PolyNarrative: A Multilingual, Multilabel, Multi-domain Dataset for Narrative Extraction from News Articles

Nikolaos Nikolaidis[1], **Nicolas Stefanovitch**[2], **Purificação Silvano**[3], **Dimitar Dimitrov**[4], **Roman Yangarber**[5], **Nuno Guimarães**[6,3], **Elisa Sartori**[7], **Ion Androutsopoulos**[1,8], **Preslav Nakov**[9], **Giovanni Da San Martino**[7], **Jakub Piskorski**[10],
[1]Department of Informatics, Athens University of Economics and Business,
[2]European Commission Joint Research Centre, [3]University of Porto,
[4]Sofia University "St. Kliment Ohridski", [5]University of Helsinki,
[6]INESC TEC, Portugal, [7]University of Padova,
[8]Archimedes, Athena Research Center, Greece, [9]MBZUAI,
[10]Institute of Computer Science, Polish Academy of Science,
nnikon@aueb.gr, jpiskorski@gmail.com

Abstract

We present **PolyNarrative**, a new multilingual dataset of news articles, annotated for narratives. Narratives are overt or implicit claims, recurring across articles and languages, promoting a specific interpretation or viewpoint on an ongoing topic, often propagating mis/disinformation. We developed two-level taxonomies with coarse- and fine-grained narrative labels for two domains: (i) climate change and (ii) the military conflict between Ukraine and Russia. We collected news articles in four languages (Bulgarian, English, Portuguese, and Russian) related to the two domains and manually annotated them at the paragraph level. We make the dataset publicly available, along with experimental results of several strong baselines that assign narrative labels to news articles at the paragraph or the document level. We believe that this dataset will foster research in narrative detection and enable new research directions towards more multi-domain and highly granular narrative related tasks.

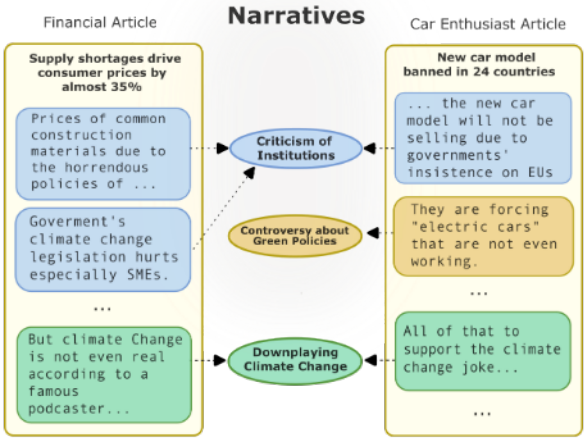

Figure 1: Two articles on different topics (finance vs. car enthusiasts), conveying three different narratives. The three narratives found in the articles are in the round shapes in the middle and the paragraphs are color-coded with the color of the narrative that they contain.

1 Introduction

As online news and social networks have radically altered the media landscape, understanding the dynamics of news propagation and the narratives that news convey is more crucial than ever. Online news outlets make it possible for recurring narratives to appear, often with very different phrasings, in multiple articles and propagate with very high velocity across audiences, languages, and countries. This is especially problematic with manipulative narratives, potentially containing strong biases, mis/disinformation, propaganda, or harmful content. Such a risk is even more pervasive when it is related to divisive issues and can have large societal implications, including destabilization, conflict, hate, or incitement to violence.

Various definitions of *narrative* can be found in the literature, depending on the context and the goal of the analysis. In this paper, we define narrative 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*.

Narratives may be embedded in thousands of articles across multiple topics, themes, languages, and news genres (e.g., news reporting or opinion). They can manifest using very different vocabularies, media frames (Entman, 2007), and tones, which makes their automatic detection and extraction very challenging. Moreover, multiple narratives can be present in the same article, or even in the same paragraph or in the same sentence.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31323–31345

An example is shown in Fig. 1, where the narrative *"Downplaying Climate Change"* occurs in two articles from very different domains: finance vs. auto enthusiasts.

Thus, the effective automatic detection, extraction, and analysis of narratives in news articles is an important challenge and remains an open problem (§2). Such capacity is crucial for analyzing the news landscape, identifying media bias, and detecting attempts to influence readers. It can also be used to warn and educate media consumers to improve their media literacy. Large Language Models (LLMs) have made the need for such capacity even more urgent, as they are vulnerable to (Xu et al., 2024) and prone to generating manipulative narratives (Vykopal et al., 2024) with modest effort.

In this work, we focus on the problem of *narrative detection*, treating it as multi-label paragraph-level classification problem (§3). The biggest challenge is the lack of annotated datasets (Nunes et al., 2024), especially for long news articles (as opposed to micro-blogs or snippets). Existing narrative classification datasets for long articles (§2) have been annotated at the document level and assign only a single label per annotation. Narratives, however, are often conveyed by small parts of the articles, and thus it is important to pinpoint those parts, which is why we model the problem as a *paragraph-level* classification.

We consider two news domains of current relevance (§3.1): climate change (CC) and the military conflict between Ukraine and Russia (URW). Unlike previous work on narrative classification, where the labels are flat and coarse (Li et al., 2023) (§2), we developed a two-level taxonomy with coarse- and fine-grained narrative labels (hereafter *narratives* and *sub-narratives*) for each topic (§3.2). We collected news articles in four languages—English, Bulgarian, Portuguese, and Russian—related to the two domains (§4.1), and manually annotated them at the paragraph level with labels from the corresponding taxonomies (§4.2). We report statistics about the resulting dataset (§4.3) and discuss its quality (§4.4). We make the dataset, dubbed **PolyNarrative**, publicly available, along with experimental results (§5) and code of several strong baselines that assign narrative labels to news articles at the paragraph or the document level; in the latter case, we use the union of the paragraph-level ground-truth labels of each document as the correct prediction.

Our contributions are as follows:

- We propose a new two-level hierarchical taxonomy for narrative extraction. We further develop a novel dataset[1] that uses this taxonomy with news article annotations from two levels (coarse- and fine-grained), at the paragraph level (previous work was document level), using multi-label annotation (previous work was mostly single-label) across two current domains (previous work was single-domain), and four languages (previous work was for a single language).
- We provide a comprehensive overview of the data acquisition and annotation process, and highlight the challenges and measures taken to address them.
- We perform evaluation using several strong multi-label classification methods, and draw conclusions that can benefit future work.

2 Related Work

2.1 Narratives in News

The term *narrative* in the context of news articles has been used in a variety of different formulations. This includes modeling them as descriptions of spatiotemporal events (Norambuena et al., 2023; Silvano et al., 2024) and capturing *narrative components* in the form of semantic relationships between entities (Nunes et al., 2024).

Related but distinct is the problem of Opinion Mining (Liu and Zhang, 2012; Cortis and Davis, 2021), where the task is to extract the opinion of an entity (opinion holder) on an entity or issue (or group thereof), based on an aspect and determine its polarity. Although a restricted subset of narratives can be expressed as opinions, the two concepts serve very different purposes.

In an effort to synthesize various formulations of the concept, Dennison (2021) proposed a refined definition of narratives as "*selective depictions of reality across at least two points that can include one or more causal claims, and are generally generalizable and can be applied to multiple situations, as opposed to specific stories*". This formulation, although well structured from a theoretical perspective, was not sufficient to properly operationalise the concept for our intention to provide a concrete taxonomy.

[1]We make the dataset available. More information on: https://nikon95.github.io/narrative-detection/

Denmark Punishing Farmers for Cow 'Emissions' to 'Fight Global Warming'

Denmark has become the first country to force farmers to comply with the goals of the World Economic Forum's (WEF) "Net Zero" agenda, to supposedly "fight global warming." ...
Criticism of institutions and authorities: Criticism of national governments

Although carbon dioxide is typically blamed for causing "climate change," globalists claim that methane traps about 87 times more heat on a 20-year timescale but top scientists have debunked these claims as a hoax. A recent peer-reviewed study provided conclusive scientific evidence proving that carbon dioxide (CO2) emissions in Earth's atmosphere cannot cause "global warming."
Questioning the measurements and science: Methodologies/metrics used are unreliable/faulty
Questioning the measurements and science: Greenhouse effect/carbon dioxide do not drive climate change

However, not only did they find that higher levels of CO2 made no difference, but they also proved that it simply isn't possible for increases in carbon dioxide to cause temperatures to rise.
Questioning the measurements and science: Greenhouse effect/carbon dioxide do not drive climate change

Nevertheless, unelected globalists at the United Nations Environment Program claim that livestock accounts for about 32% of "human-caused methane emissions." ...
Criticism of institutions and authorities: Criticism of international entities

Figure 2: Example annotated article from the Climate Change domain; the paragraph-level annotations are in **bold**.

Thus, our definition of narrative (§1), while informed by the above general formulation, is based on previous efforts to create taxonomies and datasets for practical use (Kotseva et al., 2023; Li et al., 2023; Coan et al., 2021).

2.2 Automatic Narrative Detection

To apply modern NLP methods, as highlighted by Santana et al. (2023), extraction of narratives has been approached as hierarchical event detection (Glavaš et al., 2014), as frame bias detection (Zaghouani et al., 2024), as spatiotemporal entity relationship detection (Nunes et al., 2024), and as multi-class classification (Li et al., 2023).

