NarratEX Dataset: Explaining the Dominant Narratives in News Texts

Nuno Guimarães[1,2], Purificação Silvano[1,2], Ricardo Campos[1,3], Alípio Jorge[1,2]
Ana Filipa Pacheco[1,2], Dimitar Iliyanov Dimitrov[7], Nikolaos Nikolaidis[5]
Roman Yangarber[8], Elisa Sartori[6], Nicolas Stefanovitch[5], Preslav Nakov[9]
Jakub Piskorski[10], Giovanni Da San Martino[6]

[1]INESC TEC, [2]University of Porto, [3]University of Beira Interior,
[4]Department of Informatics, Athens University of Economics and Business,
[5]European Commission Joint Research Centre, [6]University of Padova,
[7]Sofia University "St. Kliment Ohridski", [8] University of Helsinki, [9]MBZUAI,
[10]Institute of Computer Science, Polish Academy of Science

Correspondence: nuno.r.guimaraes@inesctec.pt

Abstract

We present NarratEX, a dataset designed for the task of *explaining the choice of the Dominant Narrative in a news article*, and intended to support the research community in addressing challenges such as discourse polarization and propaganda detection. Our dataset comprises 1,056 news articles in four languages, Bulgarian, English, Portuguese, and Russian, covering two globally significant topics: the Ukraine-Russia War (URW) and Climate Change (CC). Each article is manually annotated with a dominant narrative and sub-narrative labels, and an explanation justifying the chosen labels. We describe the dataset, the process of its creation, and its characteristics. We present experiments with two new proposed tasks: *Explaining Dominant Narrative based on Text*, which involves writing a concise paragraph to justify the choice of the dominant narrative and sub-narrative of a given text, and *Inferring Dominant Narrative from Explanation*, which involves predicting the appropriate dominant narrative category based on an explanatory text. The proposed dataset is a valuable resource for advancing research on detecting and mitigating manipulative content, while promoting a deeper understanding of how narratives influence public discourse.

1 Introduction

The Internet has become a powerful medium for disseminating information, but it has also amplified the spread of polarized content and deceptive and manipulated narratives. This is particularly evident in contentious topics such as geopolitical conflicts and global challenges like climate change, where dominant narratives –central perspectives or storylines–, play a significant role in shaping public opinion and fostering division.

Extremely biased and polarized discourse occurs in different regions and cultures and can be observed practically in any language found online (Pasquali et al., 2016; Addawood et al., 2019; Nakov et al., 2021a,b). The intricacies of writing can sometimes obscure the intended message, especially in cases of disinformation or highly biased political content. Therefore, making the narrative more explainable is essential for mitigating confirmation bias and strengthening readers' capacity to interpret a given piece of text (Schmitt et al., 2024). Therefore, understanding and analyzing these narratives is critical for identifying disinformation and propaganda, and for addressing their potential societal impact.

Despite the growing interest in narrative analysis (Campos et al., 2024), there is a lack of comprehensive multilingual datasets that address the challenges of narrative classification and explanation in diverse linguistic and cultural contexts. Existing corpora (Camburu et al., 2018; Kočiský et al., 2018; Lal et al., 2021) are often limited to specific aspects, such as reading comprehension or causal reasoning, and do not capture the interplay between narrative categorization and explanatory reasoning. To address this gap, we introduce NarratEX, a dataset designed to advance research in identifying and explaining dominant narratives and sub-narratives. The dataset consists of 1,056 news articles in four languages (Bulgarian, English, Portuguese, and Russian), covering the Ukraine-Russia War (URW)[1] and Climate Change (CC).

[1]While the choice of English and Russian (spoken in both Russia and Ukraine) is straightforward, we also included Bulgarian and Portuguese. This decision reflects the fact that former Soviet states and Latin American countries, including Brazil, are key targets of pro-Kremlin propaganda (see also Section 8 for further discussion of limitations).

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20408–20434
November 4-9, 2025 ©2025 Association for Computational Linguistics

Each article in our dataset is labeled with a *dominant narrative* (spanning 21 classes), a *sub-narrative* (spanning 74 sub-classes), and an *explanation* that justifies their selection for the given article. Our dataset uniquely bridges the gap between narrative classification and explanatory reasoning, enabling a deeper understanding of how disinformation and propaganda are framed, spread, and interpreted online. Our contributions are as follows:

- A novel multilingual dataset annotated for dominant narrative and sub-narrative classification and respective explanation,[2] including a description of the annotation process and the guidelines used.
- Analysis of the dataset and the annotations, exploring the distribution of dominant narratives and sub-narratives across languages and topics and the semantic similarity of the explanations.
- Comprehensive experiments for generating narrative explanations and inferring dominant narratives from explanations, leveraging state-of-the-art pre-trained generative models as well as zero-shot learning with large language models.

2 Related Work

Explaining text narratives has become particularly relevant in domains like disinformation and propaganda detection, where understanding the deeper structure of narratives is essential for reliable analysis. Since these tasks are challenging even for humans, automated models need not only accurate narrative classification, but also interpretable justifications for their predictions. As a result, several datasets have become vital in Natural Language Processing (NLP) for tasks like machine reading comprehension (Richardson et al., 2013; Huang et al., 2019; Nguyen et al., 2016; Hermann et al., 2015), text summarization (Kryscinski et al., 2022), and explainable Artificial Intelligence (AI) (Martens et al., 2025). These datasets have provided resources to train machine learning models that predict outcomes, reason explicitly, and explain their decision-making processes, ultimately enhancing our understanding of narrative structures.

For instance, the NarrativeQA dataset (Kočiský et al., 2018) offers detailed question–answer pairs focused on story comprehension, supporting tasks such as contextual reasoning and summarization. The TellMeWhy dataset (Lal et al., 2021) focuses on causal reasoning by providing explanations for the events that occur in a narrative, enhancing a model's ability to identify cause-and-effect relationships. Other datasets such as NarraSum (Zhao et al., 2022) and SummScreen (Chen et al., 2022) have further contributed by pairing narrative texts with their summaries, assisting models in learning to condense and to reframe narrative content. Additionally, rhetorical analysis datasets from shared tasks like SemEval and CheckThat! (Da San Martino et al., 2020; Piskorski et al., 2023a, 2024; Alam et al., 2025) have provided annotations that highlight persuasive strategies and logical fallacies, enabling models to identify narrative techniques, particularly in argumentative or manipulative content. Among resources emphasizing explainability, the e-SNLI dataset (Camburu et al., 2018) stands out by adding human-written justifications to natural language inference labels, thus guiding models toward transparent entailment reasoning.

While the above-described datasets from previous work have enhanced interpretability, they have often focused on specific reasoning tasks such as entailment, causality, or provide explanations indirectly through summarization or question-answer formats. In contrast, our proposed dataset introduces a novel approach by directly linking each narrative-labeled text to a concise, human-annotated explanation that explicitly justifies both the dominant narrative and the sub-narrative assignments. Unlike current datasets that focus on outcomes such as summaries or event-based explanations, our resource bridges narrative understanding with interpretability by presenting the rationale behind narrative categorization itself. This design enables a deeper exploration of the interplay between narrative structure and its communicative function, supporting models that not only identify narratives, but also explain why a particular narrative interpretation is plausible. For these reasons, our dataset makes a unique contribution to the field of explainable narrative analysis, particularly relevant in various applications requiring high-stakes narrative reasoning, such as misinformation analysis. We hope that it will represent a valuable resource for studying narratives and interpretability of misinformation detection.

[2] https://github.com/LIAAD/NarratEX/

3 Dataset Construction and Annotation

3.1 Data Acquisition

Our article selection process followed a structured approach to ensure a relevant collection of news articles across four target languages and two key subjects: the Ukraine-Russia War (URW) and Climate Change (CC), the first covering narratives surrounding Russia's large-scale invasion of Ukraine, which began in February 2022, and the second about climate change denial and activism focused on mitigating its impact. The selection of articles lasted five months, from August to December 2024. To build the dataset, we took the following steps:

1. We tailored topic-specific keyword-based queries to mine documents relevant to each subject, such as *Ukraine war*, *Ukraine-Russia*, *climate change*, *global warming*, etc. The scope was limited to articles published between January 2022 and August 2024. We used these queries to retrieve a large collection of news articles via an in-house news aggregation tool, supplemented with region-specific sources (e.g., Portuguese) to emphasize content relevance.

2. To further refine the dataset, we performed zero-shot relevance classification using the BART-large-mnli model (Lewis et al., 2020). We applied the model to each article's title and the first 300 characters of text in combination with the set of labels to get a relevance score per label. For zero-shot labels, we selected a secondary set of keyphrases, such as *Denazification of Ukraine* for URW and *Climate hoax* for CC. We further used a RoBERTa-based multi-label classifier, trained on the Persuasion Techniques dataset (Piskorski et al., 2023a,b) to assess the persuasion techniques used in each article.

3. We integrated these scores using linear weighting, ranking articles from most to least likely to contain relevant narratives. To ensure a mid-to-large size of the news articles, we filtered out articles with fewer than 250 words.

4. We manually reviewed each article to confirm its relevance to the annotation task, ensuring that all selected texts contained narratives with the potential to mislead or manipulate the reader.

3.2 Annotation Process

In order to streamline the annotation process, a specialized team was responsible for each of the four languages in our corpus: Bulgarian, English, Portuguese (the European variety), and Russian. Each team was overseen by a designated language coordinator and consisted of three to six annotators with expertise in fields such as linguistics, social sciences, international relations, or prior experience in annotation tasks. The annotators underwent thorough training, including studying detailed annotation guidelines, participating in live demonstrations, and engaging in real-time annotation exercises. Each article was annotated by two different annotators. We held weekly meetings within each team and between the language teams in order to address ambiguities, to resolve disagreements, to ensure consistency across annotations, and to refine the annotation guidelines.

The first annotation task involved identifying the dominant narrative and sub-narrative at the document level. In this context, we adopted the narrative definition from Nikolaidis et al. (2025). Therefore, a narrative is defined as "*a recurring, repetitive (across and within articles), overt or implicit claim that presents and promotes a specific interpretation or viewpoint on an ongoing (and frequently dynamic) news topic*." The annotation was based on a two-level, domain-specific taxonomy of narrative labels related to the following two topics: Ukraine-Russia War (Appendix A.1) and Climate Change (Appendix A.2) featuring coarse-grained labels (representing primary narratives) and fine-grained labels (representing sub-narratives). These taxonomies were adapted from the work of Coan et al. (2021) and Amanatullah et al. (2023) by media analysts considering their experience in media monitoring, with some modifications. The modifications included adding new sub-narratives, splitting sub-narratives that were overly broad, and consolidating sub-narratives that were highly fragmented. In addition, we did some rephrasing to phrase each narrative as a concrete claim. Note that the final taxonomies reflect the expertise of the media analyst and are not intended to be complete taxonomies covering all the narratives within the two topics. Moreover, although the taxonomies were developed to identify manipulative narratives commonly linked to disinformation, we did not assume that all claims within the taxonomy constituted disinformation.

As an example, the dominant narrative within the topic of Climate Change might be "*Downplaying climate change*", while its sub-narrative could be "*Climate changes are natural*". These labels reflect the situation where the article's authors minimize the impacts of climate change by arguing that it is a natural cyclical phenomenon. If the narratives from the taxonomy did not fully capture the main claim of a news article, the annotators would select the label *Other*.

