iLostTheCode at SemEval-2025 Task 10: Bottom-up Multilevel Classification of Narrative Taxonomies

Lorenzo Vittorio Concas and **Manuela Sanguinetti** and **Maurizio Atzori**
Department of Mathematics and Computer Science, University of Cagliari
Via Ospedale 72, Cagliari (Italy)
l.concas16@studenti.unica.it, {manuela.sanguinetti,atzori}@unica.it

Abstract

This paper describes the approach used to address the task of narrative classification, which has been proposed as a subtask of Task 10 on Multilingual Characterization and Extraction of Narratives from Online News at the SemEval 2025 campaign. The task consists precisely in assigning all relevant sub-narrative labels from a two-level taxonomy to a given news article in multiple languages (i.e., Bulgarian, English, Hindi, Portuguese and Russian). This involves performing both multi-label and multi-class classification. The model developed for this purpose uses multiple pretrained BERT-based models to create contextualized embeddings that are concatenated and then fed into a simple neural network to compute classification probabilities. Results on the official test set, evaluated using samples F_1, range from 0.15 in Hindi (rank #9) to 0.41 in Russian (rank #3). Besides an overview of the system and the results obtained in the task, the paper also includes some additional experiments carried out after the evaluation phase along with a brief discussion of the observed errors.

1 Introduction

Online news is a primary source of information and has a major role in shaping public discourse and influencing perceptions. Identifying the narratives embedded within news articles is crucial for critically analyzing their perspectives, biases, and underlying messages. For instance, this can be relevant in contexts where harmful or misleading content is present: recognizing the dominant narratives can facilitate the construction of counter-narratives (Tekiroğlu et al., 2020), in view of promoting a more constructive and less toxic online debate. Furthermore, given the abundance of information available online—from both mainstream and alternative media sources—understanding the stance underlying different narratives can be useful when navigating digital content. This requires not only recognizing the explicit claims made in a given article or social media post, but also understanding how such claims align with broader thematic views. A critical approach to narratives can help the interested reader distinguish between different perspectives and engage with news content in a more informed way. From a theoretical perspective, providing a shared framework for the definition and categorization of narratives and sub-narratives is essential (Stefanovitch et al., 2025) for more systematic analyses and comparisons across different texts and media sources. In turn, automatic systems can build upon these theoretical foundations to detect and classify narratives with the help of Natural Language Processing techniques (Santana et al., 2023).

Our work lays on these premises and it focuses on the task of Narrative Classification, proposed as part of SemEval-2025 Task 10 on Multilingual Characterization and Extraction of Narratives from Online News (Piskorski et al., 2025). More in particular, the team participated in Subtask 2 on Narrative Classification, which consists precisely in assigning one or more sub-narrative labels to a given news article in one among five languages (Bulgarian, English, European Portuguese, Hindi and Russian), and resorting to predefined narrative taxonomies. The news articles are centered around two main topics, Ukraine-Russia war and climate change, and each topic has its own taxonomy of narratives and corresponding sub-narratives.[1]

This challenge aligns with prior research in text classification, particularly in multi-label and hierarchical classification tasks, such as Task 4 from SemEval 2024, which shares similarities (Dimitrov et al., 2024).

To address this task, our system combines a sim-

[1]An overview of both taxonomies with annotation guidelines has been made available here: https://propaganda.math.unipd.it/semeval2025task10/NARRATIVE-TAXONOMIES.pdf

