Narrlangen at SemEval-2025 Task 10: Comparing (mostly) simple multilingual approaches to narrative classification

Andreas Blombach[1], Bao Minh Doan Dang[1], Stephanie Evert[1], Tamara Fuchs[2], Philipp Heinrich[1], Olena Kalashnikova[2], Naveed Unjum[1]

[1]Chair of Computational Corpus Linguistics [2]Chair of Japanese Studies
Friedrich-Alexander-Universität Erlangen-Nürnberg
[1]Bismarckstr. 6, 91054 Erlangen [2]Artilleriestr. 70, 91052 Erlangen
{firstname.lastname}@fau.de

Abstract

In SemEval 2025 Task 10, which addresses the multilingual characterisation and extraction of narratives from online news, our team *Narrlangen* focused on Subtask 2 (narrative classification), and we tried several conceptually straightforward approaches: (1) prompt engineering of LLMs, (2) a zero-shot approach based on sentence similarities, (3) direct classification of fine-grained labels using SetFit, (4) fine-tuning encoder models on fine-grained labels, and (5) hierarchical classification using encoder models with two different classification heads. We list results for all systems on the development set, which show that the best approach was to fine-tune a pre-trained multilingual model, XLM-RoBERTa, with two additional linear layers and a softmax as classification head.

1 Introduction & background

Narratives shape how information is conveyed and understood, influencing public discourse and decision-making processes. Since both corporate and state actors are actively seeking to influence public discourse on topics such as climate change, vaccinations, migration, or the war in Ukraine by pushing their own narratives, one of the greatest challenges of contemporary democracies is to identify and counteract such campaigns while upholding the ideal of freedom of expression. Identifying narratives in texts, e.g. online news or social media, is a key part of this. More generally, robust methods for identifying and classifying narratives would also provide important support for large-scale analysis of textual data, allowing researchers to track the evolution of discourses across time.

Despite significant advancements in NLP, predicting narratives remains a complex challenge due to their abstract, multi-layered nature. Traditional classification methods struggle with the implicit and evolving structures of narratives, which often span multiple sentences or paragraphs. Recent approaches, including zero-shot learning and fine-tuning of transformer models, have demonstrated promise in capturing nuanced narrative patterns without requiring extensive labelled datasets (see e.g. Heinrich et al., 2024).

The task is additionally complicated by terminological uncertainty. Over the years, scholars have proposed various interpretations for the very term *narrative* itself, reflecting the difficulty in reaching a consensus (see e.g. Santana et al., 2023). Chatman (1980) e.g. offers a structuralist perspective, defining narratives as comprising a story (a chain of events and characters) and discourse (how the content is communicated).[1] Riedl and Young (2010) see narratives and storytelling as cognitive tools for making sense of the world. Broader definitions highlight that narratives are sequences of events that form a cohesive whole, with significance derived from the relationship between events. Consequently, detailed *annotation* of narratives usually comprises key features such as participants, events, and time (Silvano et al., 2021).

In Subtask 2 of SemEval-2025 Task 10 (Piskorski et al., 2025) narrative annotation is provided as coarse- and fine-grained labels given to whole news articles. The provided data covers news in five different languages, namely English, Portuguese, Bulgarian, Russian, and Hindi, and the task is to automatically annotate texts with all (sub-)narratives in a multi-label fashion. The narratives belong to two macro topics: climate change and the War in Ukraine, both prime domains for fake news and disinformation intended to mislead the public. In our contribution, we compare a variety of state-of-the-art approaches; all our code is publicly available.[2]

[1] However, it is needless to say that also the term *discourse* is highly problematic, since "it is used in social and linguistic research in a number of inter-related yet different ways." (Baker, 2006, 3)

[2] See `https://github.com/fau-klue/narrlangen-semeval2025`.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2240–2248
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

2 Data & labels

2.1 Data sets

The task organisers provided training, development, and test data for English, Bulgarian, Portuguese, Russian, and Hindi, respectively. Table 2 in the Appendix shows how many news texts there are in each set and gives an impression of their typical lengths.