The next step involved highlighting the textual evidence that supports and validates the dominant narrative and sub-narrative. This was used as an intermediate step to achieve better convergence in the annotations and aid the curation of the explanation. Based on this evidence, the annotators were instructed to write concise explanations justifying their selection of the narrative label without explicitly naming the narratives. To guide the formulation of the explanation, we provided detailed instructions to the annotators as follows:

- We gave the explanation, written in the language of the article, should summarize the textual evidence, including arguments, counterarguments, behaviors, stances, or opinions that support the choice of narrative.
- We further gave the entities mentioned in the articles that were relevant to the dominant narrative and sub-narrative must be included in the explanation.
- The annotators were required to justify the selection of the dominant narrative and sub-narrative, addressing the question "Why were X and Y chosen as the dominant narrative and sub-narrative?"
- The explanation must be composed using the annotator's own words, avoiding direct quotations except for brief phrases or expressions, and must not exceed 80 words.

We further provided style recommendations:

- The annotators were asked to explicitly reference the entities and their actions or statements where possible to support the narrative selection.
- If explicit entities, actions, or statements were unavailable, we encouraged the annotators to use phrases such as "*the text reports*," or "*the text's author*" to justify their reasoning.
- The annotators were asked to avoid *restating* the dominant narrative and sub-narrative, and rather focus on the reasoning behind their selection.

To ensure consistency across annotators and languages, language coordinators held regular meetings to compare and harmonize explanation styles. In addition, each language had one or more curators who verified whether the predefined guidelines had been followed. These curators were experts in linguistics, computational linguistics, or computer science, with extensive experience in curating documents for various natural language processing (NLP) tasks. Some of them held master's degrees, while others are master's or PhD students in these fields. Their task was to validate the final selection of the dominant narrative and sub-narrative and the produced explanations, assess their accuracy, and select the most suitable one. If neither explanation was adequate, the curators merged the strongest elements of both or, if necessary, drafted a new explanation. This systematic process ensured high-quality and consistent explanations across all annotations.

The annotation process was carried out using the Inception platform (Klie et al., 2018). More details about the annotation process in Inception, including a full annotated example, are presented in Appendix B.

3.3 Inter-Annotator Agreement (IAA)

To highlight the relevance of this dataset for the research community, we assessed the agreement between annotators in two distinct dimensions: One that evaluates the agreement between the annotators in selecting the same dominant narrative, thus allowing us to assess the difficulty of the multi-stage annotation process; and another one that evaluates how similar the explanations of two annotators are on the same document, thus allowing us to measure the coherence and the consistency of narrative understanding across different annotators.

3.3.1 IAA on the Dominant Narrative

To assess IAA on the dominant narrative, we calculate Krippendorff's α between the annotators. The results are presented in Table 1 for all the documents on both topics (URW and CC).[3]

[3]CC was not annotated for Russian, since we did not succeed in crawling Russian-language news articles disputing climate change and global warming in the selected time period.

	Bulgarian	English	Portuguese	Russian
α	0.540	0.409	0.332	0.338

Table 1: Agreement between the annotators in the dominant narrative by language using Krippendorff's α.

The overall results are relatively low and below the recommended value ($\alpha \geq 0.667$) and can be explained by the inherent complexity and subjectivity of the annotation task. First, the high number of dominant narratives for each topic—11 in URW and 10 in CC— combined with the very fine-grained taxonomy raises disagreement in close labels (although it also enriches the dataset with more domain-specific annotations). Second, agreement in tasks that are often charged with political bias can be difficult and more complex for annotators (Stefanovitch and Piskorski, 2023; Piskorski et al., 2024). Nevertheless, the agreement obtained is still on par with the agreement on similar tasks (Stefanovitch and Piskorski, 2023; Piskorski et al., 2024). Furthermore, we stress that these values represent the agreement prior to the curation phase. Thus, the low agreement is more representative of the difficulty of the task than the quality of the final labels (consolidated by the curator). These reinforce the need and the justification for the high expertise of our curators, since they were essential to ensure the quality of the final dataset. Details about the inter-annotator agreement, including the labels where the annotators disagree the most, are presented in Appendix C, and analysis of the sub-narratives is given in Appendix D.1.

3.3.2 IAA for Explanation Writing

Building on the previous experiment, we assessed the IAA for the explanations by comparing the textual justifications by each annotator for a given document and measuring their similarity. Naturally, we only considered the cases where both annotators were in agreement with the dominant narrative, since the explanation depends on it. We further excluded the *Other* topic since the explanation was written only to justify the dominant and the sub-dominant narratives on the URW and CC topics.

To this regard, we computed the explanation agreement using precision (P), recall (R), and F1-score (F_1) resorting to BERTScore (Zhang et al., 2019) and a multilingual BERT model compatible with the four languages of the dataset.[4]

[4] https://huggingface.co/google-bert/bert-base-multilingual-cased

lang	topic	P	R	F1	lev_chr	lev_tok	lcs_chr	lcs_tok
BG	CC	0.695	0.695	0.695	139	26	66	1.74
	URW	0.721	0.720	0.720	120	25	61	1.67
EN	CC	0.741	0.736	0.738	155	32	75	1.89
	URW	0.751	0.748	0.749	152	32	91	3.36
PT	CC	0.728	0.736	0.732	206	43	111	2.11
	URW	0.734	0.733	0.732	197	41	99	1.72
RU	URW	0.694	0.690	0.691	175	30	62	1.45

Table 2: Explanation agreement results averaged by language (lang) and topic, in terms of precision (P), recall (R), and F1-score (F_1), calculated using BERTScore, Levenshtein distance at the token level (lev_tok) and at the character level (lev_chr), and longest common subsequence at the token level (lcs_tok) and at the character level (lcs_chr).

Furthermore, we calculated the Levenshtein distance (lev) and the longest common subsequence (lcs) at the token (tok) and character (chr) level. Table 2 shows the results, organized by language and topic.

The evaluation results in Table 2 show that for all language-topic pairs, we achieved an F1-score above 0.69, with the similarity between the explanations concerning the Ukraine–Russian War topic being more aligned than in the Climate Change topic for all languages except for Portuguese, which achieved the same F1 for both topics. At the language level, English has the most robust F1 in both topics and the second highest lcs on the URW topic. The Portuguese annotators differed most in the explanations written when considering Levenshtein (i.e., on average, more operations were required to convert one explanation into the other). However, Portuguese also has on average the longest common sequence of characters. Although linguistic differences can influence these metrics, English explanations, particularly in the URW topics, seem to be where the annotators were more aligned, combining a high F1-score and lcs across all language-topic pairs, despite having the third-best Levenshtein score. Although the combination of BERTScore, Levenshtein distance, and lcs provides a clearer understanding of the annotator agreement when comparing different languages and topics, for a full comprehension of the agreement, these measures are not sufficient, and thus, a more exhaustive and human-assessed evaluation must be conducted.

The next section presents a characterization of the final dataset, NarratEX, which is composed of the dominant narrative, the sub-narrative, and the explanations chosen by the curators for each language.

Bulgarian		English		Portuguese		Russian	
train	test	train	test	train	test	train	test
357	28	203	30	252	25	133	28

Table 3: Number of documents in the training and the test sets for each language.

4 NarratEX Dataset Characterization

The final dataset comprises 1,056 entries divided into 945 for training and 111 for testing. Each entry is composed of a document, a dominant narrative and sub-narrative, and an explanation that justifies the choice of the dominant narrative and sub-narrative for that document. A sample of complete documents and respective annotations are presented in Annex G. The number of entries per language in each set is shown in Table 3. Table 4 presents the frequency of the dominant narratives in the dataset.

Topic: Ukraine-Russia War	#docs
Discrediting Ukraine	195
Discrediting the West, Diplomacy	146
Praise of Russia	99
Amplifying war-related fears	94
Blaming the war on others rather than the invader	54
Russia is the Victim	36
Speculating war outcomes	34
Negative Consequences for the West	19
Distrust towards Media	13
Hidden plots by secret schemes of powerful groups	10
Overpraising the West	5
Topic: Climate Change	**#docs**
Amplifying Climate Fears	153
Criticism of institutions and authorities	60
Criticism of climate policies	35
Criticism of climate movement	33
Downplaying climate change	25
Hidden plots by secret schemes of powerful groups	23
Questioning the measurements and science	9
Controversy about green technologies	8
Climate change is beneficial	3
Green policies are geopolitical instruments	2

Table 4: Number of documents per dominant narrative.

Looking at Table 4, one can observe that *Discrediting Ukraine* and *Amplifying Climate Fears* are the most frequent dominant narratives, with 195 and 153 entries, respectively. One can further observe that on the topic of the Ukraine-Russia war, the two most frequent dominant narratives are both inclined towards discrediting countries and foreign policies (*Discrediting Ukraine* and *Discrediting the West, Diplomacy*). However, for Climate Change, the two most frequent dominant narratives, *Amplifying Climate Fears* and *Criticism of institutions and authorities*, represent both supportive and opposing perspectives on the topic, as the latter can encompass both pro- and anti-climate change viewpoints. More detailed information regarding the frequency of the individual sub-narratives is presented in Appendix D.2.

Figure 1: Similarity of the explanations by topic (URW in blue and CC in red) using LaBSE.

We further explored and analyzed the characterization of the explanations by assessing how close these are across languages by extracting textual embeddings using LaBSE (Feng et al., 2022), a language-agnostic sentence embedding, and applied t-SNE (van der Maaten and Hinton, 2008) for dimensionality reduction. The results for the explanations embeddings by topics are shown in Figure 1. We can see that the explanations in multiple languages are aligned, with the two clusters (concerning each topic) clearly separated. However, some CC explanations are closer to the border of the URW topic. These are mainly explanations of English articles, which address topics such as the association of electric vehicle production with neo-colonial practices and slavery, and attacks on political factions regarding energy policies. Thus, terms such as *slavery* and *attack* could potentially influence the placement of these documents near the cluster of the URW topic. Another exception is the explanation of an English-language URW document, which is placed near the center of the CC cluster. This explanation states that "*UN Secretary-General António Guterres comments on possible impacts on the world economy*." Thus, we hypothesize that the lack of terms directly related to war or politics concerning Russia or Ukraine may lead to the proximity of the explanation to this cluster.

5 Experiments and Evaluation

To showcase the potential of the NarratEX dataset, we conduct two experiments on two possible tasks where the dataset can be used. The first is the generation of the narrative explanation for the dominant narrative and sub-narrative.

The second is a multiclass classification task that aims to classify each explanation into the proper dominant narrative. For each task, we provide random and zero-shot baselines based on open and closed generative LLMs and their respective results. For the multiclass task, we also added a fine-tuned BERT model. Since the majority of the approaches in both tasks are unsupervised/zero-shot, we present the evaluation results in both the test set as well as in the entirety of the dataset (train+test).

5.1 Generation of Narrative Explanation

To demonstrate the value of this dataset, we designed a task that generates explanations for each document's dominant narrative and sub-narrative. In other words, given a document d and its dominant narrative n_d and sub-narrative s_d (assessed by the annotators), the goal is to generate an explanation e_d that justifies those narratives.