Li et al. (2023) presented a social media dataset based on short texts from microblogs and trained a BERT-based narrative classifier detecting COVID anti-vaccine claims from and using a flat taxonomy, while Weigand et al. (2022) focused on conspiracy narrative detection by applying a binary classifier to detect the presence of conspiracy theories and applied a topic modeling segmentation without a pre-defined taxonomy. Coan et al. (2021) released an English-only dataset using a two-level taxonomy on climate change denial, annotated at the snippet level from website fragments and using a single label per snippet, while Kotseva et al. (2023) provided a multi-lingual dataset annotated at document-level using a single label per document.

While datasets with individual characteristics (long-articles, hierarchical taxonomies, multilinguality) have been released before, to the best of our knowledge, there has been no other dataset that provides a multi-lingual corpus, annotated at paragraph level using a hierarchical taxonomy in a multi-label multi-class fashion.

2.3 Narrative Taxonomies

Several narrative taxonomies have been created to classify recurring argumentation in online news. Kotseva et al. (2023) created a three-level narrative taxonomy on COVID-19 and used it to classify and to analyze trends over time. Hughes et al. (2021) presented a taxonomy of common anti-vax narratives, organized on several common tropes and rhetorical strategies. Coan et al. (2021) presented a two-level taxonomy for common cases of climate change denial in short snippets. Amanatullah et al. (2023) presented a flat taxonomy of common pro-Russian narratives found in the alleged pro-Kremlin influence campaigns related to the war in Ukraine. In this work, we used the last two taxonomies as a starting point for the two domains of interest.

3 Task Description

Hereafter, we use *Narrative* (capitalized) to refer to a top-level (coarse) label of our taxonomies. A *sub-Narrative* is a lower-level (fine) label of our taxonomies, representing a narrower claim within a Narrative. For example, *"CO_2 is beneficial"* is a sub-Narrative of the *"Climate Change is beneficial"* Narrative of the Climate Change (CC) taxonomy.

A (sub-)Narrative can appear in a variety of contexts and topics, and can manifest using subtle and indirect argumentation or framing, and thus its detection can be very challenging.

For example, the sub-Narrative *"Western sanctions will backfire"* could be expressed throughout a financial reporting article, or in a small paragraph at the end of an opinion article commenting on military matters.

Thus, in order to better tackle this challenge, our approach focuses on highlighting *specific segment(s)* of each news article that expresses each (sub-)Narrative. In this sense, we consider a paragraph as a minimal segment. If a (sub-)Narrative is expressed in multiple paragraphs of an article (or the entire article), all the corresponding paragraphs need to be annotated accordingly. Moreover, one paragraph can express more than one Narrative and more than one level of Narratives (Narratives and sub-Narratives). In these scenarios, the paragraph needs to be annotated with *all* the applicable (sub-)Narratives. Given this formulation, the task becomes a hierarchical multi-label multi-class classification problem at the paragraph level.

3.1 Domains

We selected two domains (topics) that currently receive extensive news coverage, are presented through several different perspectives, and are highly susceptible to manipulation through the creation of artificial narratives. The first domain is Climate Change (CC) and the second one is the Ukraine-Russia War (URW). In both cases, we have numerous accounts of repeated argumentation in the form of narratives. We are particularly interested in potentially manipulative narratives, and this is reflected in the choice of the (sub-)Narratives of our taxonomies. However, we do not assume that every claim corresponding to a (sub-)Narrative of our taxonomies is necessarily a case of mis/-disinformation. This is important to highlight, as mixing legitimate with mis/disinformative claims is a frequent manipulative practice (Goel et al., 2023).

3.2 Taxonomies

To create the two taxonomies, we started from existing narrative taxonomies for URW (Amanatullah et al., 2023) and CC (Coan et al., 2021). We asked media analysts to suggest modifications based on their media monitoring experience. These modifications included additions (new (sub-)Narratives), elaborations (splitting existing (sub-)Narratives into multiple different ones), modifications (refining definitions and names) and merging (grouping into common (sub-)Narratives). We asked, when possible, to phrase each (sub-)Narrative as a concrete claim (e.g., "*Renewable energy is dangerous*" or "*Ukraine is a puppet of the West*"). When the argumentation regarding a (sub-)Narrative was very diverse or fragmented, we gave a more general descriptive label (e.g., "*Criticism of Institutions*" or "*Speculating war outcomes*"). We ended up with 38 and 36 sub-Narratives, grouped into 11 and 10 Narratives for URW and CC, respectively.

The top level of the two taxonomies (Narratives) is shown in Figure 3. The full taxonomies, including sub-Narratives and definitions, can be found in Appendix A. We note that the two taxonomies are not complete; in fact, they are not intended to be complete, rather they are representative of the practical experience of the media analysts, and different valid formulations are possible. We stress again that although our taxonomies focus on potentially manipulative narratives, we do not assume that every claim corresponding to a (sub-)Narrative of our taxonomies is necessarily mis/disinformation or indicates strong bias.

UKRAINE-RUSSIA WAR (URW)

- Blaming the war on others rather than the invader
- Discrediting Ukraine
- Russia is the Victim
- Praise of Russia
- Overpraising the West
- Speculating war outcomes
- Discrediting the West, Diplomacy
- Negative Consequences for the West
- Distrust towards Media
- Amplifying war-related fears
- Hidden plots by secret schemes of powerful groups

CLIMATE CHANGE (CC)

- Criticism of climate policies
- Criticism of institutions and authorities
- Climate change is beneficial
- Downplaying climate change
- Questioning the measurements and science
- Criticism of climate movement
- Controversy about green technologies
- Hidden plots by secret schemes of powerful groups
- Amplifying Climate Fears
- Green policies are geopolitical instruments

Figure 3: Coarse-grained Narratives for the Ukraine-Russia War (URW) and Climate Change (CC) domains. The sub-Narratives are given in Appendix A.

4 The PolyNarrative Dataset

4.1 Data Acquisition

To obtain news articles for annotation, we used an in-house online news scraping and indexing tool to retrieve articles on the two selected domains in the four target languages.[2] We considered both mainstream and "alternative" news sources (that is, sources known to repeatedly publish content

[2]The tool regularly collects and indexes news articles of most major European and many worldwide news outlets.

with misinformation). We also developed ad-hoc scoring methods, outlined below, to heuristically evaluate the relevance of each article to the corresponding domain. More specifically:

1. We created keyword-based queries for the two domains in all the languages, and used them to retrieve a large number of articles from the news article index of our in-house tool. For the URW domain, we included documents from 2021 to 2024. For the CC domain, the documents ranged from 2015 to 2024.
2. To assess the relevance of the resulting articles, we formulated another set of key phrases, corresponding to the (sub-)Narratives of the taxonomies (e.g., "Ukraine is corrupt", "Science is debunked") or other aspects we wished to emphasize or de-emphasize. We then used *bart-large-mnli*[3] to perform zero-shot classification with the title and the first 300 characters of each article, resulting in a classification score per article and key phrase.
3. We used an *XLM-RoBERTa*-based [4] multi-label classifier, trained on the Persuasion Techniques dataset (Piskorski et al., 2023a,b), and used the approach of Nikolaidis et al. (2024) to produce Persuasiveness Score metrics per article. For the CC domain, we also used a climate change denial classifier, based on the work of Coan et al. (2021), to further filter and reduce the number of articles to review.
4. We used a linear combination of the resulting scores (relevance score per key phrase, Persuasiveness Score metrics and the score from the climate change denial classifier) to automatically rank articles from most to least likely to contain relevant (sub-)Narratives.

A member of the team who was familiar with the news landscape of the language, then manually inspected the resulting set of articles, focusing on articles that seemed relevant to the defined taxonomies. Since manual inspection was time-consuming, we used the different scores computed above to reorder the articles and focus mainly on highly ranked articles.

4.2 Data Annotation

In order to clarify the annotation process, including the meaning of key concepts such as Narrative, sub-Narrative, the precise meaning of each label of the taxonomies, and good practices, we created a document with detailed annotation guidelines. This included specific definitions and examples for each label, together with a detailed description of the annotation process to be followed. The definitions and examples of the taxonomies can be found in Appendix A.

As we wanted to produce a multilingual dataset with four languages, we had multiple annotator teams, each focused on one language. We assigned two annotators per news article and a curator who handled consolidation and quality assurance before the final version. We held regular calls in which hard cases were highlighted, comparisons across annotations from different languages were made, and issues with ambiguities of the guidelines were discussed. To annotate the documents, we used the *Inception* annotation tool (Klie et al., 2018).

Each language was assigned a coordinator who had the responsibility to sort through the collected articles, assign annotators and curators, and create annotation batches incrementally. The coordinator was also responsible for training the annotators.

We instructed the annotators to read each news article paragraph by paragraph, and to annotate each of them with all the applicable Narratives and sub-Narratives before moving to the next one. When no Narrative or sub-Narrative was found, the annotator simply moved to the next paragraph. Because the sub-Narrative labels were many and difficult to memorize, we configured Inception so that for each paragraph, the annotator would first select all the applicable Narratives, and then all the applicable sub-Narratives from a menu displaying the sub-Narratives of the selected Narratives only. When a paragraph was deemed to contain a Narrative, but none of its sub-Narratives, the *"[Narrative]: Other"* label was selected at the sub-Narrative level. Similarly, the *"Other"* label was selected when no Narrative was applicable.