Training and development data sets are annotated with ten coarse-grained labels for narratives related to climate change, eleven for the Russo-Ukrainian War, as well as the label "Other". Fine-grained labels further subdivide these narratives (for example, the coarse-grained label "Downplaying climate change" allows for subnarratives like "Humans and nature will adapt to the changes" or "Climate cycles are natural"). In total, there are 96 possible fine-grained levels, including "Other" subnarratives for each coarse-grained label which are not explicitly included in the taxonomy (Stefanovitch et al., 2025). Frequencies of fine-grained labels vary wildly between languages and, in some cases, even between training and development data.[3] Each data set only contains a subset of coarse- and fine-grained labels, the most extreme case being Russian, where no narratives relating to climate change are to be found.

2.2 Narrative descriptions

We manually created paragraph-length descriptions of subnarratives to be used in the sentence-similarity based zero-shot approach outlined below.[4] By using the provided taxonomy (Stefanovitch et al., 2025), we matched each subnarrative with narrative descriptions in the form of English example sentences. The sentences are intended to concisely describe the subnarratives and should contain all discourse-relevant terms. The provided examples within the taxonomy were used as a basis; further aspects were added based on domain knowledge, academic publications, online searches, and fact-checking websites[5].

Existing research on climate change skepticism and conspiracy theories (as outlined e.g. in Tam and Chan, 2023) as well as a vast array of material available online provided the necessary information for the creation of descriptive sentences for all subnarratives related to climate change. For instance, for the subnarrative "Climate policies are ineffective", our narrative description reads as follows: "Ineffective climate policies have done more harm than CO2 emissions. Even if we reduce our CO2 emissions, it won't save the planet."

Pro-Russian narratives related to the Russo-Ukrainian war are very frequent and well outlined in the academic literature (Aleksejeva, 2023; Amanatullah et al., 2023; Kalashnikova and Schäfer, 2024; Pekar and Rashkovan, 2024), making the creation of descriptive sentences straightforward. On the other hand, pro-Western narratives are less studied and less common. Descriptive sentences here are thus derived from discussions on social media platforms (e.g. "The West belongs in the right side of history"). Note that many subnarratives are interconnected, which makes it difficult to assign description sentences to one particular subnarrative; the subnarratives "Ukraine is the aggressor" and "Ukraine is a puppet of the West" e.g. both assume "the West" as an intermediate, and both "Ukraine is associated with nazism" and "Russia actions in Ukraine are only self-defence" involve Russia's allegations of "genocide" committed by Ukraine as the justification for the invasion.

3 System overview

Machine learning (ML) baseline We initially ran simple multi-label classification experiments with classical machine learning algorithms for all languages in order to get stronger baselines than the ones based on random guessing provided by the organisers. We use logistic regression (LR) and support vector machines (SVM) on bags of words for this purpose. The baselines are computed separately for each language.

Prompt engineering LLMs (PromptEng) We also implemented a structured approach of prompt engineering large language models (LLMs), namely GPT-4o (OpenAI, 2024) and Deepseek R1-32B (Guo et al., 2025) in a step-by-step fashion. The LLMs were tasked to proceed level by level and output the response in json format.

Similarity-based zero-shot (SentSim) A simple approach to scoring texts can be constructed by looking at similarities between sentences in texts and subnarrative descriptions. If there is at least one

[3] For example, the label "Other" appears in 98 of 366 texts in the Hindi training set (26.8%), but only twice in the development set of 35 texts (5.7%).

[4] Our description sentences can be found alongside our code in the repository linked above.

[5] See http://euvsdisinfo.eu/ and https://www.weareukraine.info/.

sentence in a text which is very similar to a sentence of a subnarrative description, the text will likely contain this subnarrative. This approach can have decent zero-shot performance as we have shown in the context of detecting COVID-19 related conspiracy narratives in German Telegram posts (Heinrich et al., 2024).

Let S_d denote the sentences of text d and S_n the sentences (or paragraphs) of subnarrative n. Following Heinrich et al. (2024), we use cosine similarity between sentence embeddings $\mathbf{e}_{s_i}$ and $\mathbf{e}_{s_j}$ for scoring each pair of sentences

$$s(s_i, s_j) = \cos\left(\mathbf{e}_{s_i}, \mathbf{e}_{s_j}\right) \quad \forall s_i \in S_d, s_j \in S_n.$$

Note that individual sentences of subnarrative descriptions capture different ways of expressing the subnarrative, and since we are chiefly interested in the overall presence of subnarratives, we aggregate via taking the maximum

$$\text{score}\,(d, n) = \max_{s_i \in S_d, s_j \in S_n}\left(s(s_i, s_j)\right)$$

for getting a single value for each pair of subnarrative and text. This score is then translated into an actual prediction by determining a language- and subnarrative-specific cut-off value that maximises F_1 on all available labelled data.[6] Note that this last step of maximising the desired evaluation metric technically changes the approach to a supervised algorithm.

