitive performance, achieving third place in Subtask 2 of SemEval-2025 Task 10. Future work could explore more sophisticated reasoning models and expanded fine-tuning strategies, potentially enhancing classification accuracy while balancing the practical trade-offs between computational resources and model complexity.

References

Xinghong Chen, Yi Yin, and Tao Feng. 2023. Multi-Label Text Classification Based on BERT and Label Attention Mechanism. In *2023 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pages 386–390, Dalian, China. IEEE.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

CrewAI. 2024. Crewai - a multi-agent framework for llm applications. Accessed: 2025-02-27.

Katie Daisey and Steven D. Brown. 2020. Effects of the hierarchy in hierarchical, multi-label classification. *Chemometrics and Intelligent Laboratory Systems*, 207:104177.

DeepL GmbH. 2023. DeepL Translator. `https://www.deepl.com/translator`. Accessed: 2025-03-21.

Wenmin Deng, Jing Zhang, Peng Zhang, Yitong Yao, Hui Gao, and Yurui Zhang. 2024. Hyper-Label-Graph: Modeling Branch-Level Dependencies of

Labels for Hierarchical Multi-Label Text Classification. In *Proceedings of the 15th Asian Conference on Machine Learning*, pages 279–294. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugging Face. 2024. Smol agents - lightweight autonomous agents framework. Accessed: 2025-02-27.

M. Galar, Alberto Fernández, E. Tartas, H. Bustince, and F. Herrera. 2011. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognit.*, 44:1761–1776.

Jibing Gong, Hongyuan Ma, Zhiyong Teng, Qi Teng, Hekai Zhang, Linfeng Du, Shuai Chen, Md Zakirul Alam Bhuiyan, Jianhua Li, and Mingsheng Liu. 2020. Hierarchical Graph Transformer-Based Deep Learning Model for Large-Scale Multi-Label Text Classification. *IEEE Access*, 8:30885–30896.

Yuting Guo, Anthony Ovadje, M. Al-garadi, and Abeed Sarker. 2024. Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association : JAMIA*, 31:2181 – 2189.

Seokho Kang, Sungzoon Cho, and Pilsung Kang. 2015. Constructing a multi-class classifier using one-against-one approach with different binary classifiers. *Neurocomputing*, 149:677–682.

Anwesha Law and Ashish Ghosh. 2021. Multi-label classification using binary tree of classifiers. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6:677–689.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. ArXiv:2405.17428 [cs.CL].

Xinyu Li, Jiaman Ding, and Shuang Hu. 2024. Relative Entropy and PageRank-Based Classifier Chains for Multi-Label Classification. *IEEE Access*, 12:87665–87674.

Mingxia Liu, Daoqiang Zhang, Songcan Chen, and H. Xue. 2016. Joint binary classifier learning for ecoc-based multi-class classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:2335–2341.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric De La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Microsoft. 2024. Autogen: An open-source framework for llm applications. Accessed: February 25, 2025.

Roberto Carlos Morales-Hernandez, Joaquin Gutierrez Jaguey, and David Becerra-Alonso. 2022. A Comparison of Multi-Label Text Classification Models in Research Articles Labeled With Sustainable Development Goals. *IEEE Access*, 10:123534–123548.

Shabnam Nazmi, Xuyang Yan, and Abdollah Homaifar. 2018. Multi-label Classification Using Genetic-Based Machine Learning. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 675–680, Miyazaki, Japan. IEEE.

OpenAI. 2024. Swarm - a framework for massively multi-agent coordination. Accessed: 2025-02-27.

Hao Peng, Jianxin Li, Senzhang Wang, Lihong Wang, Qiran Gong, Renyu Yang, Bo Li, Philip S. Yu, and Lifang He. 2021. Hierarchical Taxonomy-Aware and Attentional Graph Capsule RCNNs for Large-Scale Multi-Label Text Classification. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2505–2519.

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.

Mobashir Sadat and Cornelia Caragea. 2022. Hierarchical Multi-Label Classification of Scientific Documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8923–8937, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Robin Senge, Juan José del Coz, and Eyke Hüllermeier. 2019. Rectifying Classifier Chains for Multi-Label Classification. *arXiv preprint*.