As this is a text-generation task, we experimented with five different baselines using pre-trained generative models. More specifically, we selected three open-weights LLMs: Llama-3.1-it (Grattafiori et al., 2024) with 8B parameters, Phi3-small-8k-Instruct (Abdin et al., 2024) with 7B parameters, and Gemma2-2b-it (Gemma Team, 2024) with 2B parameters, as well as two proprietary LLMs: Gemini-1.5-flash-002 and Gemini-1.5-pro (Gemini Team, 2024). Next, we wrote a prompt that resembled the instructions provided to the annotators during annotation, to be used by all five models in the evaluation process. Given the limited context length of some of the models chosen, we had to ensure that the prompt contained only the essential information. First, we required the explanation to be written in the language of the document. Second, we instructed the LLM to write the explanation in 80 words or less. The final prompt passed to the model is shown in Appendix E.1. We used truncation in cases where the prompt exceeded the context length of the model due to lengthy documents. The generated explanation was also post-processed, i.e., truncated, whenever needed, to ensure that it met the 80-word criteria. In addition, to better understand the results from the zero-shot baselines, we also added a random baseline where a sequence of 80 consecutive tokens from the original document is used as the predicted explanation.

Table 5 shows the results in terms of precision, recall, and F1-score, calculated using BERTScore (with the same multilingual BERT model as the one we used in Section 3.3.2).

		Test			Full		
Lang	**Model**	**P**	**R**	**F1**	**P**	**R**	**F1**
PT	random	0.632	0.656	0.643	0.677	0.707	0.692
	gemini-1.5-flash	0.640	0.652	0.646	0.698	0.691	0.694
	gemini-1.5-pro	0.648	0.660	0.654	0.685	0.687	0.686
	gemma2	0.645	0.658	0.651	0.683	0.681	0.682
	llama-3.1-it	0.655	0.677	0.666	0.720	0.735	0.727
	phi3	0.659	0.680	**0.669**	0.722	0.736	**0.729**
EN	random	0.606	0.645	0.625	0.632	0.689	0.659
	gemini-1.5-flash	0.660	0.678	**0.669**	0.695	0.725	**0.709**
	gemini-1.5-pro	0.662	0.674	0.668	0.682	0.721	0.701
	gemma2	0.659	0.677	0.668	0.683	0.719	0.700
	llama-3.1-it	0.648	0.674	0.661	0.674	0.721	0.696
	phi3	0.654	0.677	0.665	0.678	0.717	0.697
RU	random	0.574	0.631	0.601	0.604	0.683	0.641
	gemini-1.5-flash	0.621	0.642	0.631	0.653	0.668	0.660
	gemini-1.5-pro	0.631	0.646	0.638	0.646	0.669	0.657
	gemma2	0.623	0.643	0.633	0.649	0.668	0.658
	llama-3.1-it	0.617	0.665	0.640	0.655	0.708	**0.680**
	phi3	0.630	0.660	**0.644**	0.654	0.684	0.669
BG	random	0.581	0.640	0.609	0.616	0.690	0.651
	gemini-1.5-flash	0.617	0.630	0.623	0.652	0.662	0.657
	gemini-1.5-pro	0.623	0.637	0.630	0.643	0.664	0.653
	gemma2	0.619	0.635	0.627	0.641	0.655	0.648
	llama-3.1-it	0.625	0.669	0.646	0.668	0.724	**0.695**
	phi3	0.634	0.666	**0.649**	0.664	0.689	0.676

Table 5: Results for the narrative explanation task by language using BERTScore in the test and full dataset.

The results obtained in terms of F1-score indicate superior performance for the open-source models, except for English, where Gemini-1.5-flash achieves slightly better results. Llama and Phi achieved the best performance for the remaining languages. A deeper analysis reveals that although the prompt explicitly mentions that the explanation should be given in the language of the document, several models fail to interpret that instruction. For example, for the Portuguese documents, Phi-3 generates the majority of the explanations in Portuguese, while Gemini-1.5-flash answers the majority in English. This can impact and justify some of the results achieved since Gemini's best results are in English: F1-score of 0.669 on test and 0.709 on the full data. In addition, it is also important to highlight the scores achieved by the random baselines, which, when evaluated in the entirety of the dataset, surpass zero-shot approaches in some scenarios, notably in Portuguese and Bulgarian. A detailed analysis of the generated explanations for Bulgarian reveals that, similarly to Portuguese, the models often failed to generate explanations in Bulgarian, which negatively impacted our BERTScore-based evaluation measures. Furthermore, the small gap between random and zero-shot scores provides important context for interpreting future model evaluations and performance benchmarks. These models (and evaluation) should, however, be interpreted as baselines rather than definitive solutions, as they are meant to stimulate further research in this novel task.

To complement our BERTscore analysis, we conducted an additional experiment using human judgment to assess the similarity between the generated content and the ground-truth explanation on a scale of 1 (not at all similar) to 5 (fully similar). We did this on the test set of the English data, since English is the language with the best score on average. The guidelines and details about this evaluation are presented in Appendix F.1, while the results of the evaluation are shown in Table 6.

Note that Gemini-1.5-Pro achieves the best results with the second-lowest standard deviation. These results are closely aligned with the results from BERTScore in English, where this model achieves the second-highest F1-score with a 0.01 difference from the first (Gemini-1.5 - flash). The model that yields the worst scores and that seems to generate barely reasonable explanations is Phi3 with a 2.63 score and 0.89 standard deviation. Once again, this score is consistent with the ranking based on BERTScore's F1 in the English test data, where the Phi-3 model achieved the second lowest score among all tested models (excluding the random baseline). The remaining models produce, on average, reasonable explanations, with Llama-3.1 achieving the highest score of the three, but also the highest standard deviation, showing a lack of consistency in the explanations generated. Overall, the proprietary Gemini-1.5-Pro model produces the best explanations among the five models. LLaMA-3.1 ranks second, but shows higher variability, as indicated by its larger standard deviation. In contrast, Gemini-1.5-Flash scores are slightly lower, but it demonstrates greater robustness, with a lower standard deviation. Although there seems to be some alignment between BERTScore and human evaluation, we complement the quantitative evaluation with a qualitative error analysis by manually inspecting a subset of the generated explanations. This analysis is presented in Appendix F.2 and showcases some of the current limitations of BERTScore as an evaluation measure for this task.

Models	Average	Std
gemini15-flash	3.03	0.76
gemini15-pro	3.90	0.84
gemma2	3.00	0.98
llama31-it	3.33	1.15
phi3	2.63	0.89

Table 6: Average and standard deviation (Std) of human evaluation scores of the narrative explanation task on the English test data.

5.2 Dominant Narrative Inference from Justification

The second task is a multiclass classification task. Given a justification e and the list of possible dominant narratives N, the goal is to predict the correct narrative n, where $n \in N$. Similarly to the previous task, we approached the problem using generative models and zero-shot learning. The prompt we used is presented in Appendix E.2.1.

Note that in this formulation of the task, we provided the model with the 21 possible dominant narratives without any prior knowledge about the topic (CC or URW). We evaluated the task using accuracy and macro precision, recall, and F1-score.

This time, we selected a smaller number of models from the previous tasks. Specifically, we selected the proprietary model that had the best performance at the language level (Gemini-1.5-flash) and the two best open-weights models using the same criteria (Llama-3.1-it and Phi-3). In addition, we also added a random classifier that randomly predicts one narrative for each example. Furthermore, we fine-tuned a BERT model for the task. We chose the same multilingual BERT model used previously due to its compatibility with the languages represented in the dataset. The values of the hyperparameters we used for fine-tuning are presented in Appendix E.2.2. The results for the different baselines grouped by language are presented in Table 7. Additional results regarding the discrimination of the evaluation metrics by the narrative type, model, and language are included in the dataset repository.

We can see that, contrary to the previous task, the proprietary model (Gemini-1.5-flash) outperforms the open-weights ones in all languages in terms of both macro F1 and accuracy on the full dataset. When analyzing the test set and with the inclusion of the fine-tuned BERT model, the only difference is for Portuguese, where the fine-tuned model is slightly better. We reinforce that these results serve as baselines for the task since we hypothesize that prompt engineering adapted to each language and model, and BERT-specific hyperparameter tuning would yield better results. Nevertheless, given that we experiment with 21 dominant narratives, it is clear that the baselines provided surpass simpler ones (such as the random baseline, which has an accuracy of 5%) and thus offer a more challenging starting point for future research.

		Test				Full			
Lang	**Model**	**P_macro**	**R_macro**	**F1_macro**	acc	**P_macro**	**R_macro**	**F1_macro**	**Acc**
PT	random baseline	0.056	0.009	0.015	0.080	0.236	0.076	0.105	0.076
	gemini-1.5-flash	0.450	0.375	0.373	0.720	0.415	0.367	**0.338**	0.632
	llama-3.1-it	0.292	0.317	0.266	0.640	0.101	0.067	0.066	0.437
	phi3	0.327	0.360	0.338	0.680	0.205	0.170	0.164	0.588
	bert-base-multilingual (FT)	0.326	0.536	**0.383**	0.720				
EN	random baseline	0.028	0.009	0.014	0.033	0.069	0.039	0.048	0.039
	gemini-1.5-flash	0.319	0.296	**0.277**	0.500	0.467	0.416	**0.400**	0.536
	llama-3.1-it	0.299	0.236	0.249	0.500	0.153	0.106	0.116	0.451
	phi3	0.217	0.178	0.191	0.500	0.155	0.111	0.120	0.464
	bert-base-multilingual (FT)	0.090	0.161	0.105	0.267				
RU	random baseline	0.033	0.006	0.010	0.036	0.071	0.027	0.027	0.050
	gemini-1.5-flash	0.409	0.314	**0.320**	0.536	0.297	0.274	**0.243**	0.553
	llama-3.1-it	0.185	0.099	0.119	0.286	0.059	0.037	0.035	0.193
	phi3	0.179	0.127	0.147	0.464	0.124	0.076	0.082	0.422
	bert-base-multilingual (FT)	0.259	0.276	0.263	0.571				
BG	random baseline	0.053	0.011	0.018	0.036	0.106	0.044	0.059	0.044
	gemini-1.5-flash	0.352	0.355	**0.327**	0.607	0.239	0.267	**0.230**	0.470
	llama-3.1-it	0.205	0.245	0.211	0.500	0.064	0.067	0.052	0.374
	phi3	0.245	0.220	0.225	0.393	0.120	0.097	0.088	0.319
	bert-base-multilingual (FT)	0.168	0.267	0.195	0.536				

Table 7: Results for the narrative inference task concerning accuracy (Acc) and macro precision (P_macro), recall (R_macro) and F1-score (F1_macro) for the generative and the fine-tuned (FT) baselines on the test and on the full datasets, respectively.

6 Conclusion and Future Work

We introduced NarratEX, a new multilingual dataset designed to facilitate research on dominant narrative and sub-narrative classification and explanation in several European languages. The dataset uniquely bridges the gap between narrative categorization and explanatory reasoning. Our analysis highlights the dataset's potential to support a range of tasks, from narrative explanation generation to category inference from justifications, with baseline results using state-of-the-art LLMs. Our experiments highlighted the dataset's value as a challenging, yet important resource for advancing the detection, the explanation, and the mitigation of manipulative textual content in diverse linguistic and cultural contexts. We hope that it will pave the way for future research on narrative analysis by fostering a deeper understanding of how narratives are framed and disseminated, thus contributing to the broader goal of promoting a more informed digital society.

In future work, we plan to analyze the narrative elements (the participants, the events, the time expressions, and the relationship between them) in the news articles and in the justifications, with the aim to assess how these structures relate to specific narratives and explanations. Additionally, we want to investigate the annotated evidence used to construct these explanations and to conduct extensive analysis and experiments on this data, particularly focusing on generating explanations using these textual segments as inputs.