SetFit SetFit (Tunstall et al., 2022), a framework for few-shot fine-tuning Sentence Transformers (Reimers and Gurevych, 2019), is another approach that achieved good results in detecting COVID-19-related conspiracy narratives (cf. Heinrich et al., 2024). Coupled with the fact that it does not require prompting and is relatively fast to train, this makes it a sensible choice to use it for the task at hand. SetFit works by (1) fine-tuning a pre-trained Sentence Transformers model on contrastive pairs of labelled texts, (2) using the resulting model to encode the training data, (3) using the encoded data to train a text classification head.

Fine-tuning on fine-grained labels (FGM) In this approach, we fine-tune a multilingual masked language model and introduce two linear layers with a ReLU activation in between, followed by a final softmax function to predict fine-grained labels. This model prioritises subnarrative classification accuracy, which is considered most crucial for competitive performance. Coarse narrative labels are then inferred from the subnarratives. We trained the model on the combined data from all five languages in order to obtain a unified model that is applicable across all languages involved in the competition.

Hierarchical models We further conduct experiments with a model architecture that takes the nature of the labels into account, i.e. the relationship between narratives and subnarratives. In the course of these experiments, we fine-tune a multilingual masked LLM with two additional classification heads. The models are trained in two different manners, namely multi-label and multi-class with narratives attention. The hierarchy in the model is constructed by using the embeddings generated by the masked LLM to predict the coarse-grained labels, i.e. the narratives, and subsequently use the outputs to extract additional features to classify subnarratives.

The **multi-label model (MLHM)** takes the raw logits from the narrative level and feeds it to a sigmoid function in order to derive label probabilities, which are then utilised as additional features for the subnarratives level head. This method aims to establish a dependency between different hierarchical levels, thereby strengthening the interrelationship during training. We therefore assume that the subnarratives head is capable of optimising the probabilities for particular labels, leveraging the predictions from the narratives head.

The **narratives attention based model (NAHM)** employs separate attention mechanisms, one for each classification level. Each attention mechanism independently attends to the hidden states of the masked LLM to extract level-specific features from the text. As in MLHM, hierarchical relationships are established by feeding the output probabilities from higher levels as input features to lower levels. This architecture allows information to flow from broader categories to more specific ones. The model maintains a dictionary of label mappings which is crucial for both training consistency and interpreting predictions during inference.

In order to maintain consistent predictions throughout all hierarchical levels, a batch consistency loss function $\mathcal{L}_c$ is implemented as the learning objective. Let ℓ_m denote a loss function accord-

[6] This corresponds to the training data for predictions on the development set and to the combined training and development set for predictions on the test set.

ing to the model architecture[7] and let $i \in [1, N]$ be the current level (where N is the total number of all hierarchical levels). The consistency loss $\mathcal{L}_c$ is then defined as

$$\mathcal{L}_c = \sum_{i}^{N} \ell_m(p_\theta(\hat{y}_i|x_i), y_i)$$

where the x_i populate the feature matrix, y_i are the true labels of i-th hierarchical level, $p_\theta(\hat{y}_i|x_i)$ are the predicted probabilities, and θ are the trainable parameters of the model. For the attention based model, the mean of $\mathcal{L}_c$ is computed for each batch using

$$\bar{\mathcal{L}_c} = \mathcal{L}_c \frac{1}{|\text{batch}|}$$

as final loss that needs to be minimised during training.

4 Experimental setup

All our models make use of the combined training data during training and are evaluated on the development data. Note that since most of our models were trained to predict fine-grained narratives, we inferred the corresponding coarse-grained narratives from the fine-grained labels using regular expressions. Where applicable, we used multilingual models, which enables us to combine the full training data for all five languages to train a single model, thereby addressing the problem of data scarcity in fine-grained labels.[8]

ML We implement logistic-regression and support vector machine with a tf.idf-weighted feature matrix based on uni- and bigrams using `scikit-learn` (Pedregosa et al., 2011).