During the annotation process, several issues were highlighted. First, the distribution of labels was very uneven and varied noticeably across languages (Table 2). This reflects the reality of the news discourse, as some Narratives are more frequent in some countries than others. After each annotation batch, the collected label statistics were used by the coordinator to optimize the selection of articles for the next batch. Second, some labels were semantically very close, which led to frequent confusion. We used frequency metrics and cura-

[3] huggingface.co/facebook/bart-large-mnli
[4] huggingface.co/FacebookAI/xlm-roberta-large

tor feedback to discuss such cases in the regular meetings, where we reiterated and further refined the guidelines. For example, the sub-Narratives "*Ukraine is a puppet of the West*" and "*The West does not care about Ukraine, only their interests*" were frequently confused. To resolve this ambiguity, we added guidelines instructing the annotators to select the former when Ukraine was the subject and the latter when the West was.

4.3 Dataset Statistics

Table 1 shows statistics about the documents of the **PolyNarrative** dataset, broken down per language, and aggregated over both domains. We can see that, except for Russian, all languages have approximately 400 annotated training documents; the test documents are approximately 35 for each language. We aimed at articles of relatively long text with a median length of 500 words and a minimum length of 250 words.

Even after the rigorous data selection process, the final label distribution after annotation is highly imbalanced (see Table 2 and Appendix B for a breakdown per language). Additionally, the distribution of labels varies noticeably from language to language, due to differences in media interest across countries, as already discussed. For example, for URW, Russian is a notable outlier in terms of distribution, with "Praise of Russia" being one of the most common Narratives. One note to highlight, is that after the article selection process, the proportion of mainstream and alternative news varies widely across languages and could have an impact on label distribution.

Moreover, 44% and 54% of the documents in CC and URW, respectively, are assigned more than one label; for paragraphs, the corresponding percentages are 12% and 15%, respectively. This observation justifies our choice to perform multi-label annotation.

4.4 Dataset Quality

Table 3 reports the inter-annotator agreement scores for Narratives (coarse labels) and sub-Narratives (fine-grained labels), per language, collectively for both domains, measured as Krippendorff's α at paragraph level using the `simpledorff`[5] library. The agreement is under the recommended value of 0.667 but is higher than the IAA of tasks of similar complexity (Piskorski et al., 2023b, 2024).

[5] github.com/LightTag/simpledorff

TRAIN				
Language	#documents	#paragraphs	#sent.	#char.
BG	400	3,951	6,728	937,150
EN	394	3,683	8,012	1,176,513
PT	395	3,877	5,739	971,356
RU	133	583	1,630	206,124
TEST				
Language	#documents	#paragraphs	#sent.	#char.
BG	35	325	542	70,291
EN	41	584	1,106	144,401
PT	35	330	578	86,132
RU	32	154	478	56,872
Total	**1,465**	**13,487**	**24,813**	**3,649k**

Table 1: Document statistics per language, collectively for both domains (CC, URW).

Narrative	#para	#doc
CC: Amplifying Climate Fears	1,024	238
CC: Climate change is beneficial	29	12
CC: Controversy about green technologies	101	31
CC: Criticism of climate movement	230	65
CC: Criticism of climate policies	234	102
CC: Criticism of institutions and authorities	467	171
CC: Downplaying climate change	135	49
CC: Green policies are geopolitical instruments	14	11
CC: Hidden plots by secret schemes of powerful groups	138	50
CC: Questioning the measurements and science	83	29
URW: Amplifying war-related fears	486	185
URW: Blaming the war on others rather than the invader	282	149
URW: Discrediting Ukraine	1,069	366
URW: Discrediting the West, Diplomacy	862	331
URW: Distrust towards Media	80	39
URW: Hidden plots by secret schemes of powerful groups	53	20
URW: Negative Consequences for the West	152	79
URW: Overpraising the West	34	23
URW: Praise of Russia	602	271
URW: Russia is the Victim	318	167
URW: Speculating war outcomes	161	79
Total annotations	**6,554**	**2,467**

Table 2: Distribution statistics for coarse-grained labels (Narratives), assigned to paragraphs and documents, in both TRAIN and TEST splits. Statistics for fine-grained (sub-Narratives) are shown in Appendix B.

As one would expect, the agreement on finer labels is lower, because of their subtler differences.

We noticed that the CC domain caused more confusion between the annotators than URW. In both cases, we could see that the confusion was skewed by a small set of under-agreed labels (5 for URW and 7 for CC) that achieved disagreement above 40% and 60%, respectively. If we excluded these labels, the IAA for all languages rises to 0.567 and 0.560 for the coarse and 0.452 and 0.516 for fine-grained, for CC and URW respectively.

There were some sub-Narratives that were commonly confused. For example, in the URW subset "*Discrediting the West, Diplomacy: West is tired of Ukraine*", "*Discrediting the West, Diplomacy: The West does not care about Ukraine, only about its*

granularity	lang. domain	BG	EN	PT	RU	all
coarse	ALL	0.736	0.499	0.461	0.427	**0.571**
	CC	0.652	0.375	0.465	-	0.524
	URW	0.700	0.558	0.362	0.427	0.533
fine	ALL	0.642	0.388	0.385	0.415	**0.480**
	CC	0.541	0.283	0.331	-	0.408
	URW	0.626	0.457	0.349	0.415	0.479

Table 3: Krippendorff's α for different granularities, languages and domains on paragraph level.

interests", and "*Discrediting Ukraine: Ukraine is a puppet of the West*".

Overall, on a positive note, the majority of the disagreement between the two annotators in the sub-Narrative labels was between labels of the same Narrative (e.g. sub-Narratives under "*Discrediting the West, Diplomacy*" were frequently confused with one another) On average, out of all the individual annotator disagreements (paragraphs where the two annotators picked a different sub-Narrative), 67% was with sub-Narratives of the same Narrative. A detailed breakdown of the disagreements statistics for each label can be found in Appendix E.

5 Experiments and Evaluation

5.1 Experimental Setup

To provide baseline results for the new **PolyNarrative** dataset, we performed experiments with commonly used multi-label classification methods. We assessed the following configurations:

- **Label granularity**: We assessed the performance in both coarse (predicting Narratives) and fine-grained (predicting sub-Narratives) settings. Additionally, we assessed how the extra information about the fine-grained predictions impacted coarse-grained predictions.
- **Paragraph vs. document**: We assessed the performance when the models were trained and evaluated to classify paragraphs vs. entire documents. The document-level labels were produced by taking the union of the labels of all of the paragraphs contained in each document.
- **Prompting vs. fine-tuning**: We also assessed the capability of (open-source) LLMs to annotate (sub-)Narratives without further training, given instructions similar to the guidelines given to the human annotators, comparing performance to fine-tuned models.

We experimented with the following models:

- *XLM-RoBERTa$_{large}$* (Conneau et al., 2020), in two versions:
  - *XLM-RoBERTa-coarse*: trained using only the coarse labels.
  - *XLM-RoBERTa-fine*: trained using the fine-grained labels.
- *Llama3.1-70b* (Grattafiori et al., 2024), without fine-tuning, in four versions:
  - *Llama3.1-70b-0shot-labels*: Prompted with the names of the labels of the taxonomy only, in a zero-shot fashion.
  - *Llama3.1-70b-guidelines*: Prompted with the full annotation guidelines, which include hand-picked examples for each (sub-)Narrative.
  - *Llama3.1-70b-guidelines-labels*: Prompted with the names of the labels, followed by the full annotation guidelines.
  - *Llama3.1-70b-hierarchical*: two-step hierarchical prompting, where the model is instructed to pick the coarse-grained labels first, and then given these coarse-grained predictions, decide on the fine-grained label; full guidelines are also provided.

When using Llama, after receiving the model's response, we performed a series of post-processing operations, where the generated output labels were filtered and normalized to remove hallucinated labels and to handle cases where the model generated labels in a slightly different form (e.g., different capitalization, punctuation errors).

To extract paragraph-level annotations from Llama, we inserted extra "*[paragraph_N:]*" tags at the beginning of each paragraph, *N* being the number of the paragraph, and modified the prompts to instruct the LLM to output each extra tag followed by the relevant labels of the corresponding paragraph. After the call to the LLM, we recombined the outputs with the input paragraphs in a post-processing phase, using regular expressions. The exact prompts used are given in Appendix C.

The XLM-RoBERTa (Conneau et al., 2020) models were trained at the token level. For each token, a multi-hot vector equal in size to the number of labels was provided as ground-truth, and each

		document				paragraph			
		coarse		fine		coarse		fine	
Domain	**Model**	**F1**	**F1**$_{sample}$	**F1**	**F1**$_{sample}$	**F1**	**F1**$_{sample}$	**F1**	**F1**$_{sample}$
CC	XLM-RoBERTa-fine	0.457	0.523	**0.198**	**0.347**	0.282	0.161	**0.146**	**0.116**
	XLM-RoBERTa-coarse	0.423	0.550	-	-	0.266	**0.189**	-	-
	Llama3.1-70b-0shot-labels	0.542	**0.577**	0.144	0.198	0.349	0.149	0.076	0.043
	Llama3.1-70b-guidelines	0.503	0.493	0.151	0.142	0.326	0.128	0.083	0.044
	Llama3.1-70b-guidelines-labels	**0.573**	0.505	0.168	0.199	**0.352**	0.138	0.085	0.050
	Llama3.1-70b-hierarchical	0.534	0.548	0.130	0.226	0.302	0.146	0.055	0.053
	Stratified Random Baseline	0.106	0.223	0.016	0.090	0.062	0.071	0.008	0.026
URW	XLM-RoBERTa-fine	0.410	0.486	**0.210**	**0.380**	0.257	0.143	**0.128**	**0.105**
	XLM-RoBERTa-coarse	0.428	**0.524**	-	-	0.261	**0.185**	-	-
	Llama3.1-70b-0shot-labels	**0.475**	0.425	**0.210**	0.245	**0.297**	0.124	0.126	0.071
	Llama3.1-70b-guidelines	0.359	0.370	0.155	0.182	0.232	0.121	0.069	0.046
	Llama3.1-70b-guidelines-labels	0.432	0.373	0.190	0.249	0.291	0.125	0.117	0.078
	Llama3.1-70b-hierarchical	0.381	0.420	0.172	0.173	0.205	0.115	0.073	0.048
	Stratified Random Baseline	0.148	0.179	0.033	0.046	0.044	0.042	0.013	0.018

Table 4: Test set results when classifying documents or paragraphs, for Narratives (coarse) and sub-Narratives (fine), averaged over all languages, for Climate Change (CC) and Ukraine-Russia War (URW), respectively.

token inherited the ground-truth labels of its paragraph. At inference time, we obtained a similar predicted multi-hot vector from each token, and each label was assigned (or not) based on the majority vote of the paragraph's tokens. To bypass the 512 token limitation of XLM-R, we used a sliding window with a 50% overlap.