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 607–616

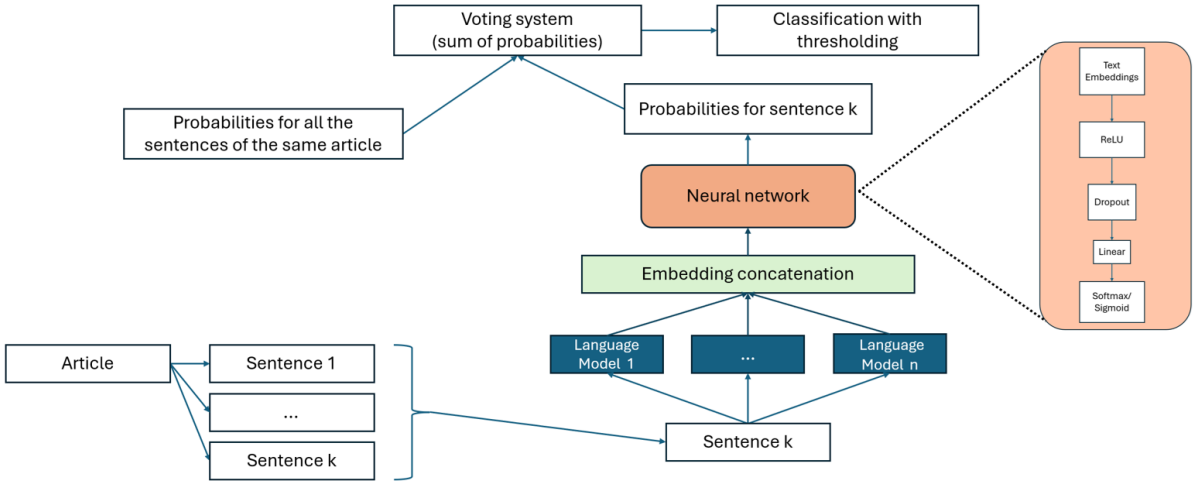

Figure 1: Model architecture.

ple neural network with a concatenation of contextual embeddings generated using multiple BERT-based models, each already fine-tuned for different tasks, though not specifically for this one. Language models such as BERT and its variations have demonstrated strong performance in various NLP tasks, using contextual embeddings to have good classification accuracy even without fine-tuning (Uppaal et al., 2023). Furthermore, approaches resorting to vector concatenation to address similar tasks were implemented in the recent past, obtaining good results (Zedda et al., 2024; Anghelina et al., 2024); with similar foundations, we designed a lightweight neural architecture that aims to balance prediction accuracy with computational demand.

The remainder of the paper describes the system architecture along with the experiment setup adopted for this task and an overview of the results obtained.

2 System Overview

The model is a straightforward combination of several independent language models, previously fine-tuned for different tasks, whose results are then concatenated and fed into a neural network, as outlined in Figure 1. This approach has been structured into several steps and modules that are described below.

Data pre-processing Each sentence of an article was extracted and treated as an individual data point (or sample), with the requirement that every sentence from the same article shares the full set of classes attributed to the entire article. This is motivated by the fact that BERT models have a constrained token limit per sentence. Treating each sentence independently, instead of using the article as a whole, allows to manage this limitation effectively, as each sentence can be processed within the model's token constraints, while still preserving the information on the associated narratives of the whole article. This approach also ensures more consistent results when using a voting system, which is indeed an additional component of the system, as explained later in this overview.

Embedding extraction Every file undergoes an embedding extraction process. Several extraction modes were tested to include a context window of previous content. After preliminary experiments, two configurations were ultimately employed for the purposes of this task, i.e., one without a context window and one with a window covering the two preceding sentences. Various fine-tuned models based on BERT architecture (Devlin et al., 2019) and available on HuggingFace were used. These models were employed only to extract textual embeddings, specifically the CLS token embedding for each sentence. Multiple instances of these models were used in parallel. It is worth pointing out that these models were selected because they were already fine-tuned for different (though relevant) tasks, but they were not retrained on this competition's dataset.

Concatenation Module The embeddings extracted from each model are simply concatenated into a single vector and stored in memory along with the labels of the classes associated with the analyzed sample.

Neural Network Once all embedding vectors are stored in memory, they serve as features for a simple neural network. The architecture consists of a non-linear ReLU layer followed by a dropout layer and finally a classification layer that uses either a Sigmoid or Softmax activation function to obtain class probabilities.

Voting system In this module, the predicted probabilities for all sentences of the same article are summed and then normalized. Although a multiplicative aggregation approach was also tested, it proved to be overly sensitive to individual sentence predictions and was therefore discarded.

Classification and class extraction After obtaining class probabilities, since a variable number of classes needs to be predicted, several methods were tested to extract the correct number of classes. In this module, different approaches were explored: initially, a secondary neural network was tested to determine the number of classes. However, this approach was discarded in favor of a classic thresholding method based on results from the development set. This is the most crucial part, as even if the neural network model performs well, choosing the wrong threshold could decrease significantly the results.