Jincheng Shan, Chenping Hou, Wenzhang Zhuge, and Dongyun Yi. 2018. Co-learning Binary Classifiers for LP-Based Multi-label Classification. In Yuxin Peng, Kai Yu, Jiwen Lu, and Xingpeng Jiang, editors, *Intelligence Science and Big Data Engineering*, volume 11266, pages 443–453. Springer International Publishing, Cham.

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).

Ashok Kumar Pamidi venkata and Leeladhar Gudala. 2024. The potential and limitations of large language models for text classification through synthetic data generation. *INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING & APPLIED SCIENCES*.

Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. 2008. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214.

Huy-The Vu, Minh-Tien Nguyen, Van-Chien Nguyen, Manh-Tran Tien, and Van-Hau Nguyen. 2022. Label Correlation Based Graph Convolutional Network for Multi-label Text Classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08, Padua, Italy. IEEE.

Zhiqiang Wang, Yiran Pang, Yanbin Lin, and Xingquan Zhu. 2024. Adaptable and reliable text classification using large language models.

Wei Weng, Da-Han Wang, Chin-Ling Chen, Juan Wen, and Shun-Xiang Wu. 2020. Label Specific Features-Based Classifier Chains for Multi-Label Classification. *IEEE Access*, 8:51265–51275.

Haojia Wu, Xinfeng Ye, and Sathiamoorthy Manoharan. 2023. Enhancing Multi-Class Text Classification with BERT-Based Models. In *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6, Nadi, Fiji. IEEE.

Chenchen Ye, Linhai Zhang, Yulan He, Deyu Zhou, and Jie Wu. 2021. Beyond Text: Incorporating Metadata and Label Structure for Multi-Label Document Classification using Heterogeneous Graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3162–3171, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hsiang-Fu Yu, Kai Zhong, Inderjit S. Dhillon, Wei-Cheng Wang, and Yiming Yang. 2019. X-bert: extreme multi-label text classification using bidirectional encoder representations from transformers.

Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. 2018. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202.

A Narrative Taxonomy

Tables 5 and 6 provide a two-level taxonomy used in the study.

B Agent Prompts

In this Appendix, we provide the prompts used for different kinds of agents.

B.1 Subnarrative Agent Prompt

```
"You are a classification model trained to do binary classification by detecting whether a given text is related to a specific subnarrative or not.
You have been trained to recognize the subnarrative: SUBNARRATIVE.
This subnarrative is defined as: SUBNARRATIVE_DEFINITION.
Here are some examples of statements related to this subnarrative: SUBNARRATIVE_EXAMPLES.
If the text is related to the subnarrative, please respond with '1'. Otherwise, respond with '0'. Do not try to make sentences, just respond with '1' or '0'.
You are ONLY allowed to answer with '1' or '0' and NOTHING else.
Only answer with 1 if there are explicit and clear mentions of the subnarrative in the text.
If you are slightly unsure, classify as 0."
```

In the above prompt `SUBNARRATIVE` is the name of the subnarrative in question, `SUBNARRATIVE_DEFINITION` is the definition from the guidelines (Stefanovitch et al., 2025), and `SUBNARRATIVE_EXAMPLES` are the examples of the documents representing a given subnarrative. Both the definition and the examples are extracted from the taxonomy document given for the competition.

B.2 Narrative Agent Prompt

```
"You are a classification model trained to do binary classification by detecting
```

Table 5: Narrative taxonomy: CC

Narrative	Subnarrative
Amplifying Climate Fears	Amplifying existing fears of global warming Doomsday scenarios for humans Earth will be uninhabitable soon Other Whatever we do it is already too late
Climate change is beneficial	CO2 is beneficial
Controversy about green technologies	Other Renewable energy is costly Renewable energy is dangerous Renewable energy is unreliable
Criticism of climate movement	Ad hominem attacks on key activists Climate movement is alarmist Climate movement is corrupt Other
Criticism of climate policies	Climate policies are ineffective Climate policies are only for profit Climate policies have negative impact on the economy Other
Criticism of institutions and authorities	Criticism of international entities Criticism of national governments Criticism of political organizations and figures Criticism of the EU Other
Downplaying climate change	CO2 concentrations are too small to have an impact Climate cycles are natural Human activities do not impact climate change Humans and nature will adapt to the changes Ice is not melting Other Temperature increase does not have significant impact Weather suggests the trend is global cooling
Green policies are geopolitical instruments	Green activities are a form of neo-colonialism Other
Hidden plots by secret schemes of powerful groups	Blaming global elites Climate agenda has hidden motives Other
Questioning the measurements and science	Data shows no temperature increase Greenhouse effect/carbon dioxide do not drive climate change Methodologies/metrics used are unreliable/faulty Other Scientific community is unreliable