We also report the results for a naïve stratified random baseline, where we randomly draw labels from a multinomial distribution, while respecting the frequency statistics in the training set.

5.2 Experimental Results

Since our task is multi-label multi-class classification with a large number of classes, and noticeable class imbalance, we opted to use suitable metrics that are robust in these conditions. Table 4 presents the F_1 macro and sample-averaged F_1 scores for all models in both domains (CC, URW). Sample-average F_1 score is calculated by calculating the F_1 score on each instance and subsequently averaging over all instances. We observe that the XLM-RoBERTa$_{large}$ classifier outperforms Llama3.1-70B in the fine-grained case both for document- and paragraph-level predictions, especially in sample F_1, where the difference is very pronounced. The Llama3.1 prompted model seems to be more competitive in coarse-grain evaluation.

Interestingly, including the full guidelines in the prompt generally achieves worse results than using the taxonomy labels only. When we add both the taxonomy labels and the guidelines, performance improves, compared to using only the guidelines, but it is unclear whether it is better to include only the taxonomy labels. Hierarchical prompting does not seem to yield competitive results.

In terms of per-label performance, the weakest model performance was measured on the sub-Narratives under "*Downplaying climate change*" and "*Green policies are geopolitical instruments*" Narratives in the CC domain and "*Russia is the Victim*" and "*Blaming the war on others rather than the invader*" in the URW domain.

One key observation is that models exhibited relatively lower performance for the "*[Narrative]: Other*" labels, that is, the paragraphs that contained a Narrative but none of the listed sub-Narratives, that may include a very diverse set of argumentation, coming from a set of heterogeneous sub-Narratives that our taxonomy does not cover.

Performance also seemed to vary noticeably per language. In case of Russian, Llamma3.1-70b models seemed to perform rather poorly on a per-sample level. On XLM-RoBERTa, English consistently exhibited poorer performance compared to other languages, with Bulgarian and Portuguese showing different performance per domain. We provide a detailed breakdown of the performance (by label and by language) of two models (XLM-RoBERTa$_{fine}$ and Llama3.1-70b-guidelines-labels) on Appendix D.

6 Conclusion and Future work

We present **PolyNarrative**, a new multi-lingual, multi-label, multi-domain dataset for extraction of narratives at the paragraph level from long news

articles. The dataset comprises 1,476 articles totaling 13,625 paragraphs in 4 languages (Bulgarian, English, Portuguese, and Russian), annotated using two expert-refined Narrative hierarchical taxonomies with 38 and 36 fine-grained and 10 and 11 coarse-grained narratives for each domain. We describe in detail the data acquisition and annotation process and highlighted noteworthy issues.

We present preliminary experimental results using both fine-tuned encoder (XLM-R) and prompting-based decoder (Llama 3.1) Language Models in multiple granularities (fine, coarse), levels (paragraph, document), and configurations (prompting strategies). We also highlight interesting findings regarding model performance.

We hope that this dataset will catalyze new research directions in narrative detection and extraction, and stimulate the development of new methods and techniques. We highlight that the taxonomies are based on expert real-world experience and they are not meant to be perceived as complete or to capture the whole breadth of the media discourse around the two domains. We encourage the research community to propose extensions that potentially capture more diverse perspectives. We encourage the direction of developing methods that are domain- and potentially taxonomy-agnostic, able to perform narrative detection in a completely unsupervised fashion, removing the dependence on human subjectivity. We believe that a multi-domain and highly granular dataset such as this one can facilitate experimentation towards this goal.

Regarding future work, we plan to experiment with ways to perform generation of new narrative labels and automatic revision of taxonomy and guidelines. One way to approach this is by analyzing the structure of narrative elements in the text relate them with narrative labels and identify new structures. We also intend to use retrieval strategies to perform narrative retrieval within large news corpora.

7 Ethics Policy

Intended Use and Misuse Potential The main drive behind the creation of our dataset was to advance research on automated narrative classification and the detection of deceptive content across multiple languages and domains. However, given that possible risk of exploiting the dataset to boost the production of biased manipulative disinformation attempts, we advise responsible use. In this context, whoever develops a Narrative detection system is also responsible for deciding which Narratives to detect, in an ethical manner.

Furthermore, this research in direction can contribute to the development of independent tools such as browser plugins or independent publicly available services that could help users contextualize the information they consume, contributing to their awareness and improving citizens' media literacy.

Environmental Impact The deployment of LLMs might have a large carbon footprint, especially when training new models. In the context of the reported experiments, we did not train any new LLMs, but only used existing trained models in an in-context zero-shot scenarios, which is relatively cheap in terms of computing.

Fairness The majority of the annotators, primarily researchers with linguistic background and prior annotation experience, come from the institutions of the co-authors of this manuscript. They were fairly remunerated as part of their job.

The remaining part of the annotator pool consisted of (a) some students from the respective academic organizations, (b) few external experienced analysts paid at rates set by their contracting institutions, and (c) experts from a contracted a professional annotation company, who were compensated according to rates based on their country of residence.

8 Limitations

Taxonomies and Dataset Representativeness Our taxonomies were edited by experienced media analysts, active in the study of misinformation and fact-checking. As such, the taxonomies overrepresent Narratives of interest of media analysts from Western institutions. The selection should not be perceived as covering the complete discourse of the two domains, but rather what such analysts encounter in practice. The original taxonomy for URW used before the heavy revision from the analysts, was not from a peer-reviewed publication but from a technical report.

The dataset presented covers two widely discussed domains around the world and a wide range of media sources. However, it should not be considered as representative of the media in any specific country, nor should it be considered as balanced in any way. Also, it is important to note that while this

work focuses on narratives potentially containing mis/disinformation, the operationalized definition of Narrative can also be used to detect neutral or desirable viewpoints.

Biases Although the annotators were trained and made acquainted with the specifics of the two domains of interest for our task and cross-language quality control mechanisms have been put in place in the annotation process, we are aware that some degree of intrinsic subjectivity will inevitably be present in the dataset. Consequently, models trained on this dataset might exhibit certain biases.

Acknowledgements

This research is partially 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.

Also, this work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

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 by 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.

9 References

References

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.

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.

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

Keith Cortis and Brian Davis. 2021. Over a decade of social opinion mining: a systematic review. *Artificial Intelligence Review*, 54(7):4873–4965.

James Dennison. 2021. Narratives: a review of concepts, determinants, effects, and uses in migration research. *Comparative Migration Studies*, 9(1):50.

Robert M. Entman. 2007. Framing bias: Media in the distribution of power. *Journal of Communication*, 57(1):163–173.

Goran Glavaš, Jan Šnajder, Parisa Kordjamshidi, and Marie-Francine Moens. 2014. HiEve: A Corpus for Extracting Event Hierarchies from News Stories. In *Proceedings of 9th language resources and evaluation conference*, pages 3678–3683. ELRA.

Pranav Goel, Jon Green, David Lazer, and Philip Resnik. 2023. Mainstream news articles co-shared with fake news buttress misinformation narratives. *Preprint*, arXiv:2308.06459.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad,

Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim

Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Brian Hughes, Cynthia Miller-Idriss, Rachael Piltch-Loeb, Beth Goldberg, Kesa White, Meili Criezis, and Elena Savoia. 2021. Development of a codebook of online anti-vaccination rhetoric to manage covid-19 vaccine misinformation. *International Journal of Environmental Research and Public Health*, 18(14).

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. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Bonka Kotseva, Irene Vianini, Nikolaos Nikolaidis, Nicolò Faggiani, Kristina Potapova, Caroline Gasparro, Yaniv Steiner, Jessica Scornavacche, Guillaume Jacquet, Vlad Dragu, Leonida Della Rocca, Stefano Bucci, Aldo Podavini, Marco Verile, Charles Macmillan, and Jens P. Linge. 2023. Trend analysis of COVID-19 mis/disinformation narratives–A 3-year study. *PLOS ONE*, 18(11):e0291423.

Yue Li, Carolina Scarton, Xingyi Song, and Kalina Bontcheva. 2023. Classifying COVID-19 vaccine narratives. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 648–657, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bing Liu and Lei Zhang. 2012. *A Survey of Opinion Mining and Sentiment Analysis*, pages 415–463. Springer US, Boston, MA.

Nikolaos Nikolaidis, Jakub Piskorski, and Nicolas Stefanovitch. 2024. Exploring the usability of persuasion techniques for downstream misinformation-related classification tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6992–7006, Torino, Italia. ELRA and ICCL.