7 Ethics and Broader Impact

Intended Use and Misuse Potential Our NarratEX dataset was developed to advance research on narrative classification, explanation generation, and the detection of manipulative content across multiple languages and topics. It aims to support researchers in understanding how narratives influence public discourse and supports the development of explainable models for detecting manipulative discourse and misinformation. The dataset is made available for research purposes upon acceptance of the associated terms and conditions. Nevertheless, misuse risks exist, including refining manipulative techniques or amplifying biased narratives. To mitigate these risks, we urge responsible use of the data.

Usage of AI AI was utilized for both code assistance and writing support. In coding, we used an AI assistant to generate small subroutines. For writing, we used an AI assistant to correct some minor grammatical errors and to simplify phrasing.

Environmental Impact LLMs usage demands significant computational power, leading to increase in CO_2 emissions. While we used LLMs in a zero-shot in-context learning setting rather than training them from scratch, their inference still depends on GPUs. We used the Google Cloud Platform to run the experiments presented in this work, incurring approximately 10 GPU hours, including all iterations and prompt refinements.

Fairness Most of our annotators and curators come from the institutions of the co-authors of this manuscript and were fairly paid as part of their job duties. Few annotators were experienced analysts with full-time consulting roles and rates set by their contracting institutions. A fraction of the annotators were students from the respective academic organizations. For two languages, a professional annotation company was contracted on rates based on the country of their residence. At the same time, some of the remaining annotators were researchers working primarily as linguists and lexicographers at their institute of affiliation and were all compensated according to local standards and accoridng to their employment contracts.

8 Limitations

Corpus: Our dataset focuses on two topics (the *Ukraine–Russia War* and *Climate Change*), and covers news articles in four languages (Bulgarian, English, Portuguese, and Russian). The articles selected for each language do not intend to be representative of the news coverage of these topics in the different countries since these news articles go through a multi-step ranking and filtering process to fit the purpose of the dataset. The dataset is also not balanced in terms of opposite views at the narrative level (i.e., the number of articles that present a narrative that is pro-Russian is not expected to be balanced with the ones that present a narrative pro-Ukraine). Moreover, the annotators and the curators can unknowingly carry their personal bias in the annotation of the dominant narrative and sub-narrative and the writing of the explanation. Although we provide detailed annotation guidelines and implement quality control measures, some degree of subjectivity may still persist, thus impacting the quality of the annotations.

Languages: Another factor which introduces certain bias in the created corpus is the choice of languages. For instance, in the context of the Ukraine–Russia war, the motivation behind choosing English and Russian is obvious, the latter being spoken in both countries of the conflict, whereas the main drive behind adding to the pool of languages Bulgarian and Portuguese[5] was due to the fact that Former Soviet bloc countries (such as Bulgaria) and Latin American countries (such as Brazil), are primary targets of Pro-Kremlin propaganda.

[5]Portuguese is the 7th most spoken language world-wide with circa 267 mln speakers.

While this specific choice of languages does cover significant part of the globe that is heavily exposed to propaganda (from both sides of the conflict), inclusion of other languages, e.g., Ukrainian, Spanish, etc. would obviously result in a yet higher degree of representativeness of the existing narratives. Nevertheless, although we strove to choose languages in such a way to "optimize" the representativeness of the dataset, i.e., narratives covered (which is a difficult task per se), our primary goal was to create a dataset to foster research in computational linguistics and natural language processing on methods for automated manipulation attempt detection and analysis thereof.

Baseline Models: In our experiments, we used various state-of-the-art pre-trained language models, including commercial and open-source options. For commercial models, we used two LLMs from the Gemini family, which may evolve or be deprecated over time, potentially hindering the replicability of our result. Nevertheless, we believe that this baseline remains valuable within the context of our work.

Evaluation: BERTScore has limitations that could potentially impact the previously discussed conclusions. In fact, being an embedding-based score, it is capable of capturing the context and the meaning of words. However, it can also fail to understand domain-specific wording and, in the case of multi-lingual BERT models (such as the one we used), be biased towards a particular language. These limitations should be taken into account when interpreting our results. In addition, human evaluation of the baseline results was solely conducted in the English examples of the test set. This limits a more overall evaluation of how well the model generated explanations compared to explanations written by humans. Thus, readers should take this limitation into account when interpreting the results and should not generalize to all remaining languages of the dataset. A more comprehensible and multilingual approach should be explored in future work.

Acknowledgments

This work was realised within the scope of the project CitiLink, with reference 2024.07509.IACDC, which is co-funded by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension

of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, measure RE-C05-i08.M04 - "To support the launch of a programme of R&D projects geared towards the development and implementation of advanced cybersecurity, artificial intelligence and data science systems in public administration, as well as a scientific training programme," as part of the funding contract signed between the Recovering Portugal Mission Structure (EMRP) and the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology), as intermediary beneficiary.[6]

This work is also funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2023.[7]

Nuno Guimarães, Purificação Silvano, Ricardo Campos and Alípio Jorge would also like to acknowledge project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC) and the Advanced Computing Project CPCAIAC/AV/594794/2023.[8]

Dimitar Iliyanov Dimitrov's work was funded by the EU NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008.

Giovanni Da San Martino would like to thank the Qatar National Research Fund, part of Qatar Research Development and Innovation Council (QRDI), for funding this work under grant NPRP14C0916-210015. He also would like to thank the European Union under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of March 15, 2022 of Italian Ministry of University and Research – NextGenerationEU; Code PE00000014, Concession Decree No. 1556 of October 11, 2022 CUP D43C22003050001, Progetto "SEcurity and RIghts in the CyberSpace (SERICS)" - Spoke 2 Misinformation and Fakes - DEcision supporT systEm foR cybeR intelligENCE (Deterrence) for also funding this work.

[6] https://doi.org/10.54499/2024.07509.IACDC

[7] https://doi.org/10.54499/UID/50014/2023

[8] https://doi.org/10.54499/CPCAIAC/AV/594794/2023

References

Marah Abdin et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Linguistic cues to deception: Identifying political trolls on social media. volume 13 of *ICWSM '19'*, pages 15–25, Munich, Germany.

Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, Vinay Setty, Megha Sundriyal, Konstantin Todorov, and V Venktesh. 2025. Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)*.

Samy Amanatullah, Serena Balani, Angela Fraioli, Stephanie LeMasters, and Mike Gordon. 2023. Tell us how you really feel: Analyzing Pro-Kremlin propaganda devices & narratives to identify sentiment implications. In *Illiberalism Studies Program Working Papers*, volume 14.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: natural language inference with natural language explanations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 9560–9572, Montréal, Canada. Curran Associates Inc.

Ricardo Campos, Alípio Jorge, Adam Jatowt, Sumit Bhatia, and Marina Litvak. 2024. Proceedings of the 7th international workshop on narrative extraction from texts: Text2Story 2024. In *Advances in Information Retrieval*, pages 391–397, Cham. Springer Nature Switzerland.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

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

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval@COLING '2020, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Aaron Grattafiori et al. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*, volume 1 of *NIPS'15*, pages 1693—-1701, Montreal, Canada. MIT Press.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '2019, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, USA. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

R. Likert. 1932. *A technique for the measurement of attitudes*. Archives of Psychology, Nova Iorque.

David Martens, James Hinns, Camille Dams, Mark Vergouwen, and Theodoros Evgeniou. 2025. Tell me a story! Narrative-driven XAI with large language models. *Decis. Support Syst.*, 191(C).

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021a. COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '2021, pages 997–1009, Held Online. INCOMA Ltd.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021b. A second pandemic? Analysis of fake news about COVID-19 vaccines in Qatar. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '2021, pages 1010–1021, Held Online. INCOMA Ltd.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated MAchine Reading COmprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches*, volume 1773 of *CEUR Workshop Proceedings*, Barcelona, Spain. CEUR-WS.org.

Nikolaos Nikolaidis, Nicolas Stefanovitch, Purificação Silvano, Dimitar Iliyanov Dimitrov, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ion Androutsopoulos, Preslav Nakov, Giovanni Da San Martino, and Jakub Piskorski. 2025. PolyNarrative: A multilingual, multilabel, multi-domain dataset for narrative extraction from news articles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31323–31345, Vienna, Austria.

Arian Pasquali, Marcela Canavarro, Ricardo Campos, and Alípio M. Jorge. 2016. Assessing topic discovery evaluation measures on Facebook publications of political activists in Brazil. In *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering*, C3S2E '2016, pages 25–32, Porto, Portugal. ACM.

Jakub Piskorski, Nicolas Stefanovitch, Firoj Alam, Ricardo Campos, Dimitar Dimitrov, Alípio Jorge, Senja Pollak, Nikolay Ribin, Zoran Fijavz, Maram

Hasanain, Purificação Silvano, Elisa Sartori, Nuno Guimarães, Ana Zwitter Vitez, Ana Filipa Pacheco, Ivan Koychev, Nana Yu, Preslav Nakov, and Giovanni Da San Martino. 2024. Overview of the CLEF-2024 CheckThat! lab task 3 on persuasion techniques. In *CLEF (Working Notes)*, volume 3740 of *CEUR Workshop Proceedings*, pages 299–310. CEUR-WS.org.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023a. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval '2023, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023b. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Vera Schmitt, Luis-Felipe Villa-Arenas, Nils Feldhus, Joachim Meyer, Robert P. Spang, and Sebastian Möller. 2024. The role of explainability in collaborative human-AI disinformation detection. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 2157–2174, Rio de Janeiro, Brazil. Association for Computing Machinery.

Nicolas Stefanovitch and Jakub Piskorski. 2023. Holistic inter-annotator agreement and corpus coherence estimation in a large-scale multilingual annotation campaign. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 71–86, Singapore. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *Preprint*, arXiv:2407.21783.

Chao Zhao, Faeze Brahman, Kaiqiang Song, Wenlin Yao, Dian Yu, and Snigdha Chaturvedi. 2022. NarraSum: A large-scale dataset for abstractive narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 182–197, Abu Dhabi, United Arab Emirates.

A Taxonomies

A.1 Ukraine-Russia War Taxonomy

```
Other

Blaming the war on others rather than
    the invader
- Ukraine is the aggressor
- The West are the aggressors

Discrediting Ukraine
- Rewriting Ukraine's history
- Discrediting Ukrainian nation and
    society
- Discrediting Ukrainian military
- Discrediting Ukrainian government and
    officials and policies
- Ukraine is a puppet of the West
- Ukraine is a hub for criminal
    activities
- Ukraine is associated with nazism
- Situation in Ukraine is hopeless

Russia is the Victim
- The West is russophobic
- Russia actions in Ukraine are only
    self-defence
- UA is anti-RU extremists

Praise of Russia
- Praise of Russian military might
- Praise of Russian President Vladimir
    Putin
- Russia is a guarantor of peace and
    prosperity
- Russia has international support from
    a number of countries and people
- Russian invasion has strong national
    support

Overpraising the West
- NATO will destroy Russia
- The West belongs in the right side of
    history
- The West has the strongest
    international support

Speculating war outcomes
- Russian army is collapsing
- Russian army will lose all the
    occupied territories
- Ukrainian army is collapsing

Discrediting the West, Diplomacy
- The EU is divided
- The West is weak
- The West is overreacting
- The West does not care about

Ukraine, only about its interests
- Diplomacy does/will not work
- West is tired of Ukraine

Negative Consequences for the West
- Sanctions imposed by Western countries
    will backfire
- The conflict will increase the
    Ukrainian refugee flows to Europe
```

Distrust towards Media
- Western media is an instrument of propaganda
- Ukrainian media cannot be trusted

Amplifying war-related fears
- By continuing the war we risk WWIII
- Russia will also attack other countries
- There is a real possibility that nuclear weapons will be employed
- NATO should/will directly intervene

Hidden plots by secret schemes of powerful groups

A.2 Climate Change Taxonomy

Other

Criticism of climate policies
- Climate policies are ineffective
- Climate policies have negative impact on the economy
- Climate policies are only for profit

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

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

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

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

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

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

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

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

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

B Annotation in Inception

We used the Inception platform (Klie et al., 2018) to carry out the annotation process. To that end, custom instances were established for each language in the study. First, the annotators selected the *Dominant_narrative* layer to mark the full title of the article. Subsequently, they selected a label from the list of dominant narratives and sub-narratives that best represented the claim made in the article. Next, the annotators proceeded to the *Evidence* layer, where they identified and annotated specific segments of the articles that supported their choice of dominant narrative and sub-narrative. Lastly, the annotators accessed the *Explanation* layer, which opened a text box for them to write their explanations. Figure 2 shows an example with all layers annotated.