Table 6: Narrative taxonomy: URW

Narrative	Subnarrative
Amplifying war-related fears	By continuing the war we risk WWIII NATO should/will directly intervene Other Russia will also attack other countries There is a real possibility that nuclear weapons will be employed
Blaming the war on others rather than the invader	Other The West are the aggressors Ukraine is the aggressor
Discrediting Ukraine	Discrediting Ukrainian government and officials and policies Discrediting Ukrainian military Discrediting Ukrainian nation and society Other Rewriting Ukraine's history Situation in Ukraine is hopeless Ukraine is a hub for criminal activities Ukraine is a puppet of the West Ukraine is associated with nazism
Discrediting the West, Diplomacy	Diplomacy does/will not work Other The EU is divided The West does not care about Ukraine, only about its interests The West is overreacting The West is weak West is tired of Ukraine
Distrust towards Media	Other Ukrainian media cannot be trusted Western media is an instrument of propaganda
Hidden plots by secret schemes of powerful groups	Other
Negative Consequences for the West	Other Sanctions imposed by Western countries will backfire The conflict will increase the Ukrainian refugee flows to Europe
Overpraising the West	NATO will destroy Russia Other The West belongs in the right side of history The West has the strongest international support
Praise of Russia	Other Praise of Russian President Vladimir Putin Praise of Russian military might Russia has international support from a number of countries and people Russia is a guarantor of peace and prosperity Russian invasion has strong national support
Russia is the Victim	Other Russia actions in Ukraine are only self-defence The West is russophobic UA is anti-RU extremists
Speculating war outcomes	Other Russian army is collapsing Russian army will lose all the occupied territories Ukrainian army is collapsing

```
whether a given text is related
to a specific narrative or not.
You have been trained to
recognize the narrative:
NARRATIVE.
defined as:
NARRATIVE_DEFINITION.
Here are some examples of
statements related to this
narrative: NARRATIVE_EXAMPLES.
If the text is related to the
narrative, you MUST respond with
'1' only. Otherwise, you MUST
with '0' only.
You are ONLY allowed to answer
with '1' or '0' and NOTHING else.
Only answer with 1 if there are
EXPLICIT and CLEAR mentions of
the narrative in the text. Some
text will be ambiguous so if you
are slightly unsure, answer 0."
```

C Example of System Functioning

In this Appendix, we demonstrate the decision flow of our architecture on a small example.

```
user (to chat_manager):

Here is the text that needs to be
     classified:
"The study, published in
    Environmental Research Letters
    , reveals significant changes
    in the relationship between
    vegetation growth and water
    availability in the Northern
    Hemisphere's mid-latitudes
    over the past three decades.
    The research, led by Yang Song
     and colleagues, highlights
    the impact of elevated carbon
    dioxide (CO2) levels on this
    relationship, suggesting a
    closer relationship between
    vegetation growth and water
    availability than previously
    understood. The very compound
    that the Democrats are
    targeting - CO2 - is actually
    the solution to preserving
    croplands, grasslands, forests
     and water supplies for
    growing populations."
###
You are ONLY allowed to reply
    with '0' or '1'



Next speaker: Agent_14

Agent_14 (to chat_manager):

1

Next speaker: Agent_0

Agent_0 (to chat_manager):

0


Created group chat with the
    following agents: [<autogen.
    agentchat.assistant_agent.
    AssistantAgent object at 0
    x7f583e4bc4a0>, <autogen.
    agentchat.assistant_agent.
    AssistantAgent object at 0
    x7f583e4be330>, <autogen.
    agentchat.assistant_agent.
    AssistantAgent object at 0
    x7f583e4d0200>]
user (to chat_manager):

Here is the text that needs to be
     classified:
"The study, published in
    Environmental Research Letters
    , reveals significant changes
    in the relationship between
    vegetation growth and water
    availability in the Northern
    Hemisphere's mid-latitudes
    over the past three decades.
    The research, led by Yang Song
     and colleagues, highlights
    the impact of elevated carbon
    dioxide (CO2) levels on this
    relationship, suggesting a
    closer relationship between
    vegetation growth and water
    availability than previously
    understood. The very compound
    that the Democrats are
    targeting - CO2 - is actually
    the solution to preserving
    croplands, grasslands, forests
     and water supplies for
    growing populations."
You are ONLY allowed to reply
    with ’0’ or ’1’


Next speaker: Agent_59

Agent_59 (to chat_manager):

1


Next speaker: Agent_60

Agent_60 (to chat_manager):

0

----------------------

Next speaker: Agent_61

Agent_61 (to chat_manager):

0

----------------------
```