Keith Norambuena, Brian Felipe, Tanushree Mitra, and Chris North. 2023. A survey on event-based news narrative extraction. *ACM Comput. Surv.*, 55(14s).

Sérgio Nunes, Alípio Mario Jorge, Evelin Amorim, Hugo Sousa, António Leal, Purificação Moura Silvano, Inês Cantante, and Ricardo Campos. 2024. Text2Story lusa: A dataset for narrative analysis in European Portuguese news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15773–15782, Torino, Italia. ELRA and ICCL.

Jakub Piskorski, Alípio Jorge, Maria da Purificação Silvano, Nuno Guimarães, Ana Filipa Pacheco, and Nana Yu. 2024. Overview of the CLEF-2024 CheckThat! Lab task 3 on persuasion techniques.

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 (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artif. Intell. Rev.*, 56(8):8393–8435.

P. Silvano, E. Amorim, A. Leal, I. Cantante, A. Jorge, R. Campos, and N. Yu. 2024. Untangling a web of temporal relations in news articles. In *Proceedings of the 7th International Workshop on Narrative Extraction from Texts (Text2Story'24). Glasgow. Scotland. March 24. pp. 77-92*.

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2024. Disinformation Capabilities of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14830–14847, Bangkok, Thailand. Association for Computational Linguistics.

Manuel Weigand, Maximilian Weber, and Johannes Gruber. 2022. Conspiracy Narratives in the Protest Movement Against COVID-19 Restrictions in Germany. A Long-term Content Analysis of Telegram Chat Groups. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 52–58, Abu Dhabi, UAE. Association for Computational Linguistics.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu,

and Han Qiu. 2024. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.

Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, Houda Bouamor, Imed Zitouni, Mona Diab, Samhaa El-Beltagy, and Muhammed AbuOdeh. 2024. The FIGNEWS Shared Task on News Media Narratives. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 530–547, Bangkok, Thailand. Association for Computational Linguistics.

A Taxonomies

This section introduces the taxonomies of narratives and subnarratives for the two domains.

A.1 Ukraine-Russia War (URW)

1. **Blaming the war on others rather than the invader:** statements attributing responsibility or fault to entities other than Russia in the context of Russia's invasion of Ukraine.
   Example: *"The economic crisis is due to Western sanctions."*
   Example: *"Ukraine's actions provoked this conflict."*

   Sub-Narratives:

   (a) **Ukraine is the aggressor**: Statements that shift the responsibility of the aggression to Ukraine instead of Russia and portray Ukraine as the attacker.
   Example: *"Ukraine secretly provoked the war because it was harassing Donbass province citizens."*

   (b) **The West are the aggressors:** Statements that shift the responsibility for the conflict and escalation to the Western block.
   Example: *"The real perpetrators were US/EU. They sabotaged Minsk II agreement only to force Russia to invade."*

2. **Discrediting Ukraine**: Statements that undermine the legitimacy, actions, or intentions of Ukraine or Ukrainians as a nation.
   Example: *"Ukraine's government is corrupt and ineffective."*
   Example: *"Ukrainian soldiers are committing atrocities."*
   Example: *"Ukrainian identity does not exist"*

   Sub-Narratives:

   (a) **Rewriting Ukraine's history:** Statements that aim to reestablish history of Ukrainian nation in a way that discredits its reputation.
   Example: *"Ukraine is not a real nation, it was a fabrication to split Russia and ally with Hitler."*

   (b) **Discrediting Ukrainian nation and society:** Statements that aggressively undermine the legitimacy and reputability of Ukrainian ethnicity and people

(c) **Discrediting Ukrainian military:** Statements that aim to undermine the capabilities, professionalism or effectiveness of the Ukrainian armed forces.

(d) **Discrediting Ukrainian government and officials and policies:** Statements that seek to delegitimize the Ukrainian government, its leaders, and its policies, portraying them as corrupt or incompetent.

(e) **Ukraine is a puppet of the West:** Claims that Ukraine is controlled or heavily influenced by Western powers, particularly the United States and European Union.

(f) **Ukraine is a hub for criminal activities:** Allegations that Ukraine is a center for illegal activities such as human trafficking, drug smuggling, or organized crime

(g) **Ukraine is associated with nazism:** Accusations that Ukrainian society or government has ties to or sympathies with Nazi ideology, often referencing historical events or extremist groups.

(h) **Situation in Ukraine is hopeless:** Statements that portray Ukraine as having no viable perspectives or no potential positive future.
Example: *"Ukraine should just give up, it is all over debt and will be exploited by the West anyway."*

3. **Russia is the Victim:** Statements that portray Russia as being unfairly targeted or victimized.
Example: *"Russia is being unfairly sanctioned."*
Example: *"The West is ganging up on Russia without justification."*
Example: *"Russia is doing what every country would do (e.g. protect its interests/honour etc.)"*

Sub-Narratives:

(a) **The West is russophobic:** Statements that claim that the negative reaction to Russia's actions are because of the negative perspective of western countries instead of Russia's own actions.
Example: *"Politicians in the West blame Russia for everything, instead of looking at their mistakes."*
Example: *"In Country X, they banned Tchaikovsky ballets and Chechov's plays because they cannot stand Russia and its culture."*

(b) **Russia actions in Ukraine are only self-defence:** Statements that justify Russia's action solely as legitimate self-defence and not a deliberate action.
Example: *"There was no other way than war to defend the Russian-speaking people in Donbass."*

(c) **UA is anti-RU extremists:** Statements claiming that Ukraine is comprised of extremist elements that are vehemently opposed to Russia.

4. **Praise of Russia:** Statements that positively highlight Russia's actions, policies, or character
Example: *"Russia is leading the way in international diplomacy."*
Example: *"The Russian economy is resilient and strong."*
Example: *"Glorifying mentions of Russia's weapon systems and military might."*

Sub-Narratives:

(a) **Praise of Russian military might:** Statements that positively highlight Russia's military institutions, equipment and scale.
Example: *"Russia has far more tanks and powerful artillery that US/EU would only dream of."*

(b) **Praise of Russian President Vladimir Putin:** Statements that present Vladimir Putin positively, including his personal and leadership qualities.
Example: *"Any country would want such a strong leader as Putin to lead the way."*

(c) **Russia is a guarantor of peace and prosperity:** Statements that portray Russia solely in a positive manner, emphasising their potential to provide peace and prosperity to those that cooperate.
Example: *"Take a look at Africa, Russia supports countries and turns them into independent nations guided by their people's interests where Western countries colonised brutally."*

(d) **Russia has international support from a number of countries and people:** Statements that emphasise the popularity and acceptance of Russia in the international stage.
Example: *"The majority of the countries population sides with Russia as per last UN General Assembly vote."*

(e) **Russian invasion has strong national support:** Statements that emphasise the popularity and acceptance of the invasion inside Russia and on Russian-speaking populations.
Example: *"The majority of the countries population sides with Russia as per last UN General Assembly vote."*

5. **Overpraising the West:** Statements that excessively and unduly laud or extol the virtues, accomplishments, and moral superiority of Western countries, particularly in the context of international relations and military.

Sub-Narratives:

(a) **NATO will destroy Russia:** Statements that suggest or claim that the North Atlantic Treaty Organization (NATO) and its allies are capable or already in the process of eradicating Russia.

(b) **The West belongs in the right side of history:** Statements that portray Western nations and their actions as morally superior and aligned with progress and justice and possess moral superiority.

(c) **The West has the strongest international support:** Statements that emphasize or claim widespread backing for Western policies and actions from the international community, potentially downplaying opposition or criticism.

6. **Speculating war outcomes:** Statements that predict or make assumptions about the potential results or consequences of a conflict

Sub-Narratives:

(a) **Russian army is collapsing:** Statements that suggest or claim that the Russian military is experiencing a significant decline in its effectiveness, strength, or morale.

(b) **Russian army will lose all the occupied territories:** Speculative statements that predict or assume the potential outcomes of the conflict, specifically regarding the possibility of the Russian military losing control of all the territories it currently occupies.

(c) **Ukrainian army is collapsing:** Statements that suggest or claim that the Ukrainian military is experiencing a significant decline in its effectiveness, strength, or morale.

7. **Discrediting the West, Diplomacy:** Statements that criticize the Western countries, or international diplomatic efforts.
Example: *"The West is hypocritical in its foreign policy."*
Example: *"Western diplomacy has failed in resolving conflicts."*
Example: *"International organizations will not solve anything because..."*

Sub-Narratives:

(a) **The EU is divided:** Statements that present the EU as a set of divided entities and interests, usually unable to take actions.
Example: *"The European Council will never vote on sanctions for Russia, since they cannot agree on even the simplest of the issues."*

(b) **The West is weak:** Statements presenting the West overall as a non-potent group of countries (that is not as powerful as it used to be).
Example: *"The weakened West is once again impotent to act in front of the will Russia."*

(c) **The West is overreacting:** Statements that claim that the West and its institutions are reacting to Russia's actions in a disproportionate manner.
Example: *"Putin did not invade the EU but Ukraine. Imposing harsh sanctions is not the way to deal with it, dialogue and debate is."*

(d) **The West does not care about Ukraine, only about its interests:** Statements that claim that the West is only interested in Ukraine for its own benefits, disregard-

ing the country's fate.
Example: *"The West has indebted Ukraine more than XX bln of dollars, a lucrative deal for western companies to exploit."*
Example: *"NATO's actions are endangering global security."*

(e) **Diplomacy does/will not work:** Statements discrediting the potential of ongoing or potential diplomatic efforts.
Example: *"Diplomats are desperately trying to figure out solutions but now it's too late, they have failed and Russia is free to do whatever."*

(f) **West is tired of Ukraine:** Claims that Western countries, particularly the United States and European nations, are becoming fatigued or disinterested in supporting Ukraine and its efforts.