C Details in Inter-Annotator agreement in Dominant-Narratives

In this section, we present some additional details on the inter-annotator agreement and we analyze some cases where the annotators more frequently disagree on the labels presented. Heatmaps regarding the number of agreements are presented in Figure 3 and 4 for *Climate Change* and *Ukraine–Russia War*, respectively.

Starting with the *Climate Change* documents, there was some disagreement between the dominant narratives *Criticism of climate polices* and *Criticism of institutions and authorities*, something noted in the annotation process and reflecting the issue that most discourse of policies is intertwined with the institutions proposing or enforcing them.

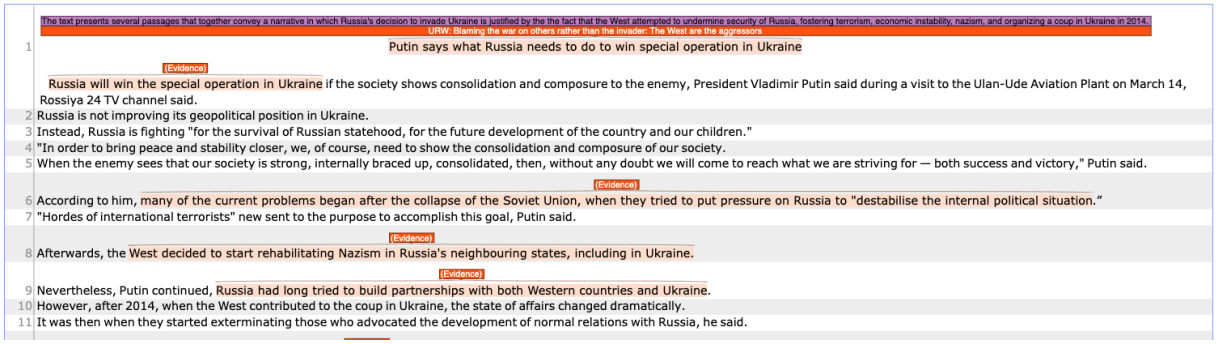

Figure 2: Annotated example from the English part of our dataset. In the title of the article, the Dominant Narrative (*Blaming the war on others rather than the invader*) and Sub-narrative (*The West as the aggressor*) are highlighted in orange, while the explanation is highlighted in purple. In the body of the article, the Evidence is highlighted in light orange.

Moreover, there were also many co-occurrences between the label *Amplifying Climate Change Fears* and the negative class *Other*. This highlights the difficulty of attributing the ulterior intentions of the author, which is to reinforce fear to the reader in this case.

Concerning the *Ukraine–Russia War* documents, the dominant narratives that we observed to be often in disagreement were *Blaming the war on others rather than the invader*, *Discrediting Ukraine* and *Discrediting the West, Diplomacy*. This stems from the fact that the argumentation present in these narratives is frequently intertwined in order to reinforce one another. For example, blaming the war on "the West" can further be justified by discrediting the intentions of the Western countries, e.g., by using the sub-Narrative *The West does not care about Ukraine, only its interests*. In practice, many disagreements and confusion were seen between the sub-Narratives *The West does not care about Ukraine, only its interests* and *Ukraine is a puppet of the West*. There were also some disagreements between *Discrediting Ukraine* and *Praise of Russia*, as it was very common to follow negative portrayals of Ukraine with overly positive ones of Russia, with the aim to highlight the moral superiority of the latter.

Finally, the label *Other* co-occurs more frequently with all other labels in both topics. This was expected since the annotators were required to use it as a negative class (*none of the above*). There, we found many cases where a (sub-)Narrative was deemed present (by both annotators), but there were different views on whether it was strong enough to include it as a dominant (sub-)Narrative or not.

	BG	EN	PT	RU
α	0.486	0.353	0.299	0.302

Table 8: Inter-annotator agreement in the sub-narratives by language, calculated using Krippendorff's α.

D Sub-Narratives Analysis

D.1 Inter-Annotator Agreement for Sub-Narratives

In this section, we present the results for the annotation agreement using Krippendorff's α in the sub-narratives. The results are presented in Table 8. The highest annotation agreement was found for Bulgarian documents, while the lowest was in the Portuguese annotations.

D.2 Number of Documents per Sub-Narrative

We present the results of the distribution of sub-narratives in the NarratEX dataset, completing the analysis made in Section 4. Table 9 presents the sub-narratives frequency concerning the *Ukraine–Russia War*, while Table 10 presents the frequency for *Climate Change*. For this analysis, we excluded the documents labeled as the *Other* sub-narrative.

In the *Ukraine–Russia War*, the most frequent sub-narratives are pro-Russian, and focus on discrediting the Ukrainian government and military, praising the Russian military, and discussing the influence of the West in the war ("*Ukraine is a puppet of the West*," and "*the West is the aggressor*"). For *Climate Change*, the most frequent sub-narrative is "*Amplifying existing fears of global warming*" with 96 documents, while "*Criticism of Political Organizations and Figures*" and "*Criticism of national governments*" hold a distant second and third place, with 18 and 17 documents, respectively.

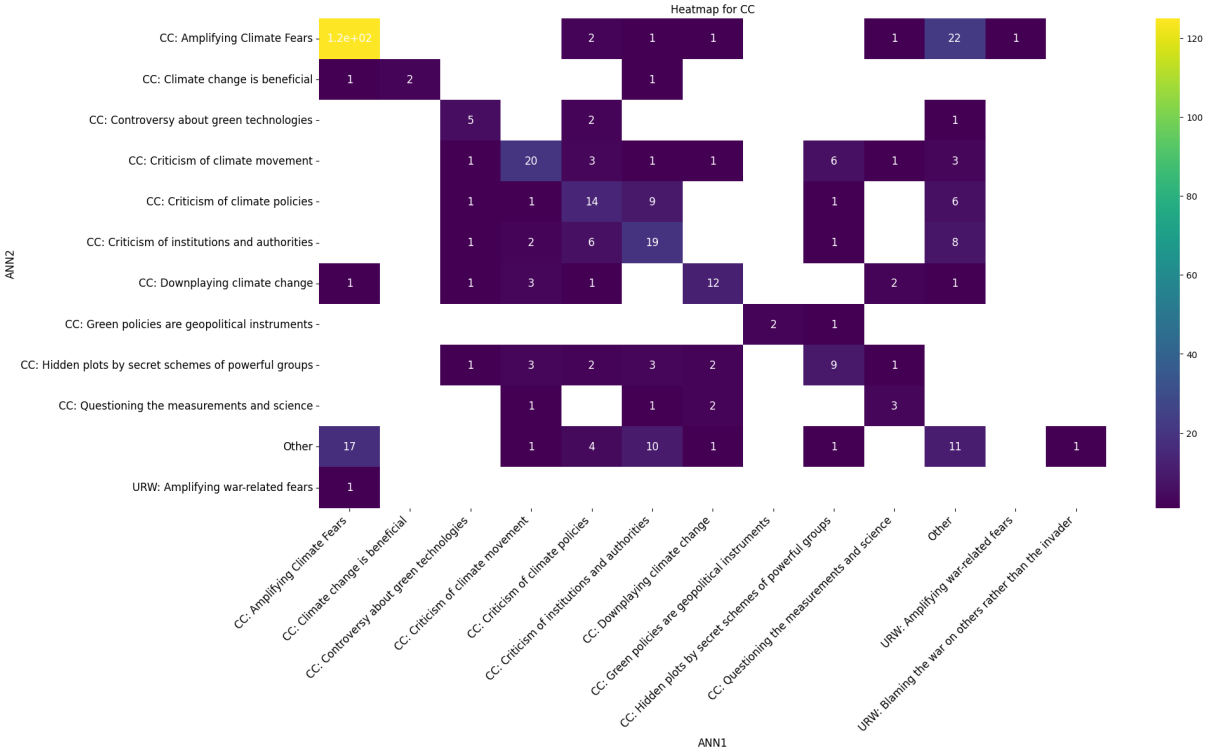

Figure 3: Co-occurrence matrix for the agreement between the dominant narratives classification in Climate Change documents.

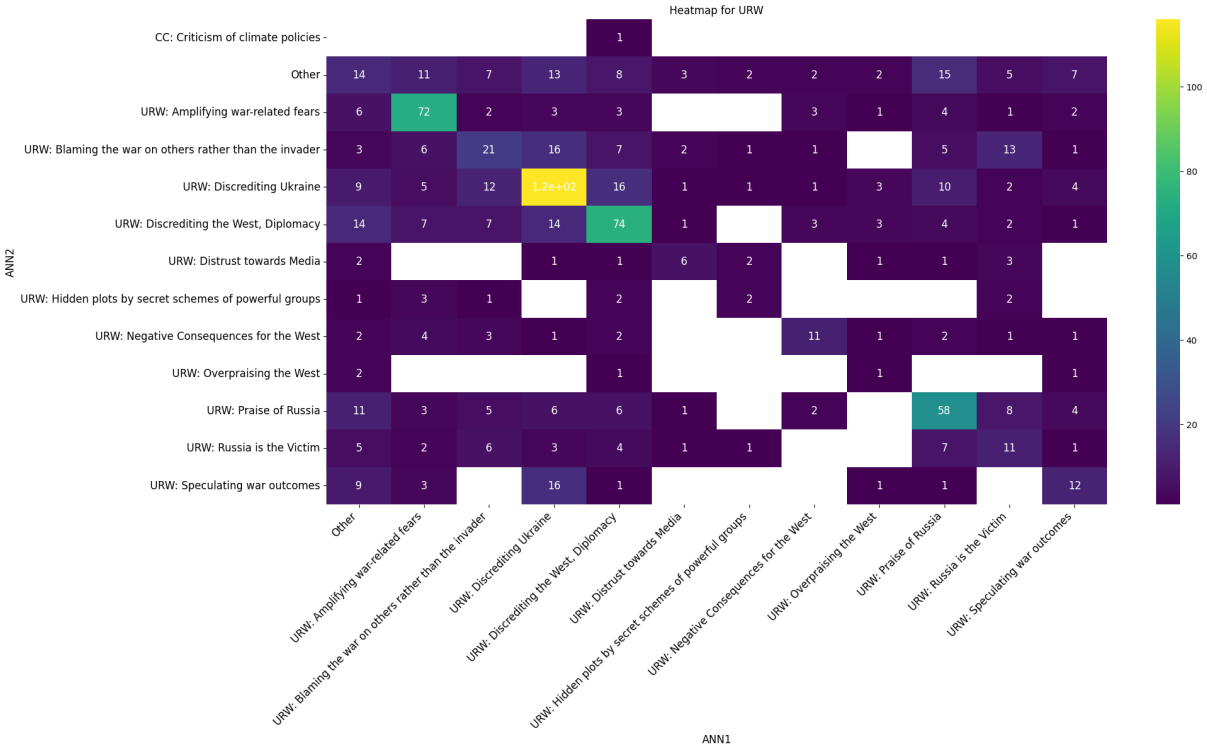

Figure 4: Co-occurrence matrix for the agreement between the dominant narratives classification in Ukraine–Russia War documents.