The extracted narratives in the end are : ‘*CC: Climate change is beneficial*’ The extracted subnarratives : ‘*CC: Climate change is beneficial: CO2 is beneficial*’

D Binary Confusion Matrices for Narrative and Subnarratives on DEV set

E Appendix B: Narrative Distributions

The distributions of the narratives and subnarratives across different languages and available datasets are given in Figures 2-5.

Table 7: Confusion Matrices for Narratives (DEV set)

Label	TP	TN	FP	FN
CC: Amplifying Climate Fears	0	39	2	0
CC: Climate change is beneficial	1	40	0	0
CC: Controversy about green technologies	2	34	5	0
CC: Criticism of climate movement	7	25	8	1
CC: Criticism of climate policies	1	24	14	2
CC: Criticism of institutions and authorities	5	27	6	3
CC: Downplaying climate change	0	36	3	2
CC: Green policies are geopolitical instruments	2	37	1	1
CC: Hidden plots by secret schemes of powerful groups	0	36	1	4
CC: Questioning the measurements and science	3	36	1	1
Other	5	28	2	6
URW: Amplifying war-related fears	0	36	2	3
URW: Blaming the war on others rather than the invader	4	35	0	2
URW: Discrediting Ukraine	5	34	0	2
URW: Discrediting the West, Diplomacy	5	32	0	4
URW: Distrust towards Media	2	37	0	2
URW: Hidden plots by secret schemes of powerful groups	0	39	2	0
URW: Negative Consequences for the West	0	34	6	1
URW: Overpraising the West	0	40	0	1
URW: Praise of Russia	0	39	0	2
URW: Russia is the Victim	0	39	0	2
URW: Speculating war outcomes	1	35	2	3