PromptEng For prompt engineering, the LLMs are provided with the multi-level labels and prompted to choose a narrative first and only choose the subnarrative within that particular narrative second. We used a few-shot learning framework within this methodological sequence by providing some examples in English.

SentSim For the zero-shot experiments, we split texts by double new lines into paragraphs and treat each paragraph as a sentence; for subnarrative description, we experiment with two approaches: using subnarrative descriptions as a whole, and splitting them into smaller segments (which mostly correspond to single sentences, but can also comprise two or three sentences). We then calculate sentence embeddings using four different multilingual SBERT (Reimers and Gurevych, 2019) models available off-the-shelf[9] from Hugging Face, namely

- paraphrase-multilingual-MiniLM-L12-v2
- paraphrase-xlm-r-multilingual-v1
- distiluse-base-multilingual-cased-v1
- paraphrase-multilingual-mpnet-base-v2[10]

We refer to these models as mini, XLM, distiluse, and mpnet, respectively.

SetFit For SetFit, we seek to find the most suitable parameters to fine-tune the model on the training dataset by using a batch size of 16, setting the learning rate $lr \in \{1\text{e-}4, 1\text{e-}6\}$ and the number of epochs $e \in \{1, 3\}$. The best model after hyperparameter tuning is subsequently evaluated using the development set. We use the mpnet model as above.

FGM The model is trained on a combined dataset of all the languages, while taking only the subnarrative labels into account. The subnarrative labels are one-hot encoded, thus enabling easier training and prediction, as well as maintaining the multi-label status. The narratives are subsequently extracted from the subnarratives. We use the base versions of both English-specific BERT[11] (Devlin et al., 2019), and multilingual XLM-RoBERTa[12], and fine-tune them for 100 epochs[13]. We apply the AdamW optimiser (Loshchilov and Hutter, 2017) configured at a learning rate of 2e-5. The learning objective is to minimise the binary Cross Entropy Loss.

Hierarchical models As the results of XLM-RoBERTa and English-specific BERT were similar

[7] This corresponds to binary cross entropy for MLHM and negative log-likelihood for NAHM

[8] Although this introduced another, albeit smaller problem, in that some labels from the taxonomy do not appear in the training and/or development sets of individual languages at all.

[9] We use the Python module `SentenceTransformers` here, see `https://www.sbert.net/`.

[10] `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`

[11] `https://huggingface.co/google-bert/bert-base-uncased`

[12] `https://huggingface.co/FacebookAI/xlm-roberta-base`

[13] During our initial experiments, we set the epoch sizes to 10 and 20 for fine-tuning FGM but did not acquire any sensible insights. The model was just able to predict the category "Other" and failed to learn the other ones. A possible reason for this issue is the scarcity of training examples for all categories.

in FGM, we decided to utilise XLM-RoBERTa as encoder model for these experiments. We adopt the AdamW optimiser with a learning rate of 3e-5, a batch size of 16 for both hierarchical models and train them for 120 (MLHM) as well as 100 (NAHM) epochs. Both models are subsequently evaluated on the provided development set.[14]

5 Results

We list results on the development set for all our systems in Table 4 in the Appendix, and visualise them in Figure 1 in comparison to other results published on the official task leaderboard[15]. Results for out submitted system (FGM) on the test set can be found in Table 1. We ranked between 3^{rd} and 6^{th} place across the five languages of the shared task, achieving 3^{rd} place in Hindi (out of 13 participating teams), 4^{th} in Russian (14 teams), 5^{th} in Bulgarian (11 teams), and 6^{th} in both English (28 teams) and Portuguese (13 teams).

ML baseline As can be seen in Figure 1, the ML baselines fail to adequately predict subnarratives, with F_1 ranging from 0.01 to 0.07 on the development set, thus barely beating random guessing.

PromptEng The strategy of prompt-engineering GPT-4o was considerably worse than our other approaches when we evaluated it on the English development set, especially looking at F_1 scores on fine-grained labels. The results from Deepseek-R1:32B were even worse, and thus this approach was not considered any further.