Topic: Ukraine-Russia War	#docs
Discrediting Ukraine: Discrediting Ukrainian government and officials and policies	72
Praise of Russia: Praise of Russian military might	44
Discrediting Ukraine: Discrediting Ukrainian military	35
Blaming the war on others rather than the invader: The West are the aggressors	32
Discrediting Ukraine: Ukraine is a puppet of the West	28
Discrediting the West, Diplomacy: The West does not care about Ukraine, only about its interests	28
Amplifying war-related fears: There is a real possibility that nuclear weapons will be employed	25
Amplifying war-related fears: By continuing the war we risk WWIII	21
Blaming the war on others rather than the invader: Ukraine is the aggressor	21
Praise of Russia: Russia has international support from a number of countries and people	19
Amplifying war-related fears: Russia will also attack other countries	18
Discrediting Ukraine: Situation in Ukraine is hopeless	18
Praise of Russia: Russia is a guarantor of peace and prosperity	18
Discrediting the West, Diplomacy: The West is weak	16
Russia is the Victim: The West is russophobic	16
Discrediting Ukraine: Ukraine is a hub for criminal activities	15
Discrediting the West, Diplomacy: The EU is divided	11
Speculating war outcomes: Russian army is collapsing	11
Distrust towards Media: Western media is an instrument of propaganda	10
Speculating war outcomes: Ukrainian army is collapsing	10
Discrediting the West, Diplomacy: Diplomacy does/will not work	9
Discrediting the West, Diplomacy: West is tired of Ukraine	9
Discrediting Ukraine: Ukraine is associated with nazism	8
Negative Consequences for the West: Sanctions imposed by Western countries will backfire	8
Praise of Russia: Praise of Russian President Vladimir Putin	6
Overpraising the West: The West belongs in the right side of history	4
Amplifying war-related fears: NATO should/will directly intervene	3
Discrediting Ukraine: Discrediting Ukrainian nation and society	3
Discrediting the West, Diplomacy: The West is overreacting	3
Russia is the Victim: Russia actions in Ukraine are only self-defence	3
Russia is the Victim: UA is anti-RU extremists	3
Discrediting Ukraine: Rewriting Ukraine's history	2

Table 9: Number of documents per sub-narrative for the Ukraine–Russia War.

We can further see in Table 9 that the first two sub-narratives against the Climate Change movement ("*Climate policies have negative impact on the economy*" and "*Climate agenda has hidden motives*") are ranked 4th and 6th with 14 and 10 documents.

E Baselines Details

In this section, we present the particular prompts we used for our experimental baselines for the two tasks we proposed, as well as the values of the hyper-parameters we used for fine-tuning the BERT model in the Dominant Narrative Inference from the Justification Task.

E.1 Prompt for the Generation of Narrative Explanation Task

Given a news article along with its dominant and sub-dominant narratives, generate a concise text (maximum 80 words) supporting these narratives without the need to explicitly mention them. The explanation should align with the language of the article and be direct and to the point. If the sub-dominant narrative is 'Other,' focus solely on supporting the dominant narrative. The response should be clear, and succinct, and avoid unnecessary elaboration.
Dominant Narrative: {**dominant narrative**}

Topic: Climate Change	#docs
Amplifying Climate Fears: Amplifying existing fears of global warming	96
Criticism of institutions and authorities: Criticism of political organizations and figures	18
Criticism of institutions and authorities: Criticism of national governments	17
Criticism of climate policies: Climate policies have negative impact on the economy	14
Criticism of institutions and authorities: Criticism of international entities	11
Hidden plots by secret schemes of powerful groups: Climate agenda has hidden motives	10
Amplifying Climate Fears: Earth will be uninhabitable soon	9
Criticism of climate movement: Ad hominem attacks on key activists	9
Hidden plots by secret schemes of powerful groups: Blaming global elites	9
Amplifying Climate Fears: Doomsday scenarios for humans	8
Criticism of climate policies: Climate policies are only for profit	5
Criticism of institutions and authorities: Criticism of the EU	5
Criticism of climate movement: Climate movement is alarmist	4
Criticism of climate policies: Climate policies are ineffective	4
Downplaying climate change: Weather suggests the trend is global cooling	4
Questioning the measurements and science: Scientific community is unreliable	4
Criticism of climate movement: Climate movement is corrupt	3
Downplaying climate change: Human activities do not impact climate change	3
Questioning the measurements and science: Methodologies/metrics used are unreliable/faulty	3
Controversy about green technologies: Renewable energy is dangerous	2
Controversy about green technologies: Renewable energy is unreliable	2
Downplaying climate change: CO2 concentrations are too small to have an impact	2
Downplaying climate change: Ice is not melting	2
Amplifying Climate Fears: Whatever we do it is already too late	1
Climate change is beneficial: CO2 is beneficial	1
Climate change is beneficial: Temperature increase is beneficial	1
Downplaying climate change: Climate cycles are natural	1
Downplaying climate change: Temperature increase does not have significant impact	1
Green policies are geopolitical instruments: Climate-related international relations are abusive/exploitative	1
Green policies are geopolitical instruments: Green activities are a form of neo-colonialism	1
Questioning the measurements and science: Greenhouse effect/carbon dioxide do not drive climate change	1

Table 10: Number of documents per sub-narrative in Climate Change.

Sub-dominant Narrative: {**sub-dominant narrative**}
Article: {**article text**}
Output:

E.2 Dominant Narrative Inference from Justification Task

E.2.1 Prompt used for the Dominant Narrative Inference from Justification Task

Goal: Identify which of the main narratives given better suits the explanation text provided. Answer only with the narrative and nothing else.
Explanation: {**explanation**}
Narratives: {**list of narratives**}
Output:

E.2.2 Fine-Tuned Hyperparameter Values for the Dominant Narrative Inference from the Justification Task

In our experiments, we used the following version of BERT: bert-base-multilingual-cased.[9] We then performed fine-tuning for three epochs with a per-device batch size of 16 during both training and evaluation. We further used a learning rate of 2e-5, which is commonly adopted in transformer fine-tuning in order to enable gradual adaptation of pre-trained weights without inducing large parameter shifts. Finally, in order to further mitigate potential overfitting and to enhance generalization, we used a weight decay coefficient of 0.01, which is again common in similar setups.

[9] https://huggingface.co/google-bert/bert-base-multilingual-cased

F Human Evaluation for the Generation of Narrative Explanation

F.1 Guidelines

The following guidelines were discussed and defined by the members of the language coordination team to support the evaluation of the model-generated explanations. An annotator, who is a master's linguistic student not part of the development or annotation of the English dataset, was tasked with evaluating how closely a language model's explanation aligned with the corresponding reference (ground truth) explanation. Each explanation was rated on a five-point Likert scale (Likert, 1932), based on the following criteria:

- **5 - Fully** The model's explanation recovers all the main arguments expressed in the ground truth. While there may be differences in wording, the justificatory content is entirely present.
- **4 - Mostly** The model's explanation conveys the same line of reasoning, but omits a secondary element or rephrases one of the main arguments from the ground truth in a less explicit way.
- **3 - Reasonably** The model's explanation captures part of the argumentation in the ground truth, overlooking one or more main elements or shifting the focus to less central aspects.
- **2 - Barely** The model's explanation fails to recover the core arguments supporting the ground truth. It may touch on secondary or tangential arguments, but remains outside the dominant line of reasoning.
- **1 - Not at all** The model's explanation provides a completely different or irrelevant justification compared to the ground truth. There is no conceptual correspondence between the arguments of the two.

F.2 Qualitative Error Analysis

In our qualitative error analysis, we focused on cases where models failed to align with the gold standard in order to better understand the nature of these divergences. Below we present three illustrative examples.

F.2.1 Gemma 2 (climate hypocrisy)

The ground truth explanation describes the article as highly critical of LA mayor Garcetti, calling him a *hypocrite* and highlighting his poor management. However, the model gives a favorable portrayal of Garcetti as a *climate leader*. This reveals a failure to capture the article's negative evaluative stance, with the model opting instead for a generic, positive summary unrelated to the actual narrative.

Ground truth explanation: *The text presents several paragraphs in which political figures and climate activists such as the LA mayor Garcetti are criticised. More specifically, the text labels Garcetti as an hypocrite polluting with his flights to climate conferences. Moreover, Garcetti is being criticised for the bad management of the city of LA.*

Gemma 2 explanation: *Los Angeles Mayor Garcetti's participation in a climate summit highlights the city's commitment to addressing the climate crisis. He emphasized the importance of global collaboration and the implementation of concrete solutions at the local level. Despite facing challenges such as high crime and homelessness, the city is actively pursuing a transition to clean energy sources and climate justice initiatives.*

F.2.2 Gemini-1.5-flash (Al Gore's rhetoric)

While the ground truth explanation emphasizes the article's ridicule of fear-inducing expressions used by Al Gore and other climate activists, Gemini15-flash instead presents Gore's statements in a matter-of-fact tone, even echoing their urgency and proposals, resulting in a mischaracterization of the article's narrative intent.

Ground truth explanation: *The article talks about Al Gore and other climate activists who use "rain bombs" and "boiling oceans" to scare people. The article emphasizes absurdity of such actions.*

Gemini-1.5-flash explanation: *Al Gore warns of a "planetary crisis" caused by "greenhouse gas pollution" and demands drastic action, including opening borders to climate refugees and replacing fossil fuels with "renewable energy," which he claims is the cheapest source of electricity.*

F.2.3 LLaMA 3.1-it (nuclear war discourse)

Here, the ground truth explanation points to a reporting tone and a chronological listing of Russian propaganda tactics, despite a fear-inducing headline. The model explanation, however, includes exaggerated pro-Kremlin rhetoric and speculative political outcomes, distorting the narrative.

Ground truth explanation: *Even though the title of the article amplifies fears on nuclear use, the article itself lists chronologically propaganda techniques used by Russian media to persuade people that Russia is winning the war and scare people with nuclear weapons. However, most of the article has reporting character.*

Llama 3.1-it explanation: *Russian state television insists that the Kremlin's "special operation" in Ukraine is progressing as planned, with President Vladimir Putin's willingness to negotiate with Ukraine seen as a demonstration of his love and respect for the Ukrainian people. The West's reluctance to intervene is presented as proof of Russia's ability to emerge victorious, with pundits speculating about the concessions Moscow will receive, including the lifting of sanctions, dissolution of NATO, and extradition of war criminals.*

These examples suggest that even in the best-performing language (English), models often fail to accurately reflect the intended stance and argumentation of the original articles. They may struggle with discursive nuance, particularly when the narrative involves irony, criticism, or propaganda techniques.

F.3 Divergence Between Automatic Metrics and Human Evaluation

Building on the qualitative examples in Subsection F.2, we now examine systematic discrepancies between BERTScore and human evaluation. When contrasting them, three systematic patterns of discrepancy emerge across the models. These patterns highlight the limits of relying solely on automatic metrics and reinforce the need for human evaluation for this task.