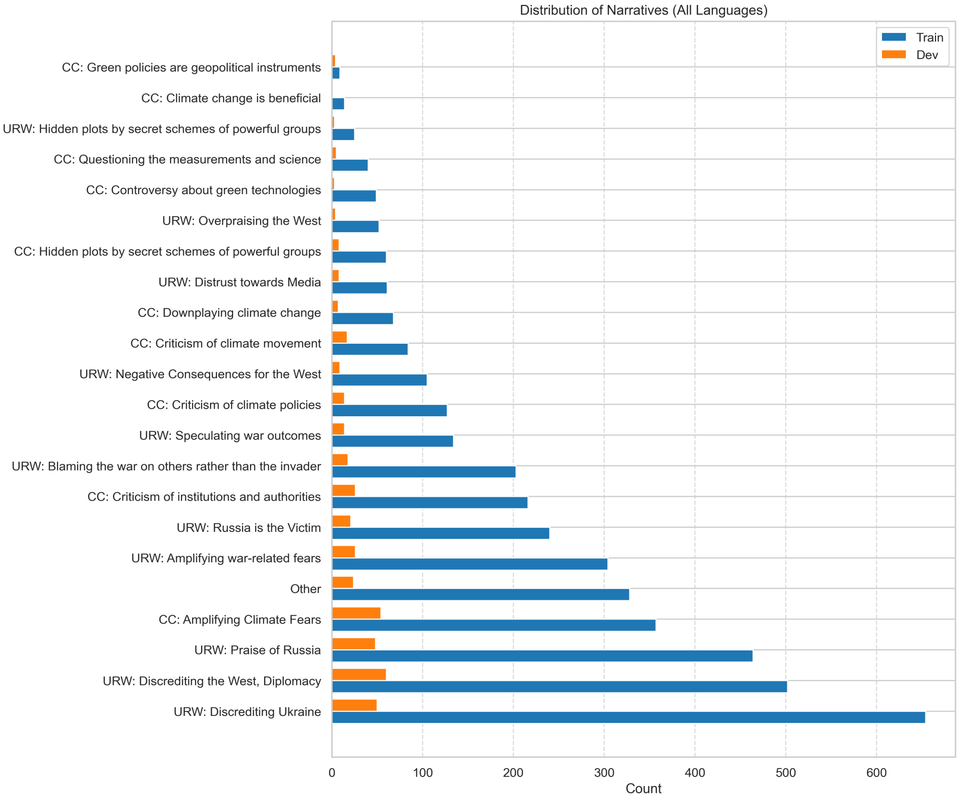

Figure 2: Narrative distribution among *train* and *dev* sets, all languages

Table 8: Confusion Matrices for Climate Change Subnarratives (DEV set)

Label	**TP**	**TN**	**FP**	**FN**
CC: Amplifying Climate Fears: Amplifying existing fears of global warming	0	40	1	0
CC: Amplifying Climate Fears: Doomsday scenarios for humans	0	39	2	0
CC: Amplifying Climate Fears: Earth will be uninhabitable soon	0	40	1	0
CC: Climate change is beneficial: CO2 is beneficial	1	40	0	0
CC: Controversy about green technologies: Other	1	38	2	0
CC: Controversy about green technologies: Renewable energy is costly	1	40	0	0
CC: Controversy about green technologies: Renewable energy is dangerous	1	39	1	0
CC: Controversy about green technologies: Renewable energy is unreliable	0	40	1	0
CC: Criticism of climate movement: Ad hominem attacks on key activists	2	37	1	1
CC: Criticism of climate movement: Climate movement is alarmist	3	36	1	1
CC: Criticism of climate movement: Climate movement is corrupt	1	37	1	2
CC: Criticism of climate movement: Other	1	34	3	3
CC: Criticism of climate policies: Climate policies are only for profit	0	37	3	1
CC: Criticism of climate policies: Climate policies have negative impact on the economy	1	40	0	0
CC: Criticism of climate policies: Other	0	39	1	1
CC: Criticism of institutions and authorities: Criticism of international entities	1	38	1	1
CC: Criticism of institutions and authorities: Criticism of national governments	1	37	1	2
CC: Criticism of institutions and authorities: Criticism of political organizations and figures	4	33	2	2
CC: Criticism of institutions and authorities: Other	0	36	4	1
CC: Downplaying climate change: Human activities do not impact climate change	0	39	0	2
CC: Downplaying climate change: Ice is not melting	0	40	1	0
CC: Downplaying climate change: Other	0	39	1	1
CC: Downplaying climate change: Weather suggests the trend is global cooling	0	40	1	0
CC: Green policies are geopolitical instruments: Climate-related international relations are abusive/exploitative	1	39	0	1
CC: Green policies are geopolitical instruments: Other	0	39	1	1
CC: Hidden plots by secret schemes of powerful groups: Blaming global elites	0	39	0	2
CC: Hidden plots by secret schemes of powerful groups: Climate agenda has hidden motives	0	40	0	1
CC: Hidden plots by secret schemes of powerful groups: Other	0	40	0	1
CC: Questioning the measurements and science: Data shows no temperature increase	0	39	1	1
CC: Questioning the measurements and science: Methodologies/metrics used are unreliable/faulty	1	38	0	2
CC: Questioning the measurements and science: Scientific community is unreliable	1	39	1	0
Other	10	24	6	1