SentSim For the approach based on sentence similarities, we compare a total of eight architectures (segmenting subnarratives in sentences or paragraphs, and four different language models). We assess their performance for each language on the development set in Table 3 in the Appendix. We quickly summarise the table as follows, focussing on the prediction of subnarratives: (1) We easily beat the baselines across all languages; (2) comparing paragraph embeddings with paragraph embeddings yields almost exclusively better results than comparing sentence embeddings with paragraph embeddings (except for XLM on Portuguese); and (3) the mpnet model performs best across all languages (except for English, where it performs remarkably poorly).[16] We include the architecture with the mpnet model and comparing paragraphs with paragraphs in Figure 1. We do not reach median performance of participating systems with this approach (except for Russian) but note that this approach does not need any labelled data (except for maximising F_1).

SetFit We report performance after training and hyperparameter optimisation[17] in Table 4 and Figure 1. Given that SetFit was used with whole labelled news texts instead of much shorter units of text, it still achieves decent results. In principle, SetFit should also work well in a zero-shot setting, using only narrative descriptions as its initial input. In our case, however, the resulting model was essentially useless. Contrary to Heinrich et al. (2024), using narrative descriptions as additional training data did not improve the model.

FGM Fine-tuning XLM-RoBERTa only on the finegrained labels has been shown to be an effective approach and the best model among all our experiments. This architecture is only outperformed by the MLHM approach for Portuguese. To be consistent, we submitted FGM predictions for all languages on the test set for final evaluation. Table 1 and Figure 1 show performance on the test set; results are overall competitive, especially for Hindi, enabling a top 3 placement. To check whether a single-language model would perform better, we also tested an English-only BERT model as the base model. The results, however, did not show an improvement over XLM-RoBERTa.

Error analysis on the development sets across task languages reveals several things.[18] First, good performance is mostly based on accurately predicting the absence of subnarratives. Second, performance would have slightly increased if predictions of labels not included in the respective language trainings sets had been removed in a post-processing step (for English, for example, the model predicts one instance of "Amplifying Climate Fears: Earth will be uninhabitable soon" while this subnarrative features neither in the English training nor in the development set). Third,

[14] We ran all our experiments with encoder models on an NVIDIA RTX 4090 with 24 GB of VRAM, with an average run time of approx. 40 minutes to complete each training.

[15] https://propaganda.math.unipd.it/semeval2025task10/leaderboard.php

[16] Note that distiluse was trained neither on Bulgarian nor on Hindi, which explains its poor performance on these languages.

[17] Hyperparameter optimisation was, however, largely inconsequential compared to out-of-the-box performance.

[18] Confusion matrices for all labels are available at https://github.com/fau-klue/narrlangen-semeval2025.

	coarse		fine	
	F_1	σ	F_1	σ
EN	0.44	0.41	0.34	0.39
BG	0.50	0.40	0.36	0.38
PT	0.48	0.34	0.29	0.26
RU	0.57	0.37	0.41	0.32
HI	0.40	0.46	0.39	0.47

Table 1: Results on test set using XLM-RoBERTa on fine-grained labels (FGM).

prediction quality for some of the underspecified "Other" subnarratives for coarse-grained labels varies wildly between languages: for "Discrediting the West, Diplomacy: Other", the model accurately predicts 4 out of 5 instances in the Bulgarian dev set whereas it only gets 1 out of 6 right for English. This likely results from large differences between instances in the training data (in this case, 61 vs. 26). The same holds true for the even more general label "Other": there are more training examples for English (169) than for all other languages combined (159), so that the model only achieves a somewhat decent performance in this language (precision .62, recall .73). Finally, results for Russian and Hindi are only better than those for other languages since they only (or mostly, in the case of Hindi) include subnarratives related to the Russo-Ukrainian War. m

Hierarchical models The results of our multi-label hierarchical model (MLHM) outperforms the attention-based approach (NAHM) across all languages for both coarse- and fine-grained labels. Especially for Portuguese, we can observe a gain of 0.21 in F_1, where MLHM also surpasses the otherwise best-performing system FGM by 0.03 in F_1. This might be due to the complexity of the attention layer and the fact that it is trained to predict one single label compared to the multi-label approach, which is not constrained to a certain narrative or subnarrative.

6 Discussion & perspectives

The results of our experiments suggest that traditional machine learning approaches provide little utility in the given task, as their performance remains significantly below that of more advanced methods. In contrast, our sentence-similarity based zero-shot approach proves to be a viable alternative, delivering competitive results without the need for domain-specific training. Nonetheless, while the zero-shot method offers a cheap alternative, fine-tuning masked LLMs remains the most effective approach, given a large amount of labelled data as in the task here. We also note that incorporating multiple languages into a single training set is a reasonable strategy, likely due to shared linguistic patterns across languages.