F.3.1 Overestimation by BERTScore

In several cases, the models produced explanations that completely inverted the polarity of the central arguments of the ground-truth, yet still obtained relatively high BERTScores. For instance, in the Garcetti's climate hypocrisy example (Gemma 2; see subsubsection F.2.1), the ground truth highlights criticism of the mayor's hypocrisy and poor management, whereas the model generated a favorable portrayal of him as a climate leader. Similarly, in Al Gore's rhetoric example (Phi-3, Gemini-1.5-flash; see below and subsubsection F.2.2), the ground truth emphasizes the exaggerated and alarmist nature of expressions such as *rain bombs* and *boiling oceans*.

At the same time, the language models presented these claims in a descriptive way, reporting them without reflecting the critical stance that the human-annotated ground truth had identified. Another case is the nuclear war discourse (LLaMA 3.1-it; see subsubsection F.2.3), where the ground truth points to a chronological listing of propaganda techniques, but, on the other hand, the model generated an embellished pro-Kremlin narrative. In all these cases, the inflated BERTScore metrics may indicate that the model relied on the same topical vocabulary as the ground truth (e.g., *climate*, *Ukraine*, *Russia*), whereas the human evaluation penalized the failure to recover the core arguments.

Example "Al Gore's Rhetoric"

Phi-3 explanation: *Al Gore emphasizes the urgent need to address the climate crisis, attributing severe weather events and rising sea levels to greenhouse gas emissions. He advocates for a transition away from fossil fuels, highlighting the Inflation Reduction Act's investment in renewable energy as a crucial step towards a sustainable future. Gore warns that without significant action, the influx of climate refugees could overwhelm national borders and threaten global stability.*

F.3.2 Underestimation by BERTScore

In other cases, the language models managed to recover the full argumentative structure of the ground truth, but scored relatively lower on BERTScore. For example, in the climate policies as profit tool explanations (Phi3, Gemini-1.5-pro; see below), the models succeeded to clearly capture the argument that green policies were instrumentalized for profit by political and corporate actors, even if phrased differently. The ground truth emphasized Democrats and green technology companies as key beneficiaries, while the models described state mandates and corporate interests bypassing local authority and exploiting taxpayer funds. Despite these differences in lexical choice and framing, the human annotators rated these explanations as fully aligned (Likert = 5). At the same time, BERTScore remained in the 0.66–0.67 range, possibly reflecting some limitations to paraphrastic variation or lexical divergence. In this case, the automatic metric underestimated the true adequacy of the model outputs.

Example "Climate policies as profit tool"

Ground truth explanation: *The text has several paragraphs arguing that climate policies are only an instrument in the hands of Democrats in order to make profit together with companies involved in the business of green technologies.*

Phi-3 explanation: *State-level bureaucrats in Michigan, Illinois, New York, and California are bypassing local governments to push large-scale green energy projects, often at the expense of rural residents and family farmers. These projects, driven by corporate interests and the promise of billions in profits, are undermining local authority and land rights, with the goal of achieving ambitious zero-carbon energy targets. The Inflation Reduction Act and other legislation are seen as facilitating this corporate-driven agenda, raising concerns about the true motives behind the green.*

Gemini-1.5-pro explanation: *State mandates are stripping local communities of their authority to reject renewable energy projects. Driven by corporate interests and fuelled by taxpayer funds, these policies prioritize large-scale ventures over the rights of rural landowners. This push for "green" initiatives bypasses local regulations, undermining community control. It's enriching corporations and benefiting foreign entities, all while claiming to fight climate change.*

F.3.3 Partial Overlap

A third group of cases sits in between. In the EU division example (LLaMA 3.1-it; see below), the model recovered the idea of internal disagreements among European institutions, but it shifted the focus by framing the debate mainly around accusations of arms profiteering, leaving aside the ground truth's emphasis on the EU's incapacity to help Ukraine effectively and the claim that only USA should provide support. Similarly, in the example of US mineral interests (Gemini-1.5-flash; see below), the model successfully identified the argument that Western support was motivated by Ukraine's mineral resources, but it downplayed the colonialist framing of this exploitation that was central to the ground truth. Similarly, in the housing costs and climate policy example (Gemini-1.5-pro), the model emphasized the rising costs, but ignored the critique of inefficiency in green technologies. These explanations received mid-range Likert scores (2–3), while BERTScore remained above 0.70.

This illustrates that while lexical overlap exists, human evaluators distinguish between tangential and central arguments, something BERTScore seemingly does not consider.

Example "EU Division"

Ground truth explanation: *The text frames EU as being divided and incapable of providing actual help to Ukraine. Through one-sided reporting the text claims that the EU could have stopped the conflict and presents only the view that support to Ukraine should be cut and left to the US.*

Llama 3.1-it explanation: *European Parliament members are at odds over continued support for Ukraine, with some accusing proponents of military aid of prioritizing arms company profits over EU citizens' interests. Critics argue that Western actions have prolonged the conflict, and that the EU is now on a collision course with Russia, making it imperative to reassess its strategy and pursue peace negotiations without external interference.*

Example "US Mineral Interests"

Ground truth explanation: *The text argues that the US is standing on Ukraine side only because it has interests in Ukraine's mineral resources. In doing so, the text claims that this is a form of colonialism, and that the US is exploiting the country.*

Gemini-1.5-flash explanation: *The West's support for Ukraine is driven by its strategic interest in securing access to Ukraine's vast mineral resources, which are essential for its economy and technological advancement. The potential for a strong economic partnership between Ukraine and the West, fuelled by these resources, highlights the importance of preventing a Russian victory.*

These discrepancies show that BERTScore is effective at capturing lexical and semantic proximity, but struggles regarding argumentative stance, claims detection, selective emphasis, and evaluative framing. Thus, further experiments, which are out of the scope of the work, would be necessary to confirm this hypothesis. Human evaluation, therefore, remains indispensable for tasks where the correct interpretation of narrative intent and argumentative justification is fundamental.

G NarratEX Dataset Examples

We complement the example in Figure 2 with complete examples of the dataset and the respective annotations in all languages.

G.1 Bulgarian

• *Dominant narrative and Sub-narrative:*

Amplifying Climate Fears; Amplifying existing fears of global warming

• *Original Document:*

Замръзналите недра на Земята може да крият опасна катаклизмична тайна

Във вечното замръзналата почва на Земята може да се крие нещо голямо.

Тъй като планетата продължава да се затопля, учените се опасяват, че множество смъртоносни болести ще бъдат отприщени от замръзналата земя, след като са останали в спящо състояние в продължение на десетилетия, векове и дори хилядолетия. Освен това, разширяването на минното дело в полярните региони може да ни приближи още повече до отварянето на тази кутия на Пандора.

Най-тревожното от всичко обаче е, че според учените нашите действия ни приближават към непозната, древна заплаха, наречена "Фактор X".

Терминът "вечно замръзнала почва" описва земята, която е била замръзнала в продължение на две или повече последователни години. Две години е минимумът, а някои райони в Сибир са замръзнали за повече от 650 000 години.

Всеки метър от тази замръзнала почва кипи от живот, като в един грам има стотици хиляди спящи микробни видове. Действителната идентичност на тези микроби обаче до голяма степен е загадка.

"Има много неща, които не знаем, и много малко хора са изследвали вечно замръзналите райони", казва пред Newsweek Биргита Евенгард, професор по инфекциозни болести в университета в Умео, Швеция.

През 2014 г. група френски и руски изследователи реактивираха гигантски вирус, който е лежал в спящо състояние под сибирската тундра в продължение на 30 000 години. Сега този конкретен вирус – известен като пандоравирус – заразява само амеби. Той не представлява заплаха за хората. Въпреки това, това проучване предоставя доказателство за концепцията.

"Ако вирусите на амебата могат да оцелеят толкова дълго във вечен мраз, това силно подсказва, че тези, които заразяват животни/човек, могат да останат заразни при същите условия", казва пред Newsweek Жан-Мишел Клавери, който е ръководител на изследването. "Освен това знаем, че ДНК на заразяващи животни/човек вируси се открива в тези места."

Други изследвания показват, че дори микроскопични животни могат да бъдат възкресени от замръзналите недра.

"Съществуват различни методи, включително фиксиране на тяхната ДНК и липидни мембрани, [които позволяват на организмите да оцелеят във вечната замръзналост,]" Кимбърли Майнър, климатолог от Лабораторията за реактивни двигатели на НАСА в Калифорния и професор в Института за климатични промени, казва пред Newsweek. "Това важи за редица микроби, които се считат за екстремофили – организми, които могат да оцелеят при екстремни температури и налягане, включително студа и налягането на вечната замръзналост."

И така, какво всъщност може да има там долу?

"Вируси от изчезнали болести като дребна шарка; винаги присъстващи-

ят антракс, чрез замърсени със спори зони; а също и ускореното разпространение на вече известни болести, които съществуват в днешна Арктика, като туларемия, сериозна бактериална инфекция, или кърлежов енцефалит", казва Клавери.

През 2016 г. огнище на антракс в Северен Сибир уби 12-годишно момче и хиляди животни. Смята се, че причината за това е необичайно топлото време в региона, което ускорява размразяването на замръзнала почва и разкрива трупа на северен елен, който се е поддал на инфекцията. Спящите спори на антракса в трупа на елена са се събудили и са се освободили, за да намерят нови гостоприемници.

Тези известни инфекции най-вероятно се намират в най-горните слоеве, но това, което се крие по-дълбоко, е още по-обезпокоително.

"Дълбоко във вечно замръзналите недра трябва да има микроби – особено вируси, но също и бактерии – които са били на Земята много преди появата на Homo sapiens", казва Евенгард.

Имунната ни система е еволюирала в контакт с трилиони микроби, съществували на Земята през живота на нашия вид. Възможно е обаче под снега и леда да има древни вируси, срещу които нямаме естествен имунитет, нито ефективни ваксини или лечения.

"Съществува Фактор Х, за който наистина не знаем много", казва Евенгард.

Всъщност тези праисторически патогени може да са допринесли за гибелта на древните ни предци. "Тези древни вируси може да са заразили неандерталски хора или мамути, причинявайки тяхното изчезване", казва Клавери.

През последните 50 години Арктика се затопля до четири пъти по-бързо от останалата част на света, а средната температура на вечната замръзналост се увеличава с около минус 17,4 градуса по Целзий на десетилетие, според Агенцията за опазване на околната среда на САЩ, което означава, че бъдещето може да крие още една неподозирана опасност, скрита дълбоко под леда.

• *Original Explanation:*

Размразяването на ледниците, причинено от глобалното затопляне може да доведе до отприщването на древни вируси, криещи се под леда.

• *Translated Document*

Earth's Frozen Ground May Hold a Cataclysmic Secret

Something big could be lurking in Earth's permafrost.

As the planet continues to warm, scientists fear that a host of deadly diseases will be unleashed from the frozen ground after lying dormant for decades, centuries, and even millennia. In addition, the expansion of mining in polar regions could bring us even closer to opening this Pandora's Box.

Most worrying of all, scientists say our actions are bringing us closer to an unknown, ancient threat called "Factor X."

The term "permafrost" describes land that has been frozen for two or more consecutive years. Two years is the minimum, and some areas in Siberia have been frozen for more than 650,000 years.

Every meter of this frozen soil is teeming with life, with hundreds of thousands of dormant microbial species per gram. The actual identities of these microbes, however, are largely a mystery.

"There's a lot we don't know, and very few people have studied permafrost," Birgitta Evengaard, a professor of infectious diseases at Umeå University in Sweden, told Newsweek.

In 2014, a team of French and Russian researchers reactivated a giant virus that

had lain dormant beneath the Siberian tundra for 30,000 years. Now, this particular virus—known as Pandoravirus—infects only amoebae. It poses no threat to humans. However, this study provides proof of concept.