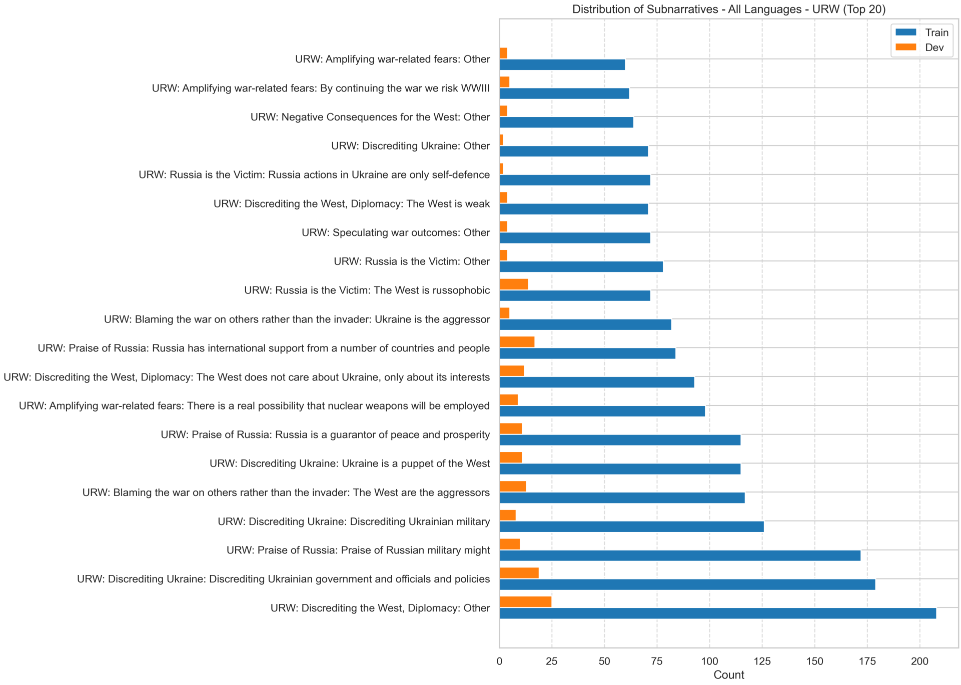

Figure 3: Subnarrative distribution among *train* and *dev* sets, all languages, Ukraine-Russia War (URW)

Table 9: Confusion Matrices for Ukraine-Russia War Subnarratives (DEV set)

Label	TP	TN	FP	FN
URW: Amplifying war-related fears: By continuing the war we risk WWIII	0	40	0	1
URW: Amplifying war-related fears: There is a real possibility that nuclear weapons will be employed	0	37	2	2
URW: Blaming the war on others rather than the invader: Other	0	40	1	0
URW: Blaming the war on others rather than the invader: The West are the aggressors	4	35	0	2
URW: Blaming the war on others rather than the invader: Ukraine is the aggressor	0	40	0	1
URW: Discrediting Ukraine: Discrediting Ukrainian government and officials and policies	3	37	1	0
URW: Discrediting Ukraine: Discrediting Ukrainian military	1	38	1	1
URW: Discrediting Ukraine: Discrediting Ukrainian nation and society	1	39	1	0
URW: Discrediting Ukraine: Other	0	38	2	1
URW: Discrediting Ukraine: Situation in Ukraine is hopeless	1	39	1	0
URW: Discrediting Ukraine: Ukraine is a hub for criminal activities	1	40	0	0
URW: Discrediting Ukraine: Ukraine is a puppet of the West	0	35	3	3
URW: Discrediting Ukraine: Ukraine is associated with nazism	2	39	0	0
URW: Discrediting the West, Diplomacy: Diplomacy does/will not work	1	38	0	2
URW: Discrediting the West, Diplomacy: Other	2	34	1	4
URW: Discrediting the West, Diplomacy: The EU is divided	1	40	0	0
URW: Discrediting the West, Diplomacy: The West does not care about Ukraine, only about its interests	2	35	2	2
URW: Discrediting the West, Diplomacy: The West is overreacting	0	39	2	0
URW: Discrediting the West, Diplomacy: The West is weak	1	40	0	0
URW: Discrediting the West, Diplomacy: West is tired of Ukraine	0	39	2	0
URW: Distrust towards Media: Western media is an instrument of propaganda	2	37	0	2
URW: Hidden plots by secret schemes of powerful groups: Other	0	40	1	0
URW: Negative Consequences for the West: Other	0	39	2	0
URW: Negative Consequences for the West: Sanctions imposed by Western countries will backfire	0	40	0	1
URW: Overpraising the West: The West belongs in the right side of history	0	40	0	1
URW: Praise of Russia: Other	0	40	0	1
URW: Praise of Russia: Praise of Russian President Vladimir Putin	0	40	0	1
URW: Praise of Russia: Praise of Russian military might	0	40	0	1
URW: Praise of Russia: Russia is a guarantor of peace and prosperity	0	40	0	1
URW: Russia is the Victim: Other	0	40	0	1
URW: Russia is the Victim: The West is russophobic	0	40	0	1
URW: Speculating war outcomes: Other	0	40	1	0
URW: Speculating war outcomes: Russian army is collapsing	0	39	0	2
URW: Speculating war outcomes: Ukrainian army is collapsing	0	38	1	2