3 Dataset

The dataset consists of different news articles and different type of texts about two main topics: Climate Change denial claims (abbreviated with the CC label) and propaganda in Ukraine-Russia war (abbreviated with the URW label). Each article is linked to a set of exclusive narratives based on its topic, and each narrative is further associated with a group of sub-narratives. The dataset has been made available in five languages (with approximately 400-450 articles with golden labels and 100 unlabeled articles to submit per language), with the same taxonomy for each language, except for Russian language where the CC topic was not present and there were far less articles (approximately 250 articles with golden labels and 60 unlabeled articles to submit). In general, 30-40 elements were used for development for each language.

For further details on the dataset development and composition we refer the reader to the task report provided by the organizers (Piskorski et al., 2025).

4 Experiment Setup

The experiments were run on a laptop with an Intel® Core™ i7-1065G7 @ 1.30GHz CPU, 36 GB RAM, CPU only.

As described in Section 2, a preliminary preprocessing step involved splitting each article into individual sentences. Furthermore, all non-English datasets of the task were then automatically translated into English using Microsoft Translate API.

The four models selected for feature extraction are RoBERTa-large fine-tuned on the TweetEmotion dataset for the emotion classification task[2], (Antypas et al., 2023) DeBERTa-v3-small[3](Sileo, 2024) and DeBERTa-v3-base[4](Laurer et al., 2024), both fine-tuned for NLI tasks, and a DistilBERT model fine-tuned for Named Entity Recognition[5](Sanh et al., 2019). On average, extracting the entire dataset for one language composed approximately of 400 samples takes 1-2 hours, which is why the extracted embeddings are saved and later used for experiments (this processing was not done in batch).

Concerning the development of the neural network, local experiments were conducted to determine its optimal configuration. The number of layers and neurons per layer was determined through extensive trial-and-error experimentation. In general, a single ReLU layer with a size 5-30 times smaller than the total number of input features—combined with a dropout rate of 0.3-0.4—was found to be sufficient, with the optimal layer size also depending on the language. Increasing the number of layers often led to overfitting and slower training, and deviations from these values generally resulted in poorer performance. For example, using a different number of neurons caused the loss function on the development dataset to decrease more slowly and converge at higher values compared to the "optimal" configuration, while using too few neurons resulted in underfitting. Overall, the development approach was to carefully adjust these parameters to optimize the neural network's classification performance on the loss function (Binary cross entropy on the one hot encoding of classes of the sample), the objec-

[2] https://huggingface.co/cardiffnlp/twitter-roberta-large-emotion-latest

[3] https://huggingface.co/sileod/deberta-v3-small-tasksource-nli

[4] https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli

[5] https://huggingface.co/dslim/distilbert-NER

Dataset	ReLU size	Window	Activation	Translated	Threshold CC	Threshold URW
English	30	2	softmax	no	0.03	0.03
Portuguese	15	0	sigmoid	yes	0.2	0.15
Russian	5	0	softmax	yes	-	0.1
Bulgarian	10	2	sigmoid	yes	0.1	0.14
Hindi	15	2	softmax	yes	0.05	0.04

Table 1: Hyper-parameters used for the final submission. **ReLU size** indicates how many times the ReLU layer is smaller than the input embeddings, **Window** indicates the number of precedent sentences included as additional embeddings, **Activation** refers to the activation function used in the classification layer, while **Threshold CC** and **URW** refer to the threshold values selected, within each language, for articles on climate change and Ukraine war, respectively.

tive was to reduce the loss on the validation dataset, or to decrease the training loss without causing an increase in the validation loss.

The training process was divided into two phases. The first phase involved using a batch size of 128 sentences and training for 10 epochs, using the Adam optimizer with Keras default settings for rapid convergence. The second phase employed Stochastic Gradient Descent with momentum (SGD) with default parameters, using a batch size equal to the entire dataset and a varying number of epochs (ranging from 1000 to 4000) with early stopping on loss function (to avoid increases on validation loss). This phased approach was chosen because Adam allows the network to learn quickly but tends to overfit after too many epochs. In contrast, SGD with a large batch size learns more slowly but continues to improve the validation loss without overfitting as rapidly on the training dataset. No distinction between sentences of different articles were made as a random shuffle of all sentences was performed before the training phase.