8. **Negative Consequences for the West:** Statements that highlight or predict adverse outcomes for Western countries and their interests.
Example: *"Sanctions against Russia will backfire on Europe."*
Example: *"The West is headed for an economic downturn."*

Sub-Narratives:

(a) **Sanctions imposed by Western countries will backfire:** Statements that catastrophize on the possible negative effects for Western sanctions of Russia.
Example: *"The winter is going to be cold and with current gas prices, we are talking of societal unrest."*

(b) **The conflict will increase the Ukrainian refugee flows to Europe:** Statements that catastrophize on the possible refugee outflows due to the conflict.
Example: *"Like we did not have refugees from the Middle East, now we will have Ukrainians stressing our housing and healthcare problems."*

9. **Distrust towards Media:** Statements that question the reliability or integrity of media organizations.
Example: *"Western media is spreading propaganda."*
Example: *"You can't trust what the news says about Russia."*

Sub-Narratives:

(a) **Western media is an instrument of propaganda:** Statements that discredit the media institutions of the West and claim that they are instruments of propaganda.
Example: *"... but you wouldn't hear this on a western channel, only the party line from State Department."*

(b) **Ukrainian media cannot be trusted:** Statements that discredit the media institutions of the Ukraine and claim that they should not be trusted for reporting on the war.
Example: *"Ukraine is conducting its own propaganda using their TV channels, news and social media."*

10. **Amplifying war-related fears:** Statements that evoke fear or anxiety about potential threats, dangers or reactions.
Example: *"The West is pushing us towards World War III."*
Example: *"It is a matter of time before war speads on the West"*
Example: *"Nuclear war is imminent"*

Sub-Narratives:

(a) **By continuing the war we risk WWIII:** Statements that warn against upsetting Russia's and its leadership, evoking fear of causing WW3.
Example: *"The Western elites with their fixation on Russia are sleapwalking towards WW3"*

(b) **Russia will also attack other countries:** Statements that claim that it is imminent that Russia will attack other countries.
Example: *"... and be sure, Ukraine is the first not the last country to be invaded. Others will follow."*

(c) **There is a real possibility that nuclear weapons will be employed:** Statements that evoke fear or anxiety about the use of nuclear weapons.
Example: *"... and if Western hypocrisy continues to provoke, Putin might be forced to press the red button... for good"*

(d) **NATO should/will directly intervene:** Statements that suggest or claim that the North Atlantic Treaty Organization (NATO) ought to or will take direct military action in a conflict, potentially implying a shift in policy or strategy.

11. **Hidden plots by secret schemes of powerful groups:** Statements that suggest hidden plots or secretive actions by powerful groups related to the war.
Example: *"There's a secret plan by the elites to control global resources."*
Example: *"The war is just a cover for something much bigger."*

A.2 Climate Change (CC)

1. **Criticism of climate policies:** Statements that question the effectiveness, economic impact, or motives behind climate policies.

   Sub-Narratives:

   (a) **Climate policies are ineffective:** Statements suggesting that climate policies fail to achieve their intended environmental goals.

   (b) **Climate policies have negative impact on the economy:** Statements claiming that climate policies lead to negative economic outcomes.

   (c) **Climate policies are only for profit:** Statements that argue climate policies are driven by financial or corporate gain rather than genuine environmental concerns.

2. **Criticism of institutions and authorities:** Statements that challenge the competence, integrity, or intentions of various institutions and authorities in relation to climate change.

   Sub-Narratives:

   (a) **Criticism of the EU:** Statements that express disapproval or distrust of the EU's role or approach to climate change or the EU in general.

   (b) **Criticism of international entities:** Statements that criticize the role and influence of international entities on climate policy.

   (c) **Criticism of national governments:** Statements that disapprove of the ways national governments handle climate change.

   (d) **Criticism of political organizations and figures:** Statements that discredit political organizations and figures in the context of climate change debate.

3. **Climate change is beneficial:** Statements that present arguments that support that changes in climate can have positive effects as well.

   Sub-Narratives:

   (a) **CO2 is beneficial:** Statements suggesting that increased CO2 levels have positive impacts on the environment.

   (b) **Temperature increase is beneficial:** Statements claiming that rising global temperatures can have positive effects.

4. **Downplaying climate change:** Statements that minimize the significance or impact of climate change.

   Sub-Narratives:

   (a) **Climate cycles are natural:** Statements suggesting that climate change is a natural and cyclical occurrence.

   (b) **Weather suggests the trend is global cooling:** Statements using local or short-term weather patterns to argue against global warming.

   (c) **Temperature increase does not have significant impact:** Statements claiming that the increase in temperature is not going to have any noticeable effect in nature.

   (d) **CO2 concentrations are too small to have an impact:** Statements claiming that the concentrations of CO2 will have a negligible effect.

   (e) **Human activities do not impact climate change:** Statements that support that climate change is not caused by human activity.

   (f) **Ice is not melting:** Statements claiming that there is not melting of ice.

   (g) **Sea levels are not rising:** Statements denying that sea levels have risen (or will rise).

(h) **Humans and nature will adapt to the changes:** Statements claiming that whatever the changes in climate humans or nature will manage to find solutions to adapt.

5. **Questioning the measurements and science:** Statements that raise doubts about the scientific methods, data, and consensus on climate change.

   Sub-Narratives:

   (a) **Methodologies/metrics used are unreliable/faulty:** Statements claiming that the scientific methodologies and metrics used to measure climate change are flawed or unreliable.

   (b) **Data shows no temperature increase:** Statements asserting that available data does not support the claim of global temperature increase.

   (c) **Greenhouse effect/carbon dioxide do not drive climate change:** Statements asserting that available data does not support the claim of global temperature increase.

   (d) **Scientific community is unreliable:** Statements discrediting scientists, the scientific community and their actions.

6. **Criticism of climate movement:** Statements that challenge the motives, integrity, or impact of the climate movement.

   Sub-Narratives:

   (a) **Climate movement is alarmist:** Statements suggesting that the climate movement exaggerates the severity of climate change for dramatic effect.

   (b) **Climate movement is corrupt:** Statements alleging that the climate movement is influenced by ulterior motives, by corruption or by unethical practices.

   (c) **Ad hominem attacks on key activists:** Statements attacking the reputation of key figures (such as scientists, activists, politicians or public figures).

7. **Controversy about green technologies:** Statements that express skepticism or criticism of environmentally friendly technologies.

   Sub-Narratives:

   (a) **Renewable energy is dangerous:** Statements claiming that renewable energy sources pose significant risks or dangers.

   (b) **Renewable energy is unreliable:** Statements asserting that renewable energy sources are not dependable for widespread adoption.

   (c) **Renewable energy is costly:** Statements asserting that renewable energy sources are too expensive, inefficient and worth adopting for widespread use.

   (d) **Nuclear energy is not climate friendly:** Statements asserting that nuclear sources are or should not be considered as good for the climate.

8. **Hidden plots by secret schemes of powerful groups:** Statements that propose secret plots or hidden agendas related to climate change initiated by powerful entities or groups.

   Sub-Narratives:

   (a) **Blaming global elites:** Statements attributing climate change agendas to secretive and powerful global elites.

   (b) **Climate agenda has hidden motives:** Claims that the push for climate action is driven by ulterior motives, such as political power or population control.

9. **Amplifying Climate Fears:** Statements that emphasize and amplify fears about the consequences of climate change.

   Sub-Narratives:

   (a) **Earth will be uninhabitable soon:** Statements predicting that the Earth will become uninhabitable in the near future due to climate change.

   (b) **Amplifying existing fears of global warming:** Statements that are using fears related to possible climate worries to spread panic.

   (c) **Doomsday scenarios for humans:** Statements presenting intense catastrophic scenarios as results of climate change.

   (d) **Whatever we do it is already too late:** Statements that minimize the urgency of addressing climate change by suggesting that any action taken at this point is futile or too late to make a meaningful impact.

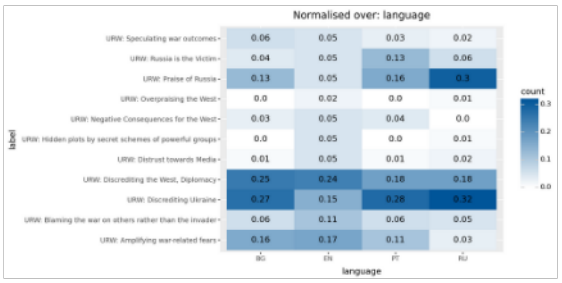

Figure 4: URW: Normalized count of coarse labels per language

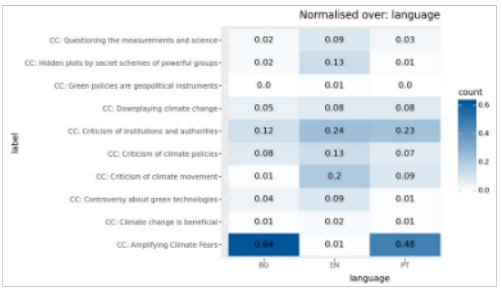

Figure 5: CC: Normalized count of coarse labels per language

10. **Green policies are geopolitical instruments:** Statements claiming that that environmental policies and initiatives are used as tools for geopolitical power and influence rather than genuine environmental concern.