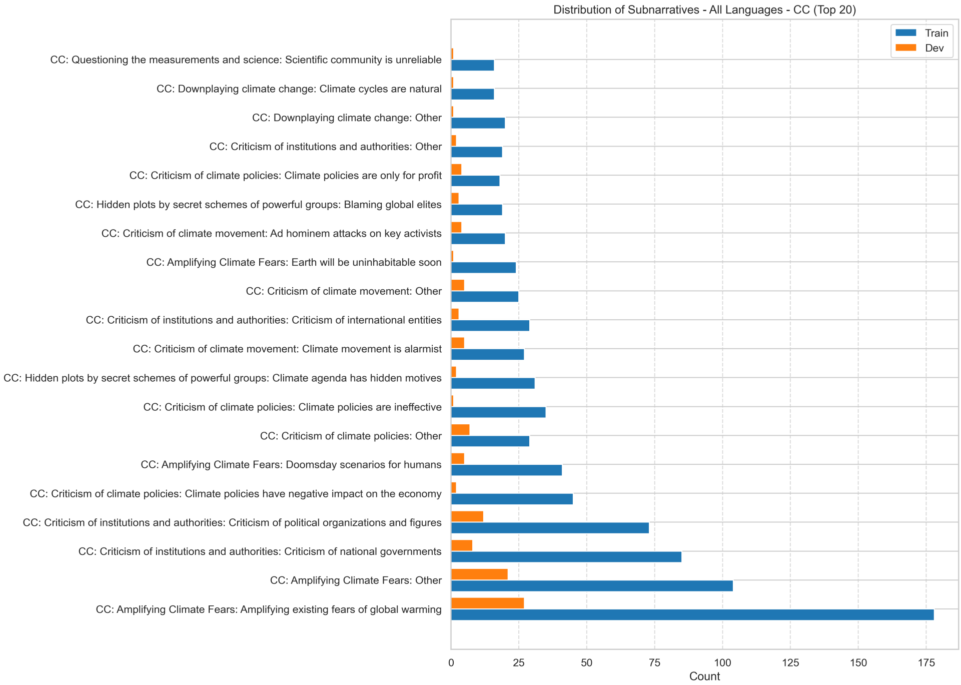

Figure 4: Subnarrative distribution among *train* and *dev* sets, all languages, Climate Change (CC)

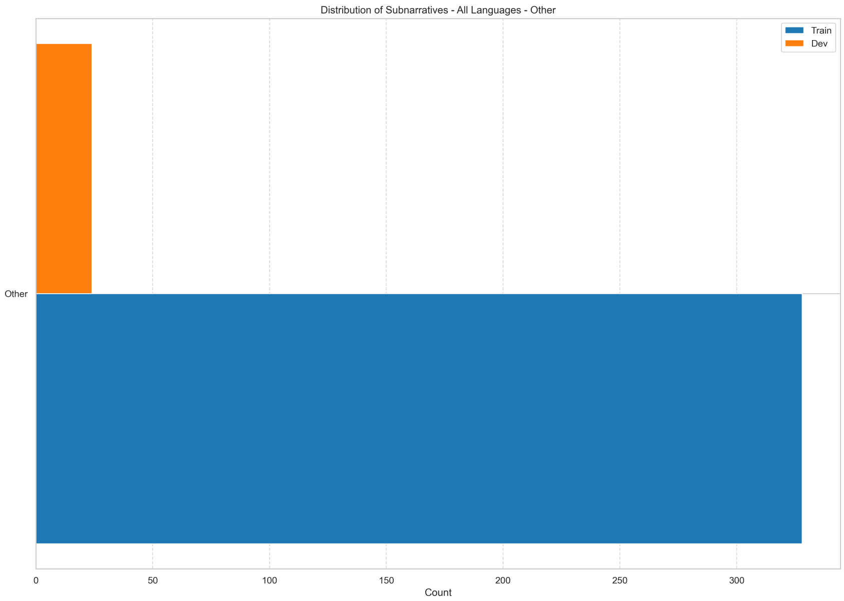

Figure 5: Subnarrative distribution among *train* and *dev* sets, all languages, Other