For future work, several directions could enhance classification performance. Firstly, since many annotated texts contain lengthy narratives including irrelevant sections, preprocessing techniques that extract relevant portions before model training could reduce noise and improve classification accuracy for all models. Similarly, SetFit could benefit from individually labelled paragraphs; it might also perform better as an ensemble of several models to predict fine-grained labels for each previously predicted coarse-grained label. For SentSim, an iterative refinement process for narrative descriptions could enhance model interpretability and predictive accuracy by incrementally improving label definitions and annotations.

While prompt engineering LLMs was not a successful strategy here, chain-of-thought prompting could improve the results (Wei et al., 2023). Findings from previous work (Jiang et al., 2024; Qi et al., 2024; He et al., 2024) suggest that LLMs have difficulty following complex instructions. Given that our task requires multi-label classification at various levels and a strict json output, this approach might not be sufficient to address the task's complexity.

Applying the narrative attention layer to the multi-label hierarchical model might strengthen the relationship between coarse- and fine-grained hierarchies. This enhancement can improve the model's ability to predict subnarratives which correspond to the classified narratives. Another possibility would be to experiment with the hierarchical loss by minimising the subnarrative loss only. We assume that focusing on subnarratives might improve the performance of our approaches.

Finally, after a thorough error analysis of the individual models presented here, it would be straightforward to use an ensemble model which could potentially improve overall performance.

Acknowledgements

This research has been partially funded by the German Research Foundation (DFG), project no. 466328567.

References

Nika Aleksejeva. 2023. Narrative warfare. How the Kremlin and Russian News Outlets Justified a War of Aggression against Ukraine. Technical report, The Atlantic Council (Digital Forensic Research Lab), Washington, D.C.

Samy Amanatullah, Serena Balani, Angela Fraioli, Stephanie M. McVicker, and Mike Gordon. 2023. Tell Us How You Really Feel: Analyzing Pro-Kremlin Propaganda Devices Narratives to Identify Sentiment Implications. Technical report, GW's Institute for European, Russian and Eurasian Studies.

Paul Baker. 2006. *Using Corpora in Discourse Analysis*. Continuum, London.

Seymour Chatman. 1980. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press, Ithaca.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Qianyu He, Jie Zeng, Qianxi He, Jiaqing Liang, and Yanghua Xiao. 2024. From complex to simple: Enhancing multi-constraint complex instruction following ability of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10864–10882, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Heinrich, Andreas Blombach, Bao Minh Doan Dang, Leonardo Zilio, Linda Havenstein, Nathan Dykes, Stephanie Evert, and Fabian Schäfer. 2024. Automatic Identification of COVID-19-Related Conspiracy Narratives in German Telegram Channels and Chats. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 1932–1943, Torino, Italy.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. Follow-Bench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand. Association for Computational Linguistics.

Olena Kalashnikova and Fabian Schäfer. 2024. Russian State-controlled Propaganda and its Proxies: Pro-Russian Political Actors in Japan. *The Asia-Pacific Journal: Japan Focus*, 33(3).

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

OpenAI. 2024. GPT-4o System Card.

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. *Journal of Machine Learning Research*, 12(85):2825–2830.

Valerii Pekar and Vladyslav Rashkovan. 2024. Seven favourite hidden narratives of Russian propaganda.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Jorge Alípio, 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.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2024. Constraint back-translation improves complex instruction following of large language models. *Preprint*, arXiv:2410.24175.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mark O. Riedl and R. Michael Young. 2010. Narrative planning: balancing plot and character. *J. Artif. Int. Res.*, 39(1):217–268.

Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artificial Intelligence Review*, 56(9):8393–8435.

Purificação Silvano, António Leal, Fátima Silva, Inês Cantante, Fatima Oliveira, and Alípio Mario Jorge. 2021. Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus. In *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 1–13, Groningen, The Netherlands (online). Association for Computational Linguistics.

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).