    Sub-Narratives:

    (a) **Climate-related international relations are abusive/exploitative:** Statements criticizing international relations related to climate change as exploitative or economically abusive.

    (b) **Green activities are a form of neo-colonialism:** Statements suggesting that green initiatives are a way for developed countries to exert control and influence over developing nations, a modern form of colonial practices.

B Dataset statistics

Figures 4 and 5 report the normalized count of coarse labels per language. There are substantial differences, notably between RU and other languages for URW and between EN and other languages for CC. These differences could be due either to bias in media of each country, or be due to bias in the sampling of the articles to annotate. The reason RU is absent for CC is that it was not possible to find enough articles to annotate.

In Table 5, we report on the difficulty of annotating labels. We report the proportion of each label to have a disagreement between annotators both with the Other label specifically, or with any other label. Based on the numbers, we categorise each label in a class of difficulty.

C Prompts used for LLMs

In this Section we provide the different prompts used in the experiments.

C.1 0shot (only taxonomy)

You are an experienced analyst making labeling articles with labels specific to the war in Ukraine.

Labels:
{codebook_urw}

You will apply a label to each paragraph if it presents at any point a narrative that is given in the above list.

The beginning of each paragraph is marked by a [paragraph_X:] tag.

Answer strictly only with the paragraph tag and the name of the labels detected and not anything else. If you find no label, just output 'Other'. Strip down the markdown artifacts. Join the labels with semicolon.

Do not output anything from the text.

C.2 Guidelines

You are an experienced analyst labeling articles with labels specific ["to the war in Ukraine"/ "Climate Change"]
You will apply a label to each paragraph if it presents at any point a narrative that is given in the annotation guidelines given below.

{guidelines_urw}

The beginning of each paragraph is marked by a [paragraph_X:] tag.

Answer strictly only with the paragraph tag and the name of the labels detected and not anything else. If you find no label, just output 'Other'. Strip down the markdown artifacts. Join the labels with semicolon.

Do not output anything from the text.

C.3 Taxonomy and guidelines

You are an experienced analyst making labeling articles with labels specific to the war in Ukraine.

Labels:
{codebook_urw}

You will apply a label to each paragraph if it presents at any point a narrative that is given in the annotation guidelines given below.

{guidelines_urw}

The beginning of each paragraph is marked by a [paragraph_X:] tag.

Answer strictly only with the paragraph tag and the name of the labels detected and not anything else. If you find no label, just output 'Other'. Strip down the markdown artifacts. Join the labels with semicolon.
Do not output anything from the text.

C.4 Hierarchical prompting

First prompt:

You are an experienced analyst making labeling articles with labels specific to the war in Ukraine.

Labels:
{codebook_urw_coarse}

The beginning of each paragraph is marked by a [paragraph_X:] tag.

Answer strictly only with the paragraph tag and the name of the labels detected and not anything else. If you find no label, just output 'Other'. Strip down the markdown artifacts. Join the labels with semicolon.
Do not output anything from the text.

Second prompt:

You are an experienced analyst making labeling articles with labels specific to the war in Ukraine.

Labels:
{codebook_urw}

You are given the coarse label and you are asked to return the fine label.
You will apply a label if the article presents at any point a narrative that is given in the annotation guidelines given below.

{guidelines_urw}

The beginning of each paragraph is marked by a [paragraph_X:] tag.

Answer strictly only with the paragraph tag and the name of the labels detected and not anything else. If you find no label, just output 'Other'. Strip down the markdown artifacts. Join the labels with semicolon.
Do not output anything from the text.

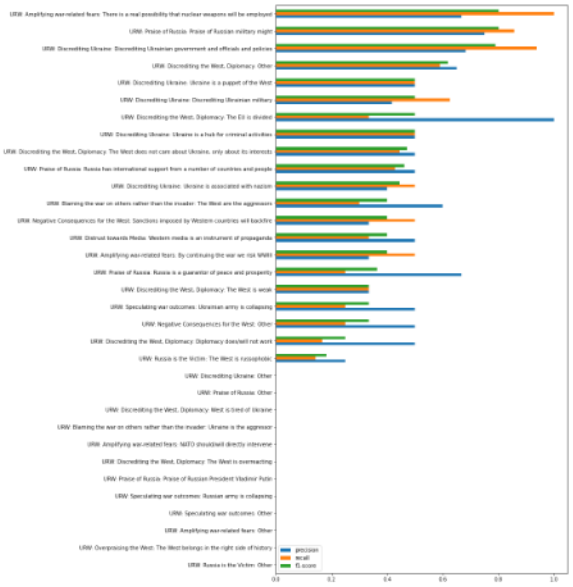

Figure 6: Performance of XLM-RoBERTa_fine on each sub-Narrative (fine) of the URW domain. Only labels with >2 support are shown

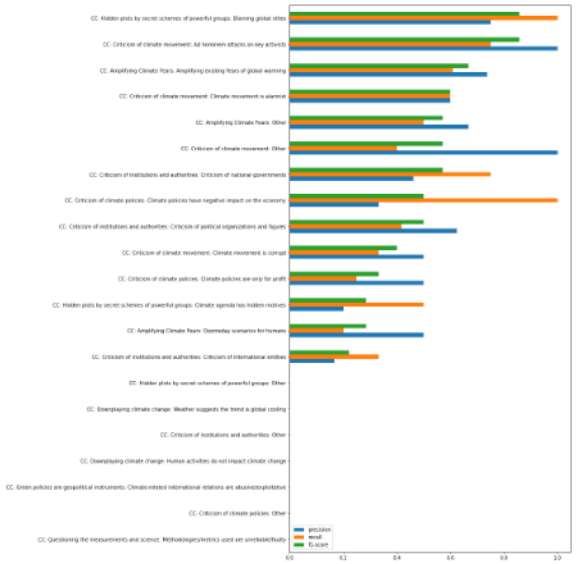

Figure 7: Performance of XLM-RoBERTa_fine on each sub-Narrative (fine) of the CC domain. Only labels with >2 support are shown

D Selected model performance per label and language

We provide detailed breakdown of the performance of our top-performing model on fine-grained level (sub-Narratives). We provide both the detailed scores per-label (Figures 6, 7) and per-language (Figures 8, 9, 10, 11) for the two domains (CC and URW).

E Breakdown on Annotator disagreements

We report on the detailed dissagreement statistics of the annotators per each fine-grained label, and highlight the difficulty of each label, in Table 5.

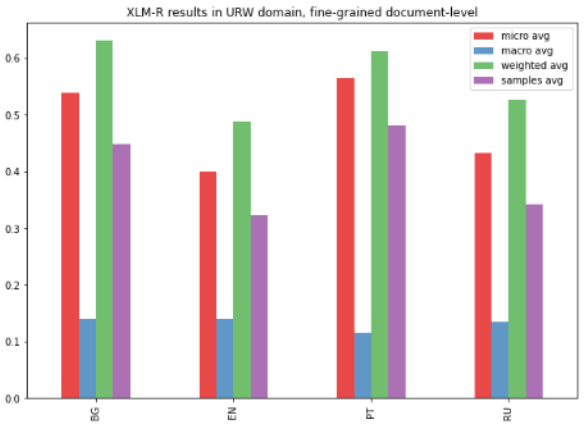

Figure 8: F1 scores of XLM-RoBERTa_fine on each language of the URW domain.

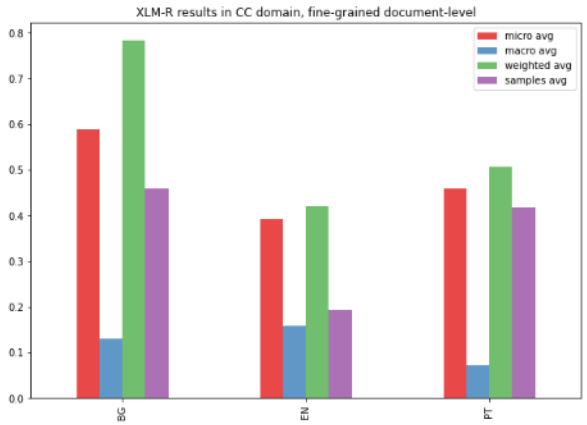

Figure 9: F1 scores of XLM-RoBERTa_fine on each language of the CC domain.

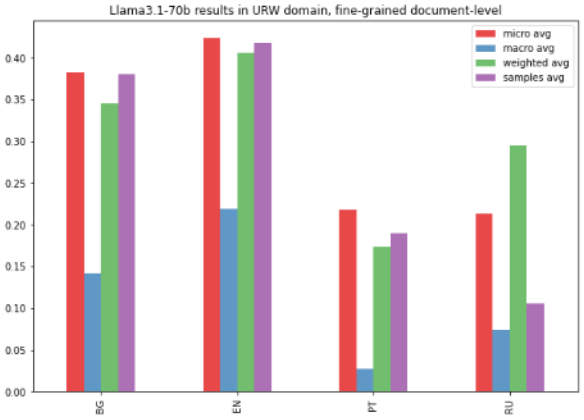

Figure 10: F1 scores of Llama-3.1-70b on each language of the URW domain.