The final threshold was selected by running the system multiple times on the development set and choosing the threshold value that maximized the result according to the official ranking metric, which was samples F_1 score (further explained below). Table 1 summarizes all the hyper-parameters selected and eventually used for the evaluation phase.

For what concerns the evaluation metrics adopted for Subtask 2, submitted results were computed considering two measures: Coarse F_1, which reports how well the model predicts the narratives (without considering the sub-narratives), and Samples F_1, which instead measures how well the model predicts both narratives and sub-narratives, meaning that both aspects should be correct for the prediction to be considered correct. For both coarse and samples F_1 the values are first averaged at document level and then across all the documents of the set. Standard deviation is finally included, in order to measure how much the scores vary across the documents.

5 Results

In this section, we present the results of our experiments across all phases of the campaign, i.e. development, evaluation and post-evaluation, where the leaderboard was made available for participants to continue testing their models following the competition.

Development phase In development phase we noticed a discrepancy on the internal development dataset results and the sent prediction results. We realized that this was mainly due to problems in the evaluation from the server (a critical bug was discovered after the start of the official evaluation phase), thus invalidating this phase actual results.

Test phase As shown in Table 2, while the model did not achieve top-ranking positions, it demonstrated consistent performance across most languages. In four out of five languages, it ranked above the average position among participants. Overall, the best results among the ones submitted by our team, were the ones for Russian, where the model consistently performed better both for the sole narratives and narratives and sub-narratives. Conversely, Hindi exhibited the lowest scores in both metrics.

Post-Test phase Upon reopening the leaderboard, therefore after the evaluation phase officially closed, further tests were conducted to investigate why the model underperformed in certain languages. At the time of writing, simply adjust-

Language	Rank	Participants	F1 coarse	st. dev.	F1 samples	st. dev.
English	9	28	0.467	0.356	0.32	0.321
Portuguese	5	13	0.536	0.273	0.293	0.206
Russian	**3**	14	0.618	0.312	**0.411**	0.308
Bulgarian	**3**	11	0.557	0.335	**0.369**	0.308
Hindi	10	13	0.207	0.313	0.147	0.292

Table 2: Official evaluation results obtained on Subtask 2 across different languages.

Language	F1 coarse	st. dev.	F1 samples	st. dev.	Thresh. CC	Thresh. URW
English	-	-	-	-	-	-
Portuguese	0.547	0.228	0.319	0.182	0.1	0.1
Russian	**0.597**	0.279	**0.44**	0.251	-	0.05
Bulgarian	0.558	0.351	0.391	0.352	0.2	0.19
Hindi	0.227	0.378	0.174	0.339	0.05	0.1

Table 3: Post-task results, empty field means no changes applied to the threshold values nor to the score values.

ing the thresholds without re-training the whole model-—ie., manually increasing or decreasing the values relative to the number of classes for some languages—-led to a slight improvement in the score as reported in Table 3.

Moreover, we further tested the system using alternative language models for the embeddings to understand whether model selection (and, as a result, the chosen type of embeddings) could offer competitive advantages compared to other general newer models. We used in particular ModernBERT-Large pretrained with zero-shot classification [6] (Warner et al., 2024). During testing, the optimal ReLU size was determined through trial and error. Thresholds were manually optimized after a first automatic search, and the number of epochs was 4000 for every language during the training of the neural network. Contrary to the previous experiments, with the Portuguese data we used the sigmoid function instead of softmax (see Table 1). As shown in Table 4, most result are lower than the ones obtained with the other models used for the campaign, however it is worth pointing out that English ModernBERT got a relatively higher score (ranking 5th in the post-task leaderboard[7]).

[6] https://huggingface.co/MoritzLaurer/ModernBERT-large-zeroshot-v2.0

[7] https://propaganda.math.unipd.it/semeval2025task10/leaderboard.php, as of April 23rd, 2025

6 Discussion and Error Analysis

While the Bulgarian and Russian datasets performed better than the other languages, despite being translated into English, surprisingly enough, the same model produced worse results on the original English dataset. Upon the re-opening of the submissions, comparing Table 2 and Table 3 several conclusions can be drawn: the thresholding mechanism appears to work but may not always yield globally optimal results. Additionally, it has been observed that as the prevalence of the "Other" class increases, the model's overall performance declines compared to that of other participants, in fact the lowest score was obtained in datasets with prevalence of this class. We can also note that, for some reason, top-ranking models in other languages similarly achieved worse results on the Indian language task. This could be due to semantic discrepancies between different dataset languages. Overall, despite the model's limitations, we hypothesize that the issue stem more from the dataset than the model itself. This is largely due to the fact that all languages were translated into English before processing.