Kim-Pong Tam and Hoi-Wing Chan. 2023. Conspiracy theories and climate change: A systematic review. *Journal of Environmental Psychology*, 91:102129.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

A Data sets

			text length		
	lang.	#texts	mean	median	σ
train	EN	399	2984.2	2960	858.1
	PT	400	2459.9	2420	566.2
	BG	401	2341.1	2001	1037.1
	RU	215	2035.8	2422	931.0
	HI	366	2367.5	1717	2297.6
dev	EN	41	3562.3	3102	1753.3
	PT	35	2486.5	2562	511.1
	BG	35	2026.7	1628	981.9
	RU	32	1791.9	1470	850.3
	HI	35	3283.6	2681	1545.6
test	EN	101	3791.9	3442	1533.3
	PT	100	2573.3	2545	665.0
	BG	100	3669.6	3621	1003.3
	RU	60	2801.8	2748	485.3
	HI	99	1447.1	1320	615.1

Table 2: Data set overview for all sets and languages in terms of number of texts (#texts) and number of characters (columns text length).

B Results for sentence similarity architectures

			coarse		fine	
lang	model	segm.	F_1	σ	F_1	σ
EN	distiluse	p	0.41	0.33	**0.26**	0.28
		s	0.36	0.32	0.17	0.20
	mpnet	p	0.32	0.32	0.16	0.24
		s	0.32	0.30	0.13	0.19
	mini	p	**0.43**	0.33	0.24	0.31
		s	0.37	0.32	0.19	0.25
	xlm	p	0.35	0.33	0.22	0.27
		s	0.34	0.35	0.18	0.26
PT	distiluse	p	0.31	0.26	0.15	0.19
		s	0.33	0.26	0.16	0.21
	mpnet	p	**0.42**	0.32	**0.19**	0.22
		s	0.31	0.25	0.16	0.18
	mini	p	0.31	0.23	0.17	0.20
		s	0.34	0.21	0.16	0.16
	xlm	p	0.38	0.32	0.14	0.19
		s	0.40	0.28	**0.19**	0.23
BG	distiluse	p	0.19	0.17	0.08	0.10
		s	0.21	0.16	0.11	0.13
	mpnet	p	**0.28**	0.22	**0.14**	0.18
		s	0.27	0.22	**0.14**	0.19
	mini	p	0.21	0.20	0.10	0.17
		s	0.23	0.20	0.11	0.15
	xlm	p	0.25	0.21	0.11	0.15
		s	0.26	0.22	0.12	0.17
RU	distiluse	p	0.50	0.32	0.21	0.22
		s	**0.51**	0.28	0.20	0.23
	mpnet	p	0.46	0.33	**0.28**	0.29
		s	0.44	0.32	0.27	0.28
	mini	p	0.44	0.29	0.21	0.25
		s	0.49	0.29	0.24	0.24
	xlm	p	0.45	0.32	0.25	0.26
		s	0.48	0.30	0.26	0.29
HI	distiluse	p	0.18	0.18	0.07	0.10
		s	0.20	0.17	0.09	0.11
	mpnet	p	0.25	0.25	**0.18**	0.23
		s	**0.32**	0.28	0.18	0.22
	mini	p	0.25	0.19	0.15	0.13
		s	0.28	0.22	0.15	0.17
	xlm	p	0.26	0.21	0.13	0.15
		s	0.23	0.19	0.12	0.14

Table 3: Results for the sentence similarity approach on the development set using different system architectures. The mpnet model scores best across all languages except English. Comparing paragraphs with paragraphs usually outperforms comparing paragraphs with sentences.