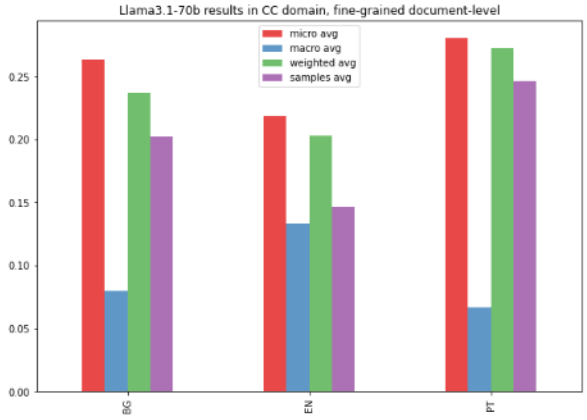

Figure 11: F1 scores of Llama-3.1-70b on each language of the CC domain.

label	count	%conf	%other	mcl	difficulty
1: URW: Blaming the war on others rather than the invader: Ukraine is the aggressor	169	0.309	0.315	5	Medium
2: URW: Blaming the war on others rather than the invader: The West are the aggressors	260	0.377	0.390	7	Medium
3: URW: Discrediting Ukraine: Rewriting Ukraine's history	23	0.272	0.111	28	Medium *
4: URW: Discrediting Ukraine: Discrediting Ukrainian nation and society	38	0.380	0.053	6	Medium *
5: URW: Discrediting Ukraine: Discrediting Ukrainian military	316	0.314	0.217	1	Medium *
6: URW: Discrediting Ukraine: Discrediting Ukrainian government and officials and policies	502	0.215	0.253	8	Easy
7: URW: Discrediting Ukraine: Ukraine is a puppet of the West	219	0.235	0.165	28	Easy *
8: URW: Discrediting Ukraine: Ukraine is a hub for criminal activities	127	0.368	0.243	6	Medium *
9: URW: Discrediting Ukraine: Ukraine is associated with nazism	97	0.189	0.080	6	Easy *
10: URW: Discrediting Ukraine: Situation in Ukraine is hopeless	107	0.303	0.220	6	Medium *
11: URW: Russia is the Victim: The West is russophobic	167	0.343	0.427	2	Medium
12: URW: Russia is the Victim: Russia actions in Ukraine are only self-defence	130	0.518	0.273	1	Hard *
13: URW: Russia is the Victim: UA is anti-RU extremists	34	0.477	0.406	11	Hard *
14: URW: Praise of Russia: Praise of Russian military might	466	0.140	0.346	5	Easy
15: URW: Praise of Russia: Praise of Russian President Vladimir Putin	100	0.157	0.348	17	Easy
16: URW: Praise of Russia: Russia is a guarantor of peace and prosperity	257	0.352	0.385	1	Medium
17: URW: Praise of Russia: Russia has international support from a number of countries and people	228	0.160	0.245	16	Easy
18: URW: Praise of Russia: Russian invasion has strong national support	20	0.333	0.250	17	Medium *
19: URW: Overpraising the West: NATO will destroy Russia	11	0.263	0.400	20	Medium
20: URW: Overpraising the West: The West belongs in the right side of history	36	0.263	0.133	28	Medium *
21: URW: Overpraising the West: The West has the strongest international support	24	0.254	0.308	7	Medium
22: URW: Speculating war outcomes: Russian army is collapsing	51	0.313	0.476	23	Medium
23: URW: Speculating war outcomes: Russian army will lose all the occupied territories	10	0.312	0.200	22	Medium *
24: URW: Speculating war outcomes: Ukrainian army is collapsing	93	0.242	0.118	5	Easy *
25: URW: Discrediting the West, Diplomacy: The EU is divided	105	0.164	0.435	28	Easy
26: URW: Discrediting the West, Diplomacy: The West is weak	145	0.301	0.267	14	Medium *
27: URW: Discrediting the West, Diplomacy: The West is overreacting	34	0.395	0.294	11	Medium *
28: URW: Discrediting the West, Diplomacy: The West does not care about Ukraine, only about its interests	211	0.381	0.262	7	Medium *
29: URW: Discrediting the West, Diplomacy: Diplomacy does/will not work	109	0.550	0.434	28	Hard *
30: URW: Discrediting the West, Diplomacy: West is tired of Ukraine	43	0.583	0.171	28	Hard *
31: URW: Negative Consequences for the West: Sanctions imposed by Western countries will backfire	79	0.338	0.723	12	Medium
32: URW: Negative Consequences for the West: The conflict will increase the Ukrainian refugee flows to Europe	12	0.000	NaN		Easiest
33: URW: Distrust towards Media: Western media is an instrument of propaganda	104	0.180	0.452	11	Easy
34: URW: Distrust towards Media: Ukrainian media cannot be trusted	23	0.500	NaN	6	Hard
35: URW: Amplifying war-related fears: By continuing the war we risk WWIII	135	0.285	0.339	37	Medium
36: URW: Amplifying war-related fears: Russia will also attack other countries	133	0.211	0.270	2	Easy
37: URW: Amplifying war-related fears: There is a real possibility that nuclear weapons will be employed	292	0.259	0.486	2	Medium
38: URW: Amplifying war-related fears: NATO should/will directly intervene	63	0.305	0.273	36	Medium *
39: CC: Criticism of climate policies: Climate policies are ineffective	74	0.407	0.174	44	Hard *
40: CC: Criticism of climate policies: Climate policies have negative impact on the economy	96	0.337	0.333	45	Medium *
41: CC: Criticism of climate policies: Climate policies are only for profit	58	0.508	0.069	68	Hard *
42: CC: Criticism of institutions and authorities: Criticism of the EU	54	0.275	NaN	45	Medium
43: CC: Criticism of institutions and authorities: Criticism of international entities	104	0.434	0.196	44	Hard *
44: CC: Criticism of institutions and authorities: Criticism of national governments	225	0.375	0.274	45	Medium *
45: CC: Criticism of institutions and authorities: Criticism of political organizations and figures	190	0.487	0.172	44	Hard *
46: CC: Climate change is beneficial: CO2 is beneficial	19	0.071	NaN	51	Easiest
47: CC: Climate change is beneficial: Temperature increase is beneficial	13	0.266	0.500	54	Medium
48: CC: Downplaying climate change: Climate cycles are natural	36	0.294	0.300	50	Medium
49: CC: Downplaying climate change: Weather suggests the trend is global cooling	13	0.406	0.308	70	Hard *
50: CC: Downplaying climate change: Temperature increase does not have significant impact	5	0.736	0.286	56	Hardest *
51: CC: Downplaying climate change: CO2 concentrations are too small to have an impact	18	0.500	0.143	59	Hard *
52: CC: Downplaying climate change: Human activities do not impact climate change	34	0.387	0.083	60	Medium *
53: CC: Downplaying climate change: Ice is not melting	18	0.117	NaN	62	Easy
54: CC: Downplaying climate change: Sea levels are not rising	2	0.500	0.500	47	Hard
55: CC: Downplaying climate change: Humans and nature will adapt to the changes	5	1.000	0.400	50	Hardest *
56: CC: Questioning the measurements and science: Methodologies/metrics used are unreliable/faulty	51	0.333	0.118	59	Medium *
57: CC: Questioning the measurements and science: Data shows no temperature increase	10	0.562	0.333	56	Hard *
58: CC: Questioning the measurements and science: Greenhouse effect/carbon dioxide do not drive climate change	5	0.750	1.000		Hardest
59: CC: Questioning the measurements and science: Scientific community is unreliable	45	0.540	0.185	56	Hard *
60: CC: Criticism of climate movement: Climate movement is alarmist	67	0.619	0.386	61	Hardest *
61: CC: Criticism of climate movement: Climate movement is corrupt	35	0.869	0.463	60	Hardest *
62: CC: Criticism of climate movement: Ad hominem attacks on key activists	78	0.234	0.217	60	Easy *
63: CC: Controversy about green technologies: Renewable energy is dangerous	18	0.421	0.250	65	Hard *
64: CC: Controversy about green technologies: Renewable energy is unreliable	42	0.344	0.333	45	Medium *
65: CC: Controversy about green technologies: Renewable energy is costly	27	0.512	0.100	45	Hard *
66: CC: Controversy about green technologies: Nuclear energy is not climate friendly	3	1.000	0.750	45	Hardest *
67: CC: Hidden plots by secret schemes of powerful groups: Blaming global elites	47	0.438	0.217	68	Hard *
68: CC: Hidden plots by secret schemes of powerful groups: Climate agenda has hidden motives	74	0.548	0.094	45	Hard *
69: CC: Amplifying Climate Fears: Earth will be uninhabitable soon	51	0.432	0.188	70	Hard *
70: CC: Amplifying Climate Fears: Amplifying existing fears of global warming	864	0.124	0.477	71	Easy
71: CC: Amplifying Climate Fears: Doomsday scenarios for humans	104	0.435	0.239	70	Hard *
72: CC: Amplifying Climate Fears: Whatever we do it is already too late	18	0.264	0.111	70	Medium *
73: CC: Green policies are geopolitical instruments: Climate-related international relations are abusive/exploitative	7	0.785	0.273	68	Hardest *
74: CC: Green policies are geopolitical instruments: Green activities are a form of neo-colonialism	10	0.375	0.333	68	Medium *

Table 5: Label difficulty, %conf is the percentage of annotation with that label that results in an inconsistency, while %other is the percentage that resulted in an inconsistency with label other. "mcl" is the most common label with which a specific label is confused. "difficulty" make assess the difficulty of the label based on %conf: Easiest ($\leq .1$), Easy ($\leq .25$), Medium ($\leq .4$), and Hard ($\leq .4$) or Hardest (> 0.6), a star indicate that there is more confusion within the labels of the taxonomy than with the Other class, meaning that it is extra difficult for the annotator