The confusion matrices generated for the results on each development set (see Appendix A) reveal that the model tends to produce a high number of false positives in certain classes. However, these classes also show a higher correct prediction rate, which suggests a significant class imbalance in the dataset. In contrast, some other classes are consistently missed, resulting in a high number of false

Language	ReLU size	F1 coarse	st. dev.	F1 samples	st. dev.	Thresh. CC	Thresh. URW
English	10	0.498	0.363	**0.373**	0.373	0.01	0.01
Portuguese	10	0.448	0.274	0.234	0.185	0.08	0.1
Russian	2	0.59	0.256	0.329	0.218	-	0.06
Bulgarian	10	0.381	0.383	0.256	0.355	0.13	0.14
Hindi	15	0.174	0.277	0.088	0.231	0.04	0.06

Table 4: Post-task results using only modernBERT for the embeddings. Empty field means no changes in threshold values.Improved results (in terms of samples F_1) with respect to the task evaluation phase are highlighted in bold.

negatives distributed across various classes, however the false negatives are low within each individual class, further confirming the imbalance. Relatively very few cases have completely correct predictions.

7 Conclusions and Future Work

The approach described in this paper consisted in feeding a simple neural network with a concatenation of multiple contextual embeddings from different fine-tuned BERT-based models to tackle the challenges deriving from a multi-class and multi-label classification task. Quantitatively, the model performed reasonably well in some languages, but underperformed in others, as the model seems to struggle to find optimal values of thresholding and it is sensitive to class definitions.

In terms of possible improvements over this approach, we observe that the Russian dataset–that only included URW-related topics–performed better than other languages, suggesting that developing separate classifiers for each topic (CC *vs* URW) might further improve results. Additionally, using a single-language merged dataset approach could also yield better performance. Another unexplored approach is a top-down hierarchical strategy. However, given that the narrative/coarse score was not particularly low across different languages, this approach may not be necessary.

Code availability

The code of the model is avaiable on Github in the repository: `https://github.com/demon-prin/iltc-narrative-classification`

Acknowledgements

The work has been partially supported by the project DEMON "Detect and Evaluate Manipulation of ONline information" funded by MIUR under the PRIN 2022 grant 2022BAXSPY (CUP F53D23004270006, NextGenerationEU), and by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU (NextGenerationEU).

References

Ion Anghelina, Gabriel Buță, and Alexandru Enache. 2024. SuteAlbastre at SemEval-2024 task 4: Predicting propaganda techniques in multilingual memes using joint text and vision transformers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 443–449, Mexico City, Mexico. Association for Computational Linguistics.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

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 American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Moritz Laurer, van Atteveldt, Wouter, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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(8):8393–8435.

Damien Sileo. 2024. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia. ELRA and ICCL.

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).

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Rheeya Uppaal, Junjie Hu, and Yixuan Li. 2023. Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12813–12832, Toronto, Canada. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Luca Zedda, Alessandra Perniciano, Andrea Loddo, Cecilia Di Ruberto, Manuela Sanguinetti, and Maurizio Atzori. 2024. Snarci at SemEval-2024 task 4: Themis model for binary classification of memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 853–858, Mexico City, Mexico. Association for Computational Linguistics.