C Comparison of results on development and test set

		coarse		fine	
		F_1	σ	F_1	σ
LR	EN	0.27	0.44	0.04	0.12
	BG	0.17	0.37	0.02	0.08
	PT	0.03	0.17	0.07	0.15
	RU	0.13	0.33	0.01	0.05
	HI	0.06	0.23	0.01	0.05
SVM	EN	0.27	0.44	0.04	0.12
	BG	0.17	0.38	0.01	0.07
	PT	0.03	0.17	0.04	0.13
	RU	0.13	0.33	0.04	0.12
	HI	0.06	0.23	0.01	0.07
SetFit	EN	0.39	0.42	0.32	0.40
	BG	0.39	0.46	0.36	0.44
	PT	0.30	0.43	0.19	0.34
	RU	0.24	0.40	0.23	0.39
	HI	0.29	0.43	0.22	0.36
NAHM	EN	0.45	0.36	0.28	0.33
	BG	0.41	0.42	0.27	0.35
	PT	0.44	0.39	0.25	0.35
	RU	0.49	0.37	0.21	0.30
	HI	0.41	0.40	0.28	0.34
FGM	EN	0.40	0.40	**0.35**	0.39
	BG	**0.51**	0.42	**0.42**	0.40
	PT	0.62	0.35	0.43	0.34
	RU	0.54	0.36	**0.35**	0.34
	HI	0.45	0.39	**0.31**	0.37
MLHM	EN	**0.49**	0.41	0.32	0.39
	BG	0.49	0.41	0.35	0.37
	PT	**0.69**	0.28	**0.46**	0.35
	RU	**0.59**	0.30	0.29	0.31
	HI	**0.54**	0.40	0.28	0.36
PromptEng	EN	0.23	0.32	0.16	0.28
SentSim	EN	0.32	0.30	0.13	0.19
	BG	0.27	0.22	0.14	0.19
	PT	0.31	0.25	0.16	0.18
	RU	0.44	0.32	0.27	0.28
	HI	0.32	0.28	0.18	0.22

Table 4: Results of all our systems on the development set. We indicate the language-specific top-score in bold. Two LLM-based models (FGM and MLHM) score best across all languages.

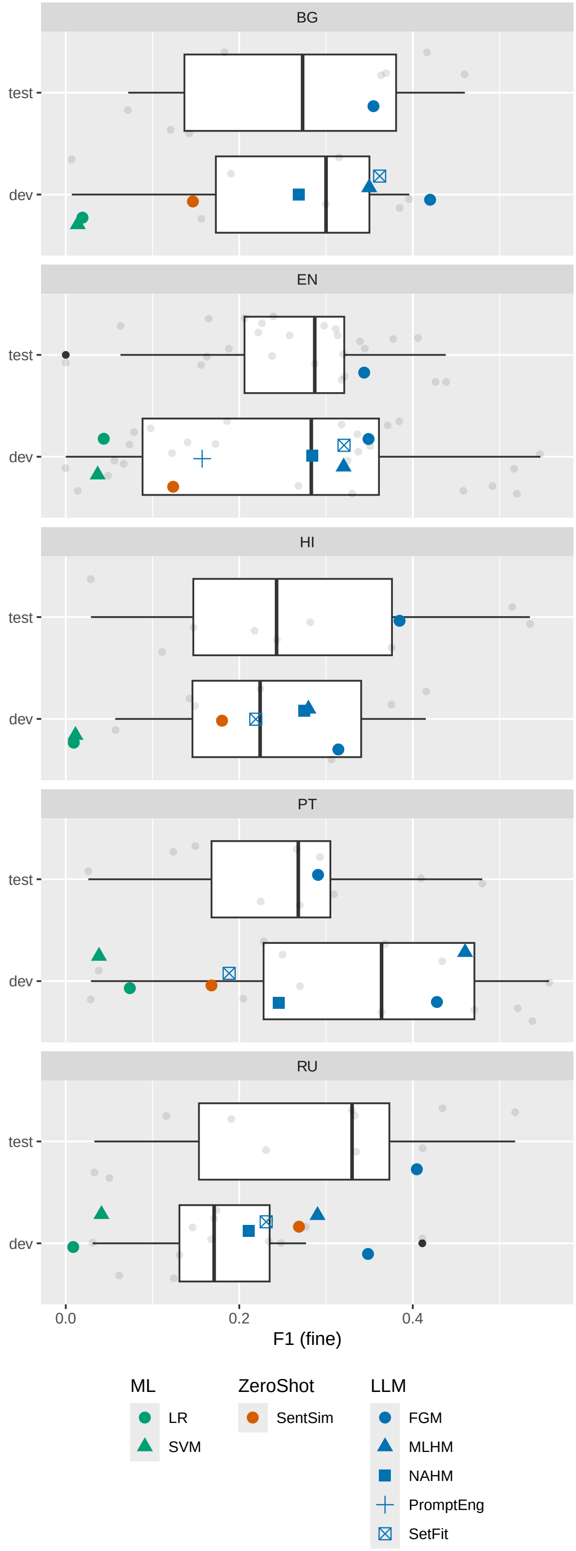

Figure 1: Results of our systems (indicated in colour) compared to other results published on the official leaderboard (visualised as boxplots) for both development and test set, split by language.