"If amoeba viruses can survive this long in permafrost, it strongly suggests that those that infect animals/humans can remain infectious under the same conditions," Jean-Michel Claverie, who led the study, told Newsweek. "We also know that DNA from animal/human viruses is found in these places."

Other studies have shown that even microscopic animals can be resurrected from the frozen depths.

"There are various methods, including fixing their DNA and lipid membranes, [which allow organisms to survive in permafrost,]" Kimberly Miner, a climatologist at NASA's Jet Propulsion Laboratory in California and a professor at the Climate Change Institute, told Newsweek. "This is true for a number of microbes that are considered extremophiles—organisms that can survive extreme temperatures and pressures, including the cold and pressure of permafrost."

So what could be down there?

"Viruses from extinct diseases like smallpox; the ever-present anthrax, via spore-contaminated areas; and also the accelerated spread of already known diseases that exist in today's Arctic, such as tularemia, a serious bacterial infection, or tick-borne encephalitis," says Claverie.

In 2016, an outbreak of anthrax in northern Siberia killed a 12-year-old boy and thousands of animals. The cause is thought to be unusually warm weather in the region, which accelerated the thawing of frozen ground and exposed the carcass of a reindeer that had succumbed to the infection. Dormant anthrax spores in the reindeer's carcass have awakened and broken free to find new hosts.

These known infections are likely to be found in the upper layers, but what lies deeper is even more troubling.

"Deep in the permafrost, there must be microbes—especially viruses, but also bacteria—that were on Earth long before Homo sapiens," says Evengard.

Our immune systems evolved in contact with trillions of microbes that have existed on Earth during the lifetime of our species. But it's possible that ancient viruses lurk beneath the snow and ice for which we have no natural immunity, no effective vaccines, or treatments.

"There's an X-factor that we really don't know much about," says Evengard.

In fact, these prehistoric pathogens may have contributed to the demise of our ancient ancestors. "These ancient viruses may have infected Neanderthals or mammoths, causing their extinction," says Claverie.

Over the past 50 years, the Arctic has been warming up to four times faster than the rest of the world, and the average temperature of permafrost is increasing by about minus 17.4 degrees Celsius per decade, according to the US Environmental Protection Agency, meaning the future may hold another unsuspected danger hidden deep beneath the ice.

- *Translated Explanation*

  Melting glaciers caused by global warming could unleash ancient viruses hiding under the ice.

G.2 Portuguese

- *Dominant narrative and Sub-narrative:*

  Amplifying war-related fears; There is a real possibility that nuclear weapons will be employed

- *Original Document:*

  Objetivo do treino da NATO é uma guerra nuclear na Europa, diz analista russo

  O objetivo dos exercícios nucleares da NATO é a preparação para uma guerra nuclear limitada na Europa, com ataques de bombas aéreas nucleares táticas dos

EUA por aviões das forças aéreas nacionais da NATO, em alvos na Rússia, disse à Sputnik Igor Korotchenko, editor-chefe da revista russa Natsionalnaya Oborona (Defesa Nacional).

O secretário-geral da NATO, Mark Rutte, anunciou na quinta-feira (10) que os exercícios nucleares anuais da aliança, Steadfast Noon, terão início na segunda-feira (14). As manobras decorrerão sobretudo no Reino Unido, Mar do Norte, Bélgica e Países Baixos.

"Os exercícios envolvem a prática de procedimentos relacionados com a transferência de instalações de armazenamento para a força aérea de vários Estados-membros da NATO de bombas nucleares táticas americanas B61-13, armazenadas em aeródromos de vários países da NATO e sob controlo americano. Na simulação de um conflito nuclear com a Rússia e de uma guerra nuclear limitada na Europa, as bombas aéreas são transferidas para as forças aéreas nacionais dos países membros da NATO envolvidos", explicou Korotchenko.

Segundo o mesmo, os exercícios incluirão a aquisição de tarefas de combate e planos de ataque contra instalações no território russo ao longo de toda a sua fronteira com os países da NATO.

"Todos estes procedimentos têm um objetivo – uma guerra nuclear limitada na Europa, com a utilização de ataques reais utilizando armas nucleares táticas dos EUA por aviões das forças aéreas nacionais dos países da NATO, em alvos no território da Rússia", disse Korotchenko.

Ao mesmo tempo, o analista acredita que, no caso de um ataque real com armas nucleares táticas dos EUA no território da Rússia, mesmo às mãos dos aliados europeus dos EUA, deve ser realizado um ataque nuclear limitado no território dos EUA.

- *Original Explanation:*

O texto, referenciando Igor Korotchenko, apresenta a possibilidade de os exercícios nucleares anuais da NATO serem, na verdade, um prelúdio de uma guerra nuclear contra a Rússia, a ser instituída na Europa, com os EUA como incitadores táticos. Korotchenko afirma, ainda, que "um ataque real com armas nucleares dos EUA no território da Rússia" terá uma resposta, na mesma escala, por parte do governo russo. The text, referencing Igor Korotchenko, presents the possibility that NATO's annual nuclear exercises are in fact a prelude to a nuclear war against Russia, to be waged in Europe, with the US as the tactical inciter. Korotchenko also claims that "a real US nuclear weapons attack on Russian territory" will be met with a response on the same scale from the Russian government.

- *Translated Document*

Aim of Nato training is nuclear war in Europe, says Russian analyst

The goal of NATO's nuclear exercises is preparation for a limited nuclear war in Europe, with US tactical nuclear air bomb strikes by NATO national air forces' aircraft on targets in Russia, Igor Korotchenko, editor-in-chief of the Russian magazine Natsionalnaya Oborona (National Defense), told Sputnik.

NATO Secretary General Mark Rutte announced on Thursday (10) that the alliance's annual nuclear exercises, Steadfast Noon, will begin on Monday (14). The maneuvers will take place mainly in the United Kingdom, the North Sea, Belgium and the Netherlands.

"The exercises involve practicing procedures related to the transfer from storage facilities to the air force of several NATO member states of American B61-13 tactical nuclear bombs, stored at airfields in several NATO countries and under American control. In the simulation of a nuclear conflict with Russia and a limited nuclear war in Europe, the aerial bombs are transferred to the national air forces of the NATO member states involved," Korotchenko explained.

According to him, the exercises will include the acquisition of combat tasks

and attack plans against facilities on Russian territory along its entire border with NATO countries.

"All these procedures have one goal - a limited nuclear war in Europe, with the use of real strikes using US tactical nuclear weapons by aircraft of the national air forces of NATO countries on targets on the territory of Russia," said Korotchenko.

At the same time, the analyst believes that in the event of a real strike with US tactical nuclear weapons on Russia's territory, even at the hands of the US's European allies, a limited nuclear strike should be carried out on US territory.

• *Translated Explanation*

The text, referencing Igor Korotchenko, presents the possibility that NATO's annual nuclear exercises are, in fact, a prelude to a nuclear war against Russia, to be waged in Europe, with the US as the tactical inciter. Korotchenko also claims that "a real US nuclear weapons attack on Russian territory" will be met with a response on the same scale from the Russian government.

G.3 Russian

• *Dominant narrative and Sub-narrative:*

URW: Praise of Russia: Praise of Russian military might

• *Original Document:*

Дмитрий Рогозин: "Харьковское наступление наших войск сильно взбаламутило украинское военное начальство, которое спешно стало перебрасывать резервы на это направление."

Харьковское наступление наших войск сильно взбаламутило украинское военное начальство, которое спешно стало перебрасывать резервы на это направление. По моей оценке, только с нашего Запорожского фронта туда уехало не менее трети батальонов ВСУ. Как итог провисающего украинского фронта—он стал двигаться под натиском 58-й гвардейской армии и подразделений ВДВ, пусть не так радикально, но всё же.

Продвижение есть, взяли и, что самое важное, удержали несколько опорных пунктов, освободили десятки км, расширили свою территорию на километры в глубину и по фронту. Бои на нашем направлении идут круглосуточно. Досаждают нам БпЛА противника, их на два порядка больше, но, как говорится, "власть развращает, а абсолютная власть развращает абсолютно". Численное превосходство противника в воздухе привело к его самонадеянности, что дало шанс наших штурмовым группам сблизиться с врагом на расстояние стрелковых боев. А вот тут русским равных нет. С той стороны тоже многих русских по происхождению, но власовцы - уже не русские, бандеровцы - тем более. В итоге все стрелковые бои за последнюю неделю наши выиграли в сухую. Имел возможность наблюдать за действиями наших ребят. Ну просто молодцы! Гордость берет.

Активизировалось и бандитское подполье в Запорожской области. Оно и понятно: отсутствие военных успехов заставляет противника палить агентуру. Только с середины апреля нашими чекистами, военной контрразведкой и силами военной полиции было выявлено более 110 схронов оружия и боеприпасов, обезврежено несколько террористических ячеек, совершавших нападения и подрывы наших военнослужащих и русских активистов. Тем не менее, надо признать, что еще год назад ситуация в Мелитополе, Бердянске и в целом в районах зоны территориальной обороны была намного сложнее. И это снижение активности и количества "спящих"—прямое свидетельство высококвалифицированной работы ФСБ и Воен-

ной комендатуры Запорожской области.

- *Original Explanation:*

  У России огромные успехи на фронте и чистые победы, у украинцев же нет побед и им нужна помощь американцев. Россия также успешно борется с терроризмом и бандитами из Украины.

- *Translated Document*

  Dmitry Rogozin: "The Kharkov offensive of our troops has greatly stirred up the Ukrainian military command, which has been hastily transferring reserves to this direction."

  The Kharkov offensive of our troops has greatly stirred up the Ukrainian military command, which has been hastily transferring reserves to this direction. In my opinion, at least a third of the Ukrainian Armed Forces battalions have left there from our Zaporizhzhya Front alone. As a result of the sagging Ukrainian Front, it has begun to move under the pressure of the 58th Guards Army and Airborne Forces units, albeit not as radically, but still.

  There is progress, we have taken and, most importantly, held several strongholds, liberated tens of kilometers, expanded our territory by kilometers in depth and along the front. The fighting in our direction is ongoing around the clock. We are being harassed by enemy UAVs, there are two orders of magnitude more of them, but, as they say, "power corrupts, and absolute power corrupts absolutely." The enemy's numerical superiority in the air led to his arrogance, which gave our assault groups a chance to get close to the enemy at a range of small arms combat. But here the Russians have no equal. On that side, too, there are many Russians by origin, but the Vlasovites are no longer Russian, and the Banderites are even less so. As a result, our guys won all the small arms combats over the past week without losing a single blow. I had the opportunity to observe the actions of our guys. Well done! I am proud.

  The bandit underground in the Zaporozhye region has also become more active. This is understandable: the lack of military success forces the enemy to expose their agents. Since mid-April alone, our security officers, military counterintelligence and military police have discovered more than 110 caches of weapons and ammunition, and neutralized several terrorist cells that had been carrying out attacks and blowing up our servicemen and Russian activists. Nevertheless, it must be acknowledged that a year ago the situation in Melitopol, Berdyansk and in general in the areas of the territorial defense zone was much more complicated. And this decrease in activity and the number of "sleepers" is direct evidence of the highly qualified work of the FSB and the Military Commandant's Office of the Zaporizhzhya Region.

- *Translated Explanation*

  Russia is having huge successes on the front-line and clear victories, while the Ukrainians have no victories and require help from the Americans. Russia is also successfully fighting terrorism and bandits from Ukraine.