A Confusion Matrices

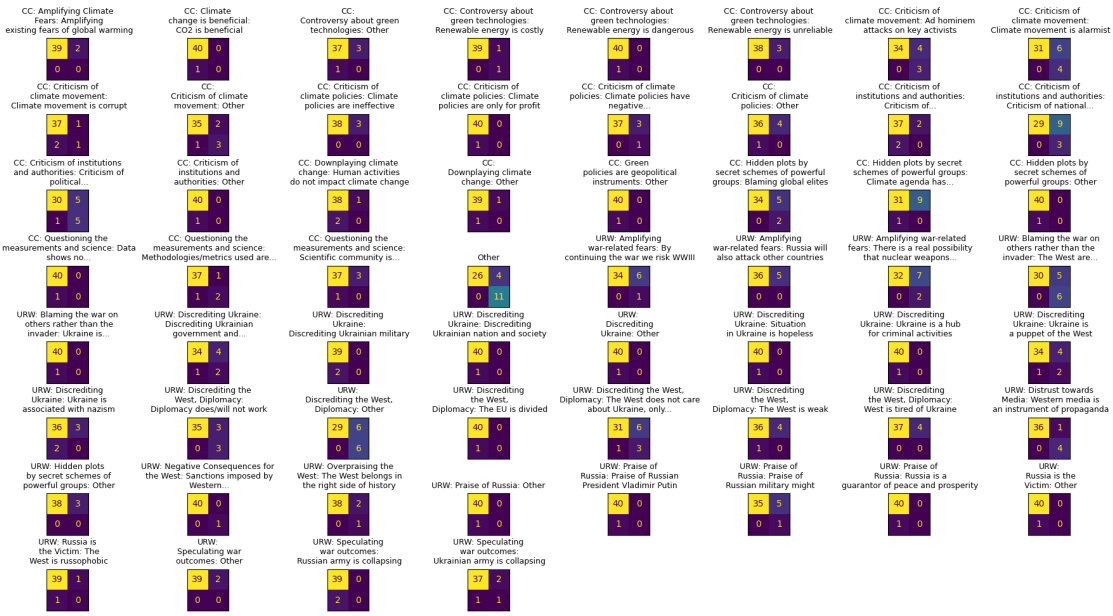

Figure 2: Confusion matrix on English dev set with highest score in post-task (relevant entries).

Figure 3: Confusion matrix on Portuguese dev set with highest score in post-task (relevant entries).

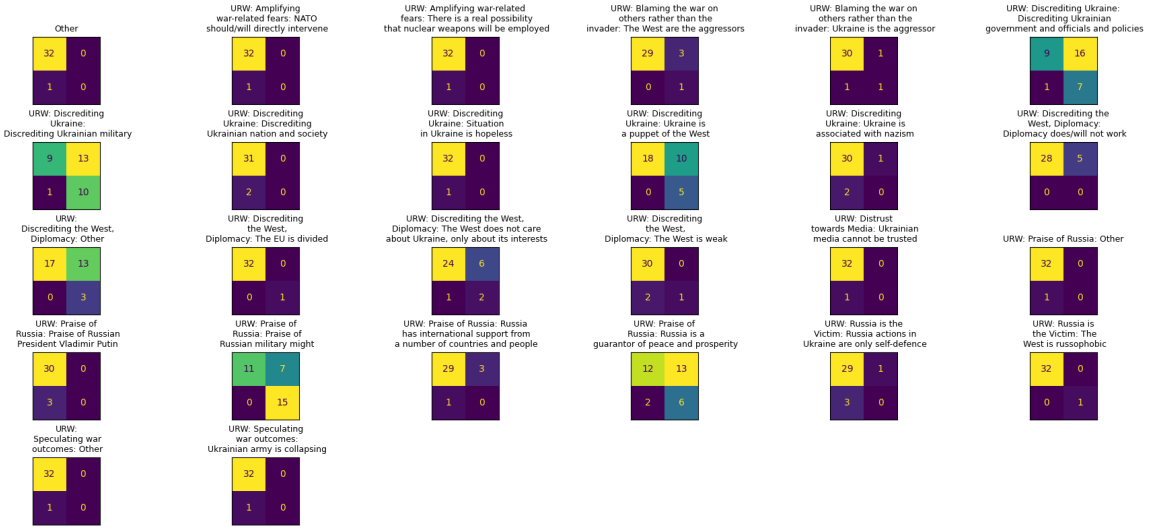

Figure 4: Confusion matrix on Russian dev set with highest score in post-task (relevant entries).

Figure 5: Confusion matrix on Bulgarian dev set with highest score in post-task (relevant entries).

Figure 6: Confusion matrix on Hindi dev set with highest score in post-task (relevant entries).