Entity Framing and Role Portrayal in the News

Tarek Mahmoud[1], Zhuohan Xie[1], Dimitar Dimitrov[2], Nikolaos Nikolaidis[3],
Purificação Silvano[4], Roman Yangarber[5], Shivam Sharma[6], Elisa Sartori[7],
Nicolas Stefanovitch[8], Giovanni Da San Martino[7], Jakub Piskorski[9], Preslav Nakov[1]

[1]MBZUAI, UAE   [2]Sofia University "St. Kliment Ohridski", Bulgaria
[3]Athens University of Economics and Business, Greece   [4]University of Porto, Portugal
[5]University of Helsinki, Finland   [6]Indian Institute of Technology Delhi, India
[7]University of Padova, Italy   [8]European Commission Joint Research Centre, Italy
[9]Institute of Computer Science, Polish Academy of Sciences, Poland
{tarek.mahmoud, preslav.nakov}@mbzuai.ac.ae

Abstract

We introduce a novel multilingual hierarchical corpus annotated for entity framing and role portrayal in news articles.[1] The dataset uses a unique taxonomy inspired by storytelling elements, comprising 22 fine-grained roles, or archetypes, nested within three main categories: *protagonist*, *antagonist*, and *innocent*. Each archetype is carefully defined, capturing nuanced portrayals of entities such as guardian, martyr, and underdog for protagonists; tyrant, deceiver, and bigot for antagonists; and victim, scapegoat, and exploited for innocents. The dataset includes 1,378 recent news articles in five languages (Bulgarian, English, Hindi, European Portuguese, and Russian) focusing on two critical domains of global significance: the Ukraine–Russia War and Climate Change. Over 5,800 entity mentions have been annotated with role labels. This dataset serves as a valuable resource for research into role portrayal and has broader implications for news analysis. We describe the characteristics of the dataset and the annotation process, and we report evaluation results on fine-tuned state-of-the-art multilingual transformers and hierarchical zero-shot learning using LLMs at the level of a document, a paragraph, and a sentence.

1 Introduction

The rapid proliferation of massive amounts of content on social media has dramatically transformed the information landscape, providing immediate access to news, and allowing anyone to propagate their narratives across the globe. While this connectivity functions as a convenient avenue for information dissemination in a very democratic way, it also heightens the risk of online users of being exposed to biased reporting, propaganda, and narrative manipulation.

These risks are particularly pronounced during periods of conflict and political upheavals, where the framing of entities—individuals, organizations, or groups—can profoundly influence public perception and decision-making. Understanding how entities are portrayed in the news is essential for fostering media literacy, identifying bias, and ensuring transparent news consumption.

Social science highlights the role of emotion in framing—selecting elements that evoke affective responses to shape perceptions (Gross and D'Ambrosio, 2004). Emotional framing often leverages language that elicits specific feelings, such as fear, anger, or compassion, to influence how entities are understood (Iyengar, 1991; Nabi, 2003; Semetko and Valkenburg, 2000; Brader, 2006). For instance, referring to a group as "freedom fighters" versus "terrorists" not only frames their role but also activates distinct emotional reactions. Research has shown that emotional appeals are powerful tools for reinforcing or challenging public attitudes (Westen, 2008; Lerner et al., 2015). Such framing can manifest through specific linguistic cues and includes the portrayal of entities also defined by Schneider and Ingram (1993) as the *Social Construction of Target Populations*.

Natural language processing research has increasingly been applied to analyze the emotional dimensions of framing (Troiano et al., 2023), including the identification of sentiment (Zhang et al., 2024; Tan et al., 2023) and emotion-laden narratives (Mousavi et al., 2022) in text. Understanding these emotional components provides deeper insight into how media narratives construct and perpetuate particular representations of entities, which ultimately shapes public perception and societal discourse.

Given the large scale and complexity of the modern news ecosystem, effective analysis of entity framing requires automated tools, which depend on high-quality annotated data.

---

[1]Our dataset is available at
https://mbzuai-nlp.github.io/entity-framing/

Figure 1: Annotated example color-coded according to the main roles in the taxonomy: red for *antagonist*, blue for *protagonist*, and green for *innocent*.

In this context, we introduce a new multilingual dataset designed to develop tools for the study of entity framing and role portrayal in news articles. Our dataset uses a unique, hierarchical taxonomy inspired by elements of storytelling containing a set of 22 carefully defined archetypal roles nested under three main frames: *protagonist*, *antagonist*, and *innocent*.

The corpus spans 1,378 recent news articles in five languages (Bulgarian, English, Hindi, European Portuguese, and Russian) and focuses on two globally significant domains: the Ukraine-Russia war and climate change. We annotated over 5,800 entity mentions with detailed role labels.

Our contributions can be summarized as follows:

- We release a novel multilingual dataset annotated for entity framing and role portrayal, complete with detailed annotation guidelines.

- We introduce a comprehensive hierarchical taxonomy for entity roles validated on a large set of documents, supporting analysis at both the coarse and the fine-grained levels.

- We provide comprehensive dataset statistics, and analyze co-occurrence of roles as well as role transitions.

- We set benchmarks using state-of-the-art multilingual transformer models, and hierarchical zero-shot learning with LLMs.

2 Related Work

2.1 A Hierarchical Taxonomy of Entity Roles

The identification of narrative frames such as heroes, villains, and victims has been studied across various modalities and domains. Sharma et al. (2023) introduced a dataset for identifying heroes, villains, and victims in memes, focusing on visual content. Similarly, (Stammbach et al., 2022) demonstrated that LLMs can detect these frames from narrative texts using zero-shot question-answering prompts across diverse domains such as newspaper articles, movie plot summaries, and political speeches. Frermann et al. (2023) integrated elements like conflict and resolution with the framing of key entities as heroes, victims, or villains.

These studies primarily focus on the coarse-grained categorization of entities into the hero-villain-victim frames. Bergstrand and Jasper (2018) analyze this triad and show how these frames are imbued with moral attributes where heroes are shown as benevolent and strong, villains as malevolent and strong, and victims as benevolent and weak. In contrast, our framework redefines these frames based on narrative function, rather than moral judgment. By introducing the frames of protagonist, antagonist, and innocent, we focus on the functional roles entities occupy within a narrative, independent of their moral aspect.

303

This approach allows for a more nuanced and flexible analysis of entity roles. We further present a hierarchical taxonomy of 22 roles nested within these narrative functional frames. This rich set of roles captures the diverse ways entities are portrayed in textual narratives. Our taxonomy was validated through human annotations across diverse news domains.

2.2 The Entity Framing Task Formulation

Aspect-Based Sentiment Analysis (ABSA) Prior work on ABSA and entity-level sentiment analysis, which is a special case of ABSA and the most relevant one to our research, explores both the assignment of sentiment to particular attributes of entities and to entities as a whole (Saeidi et al., 2016; Jiang et al., 2019; Orbach et al., 2021; Zhang et al., 2023; Chebolu et al., 2024; Kuila and Sarkar, 2024; Bastan et al., 2020; Rønningstad et al., 2022). For instance, Tang et al. (2023) presented a dataset for entity-level sentiment classification in financial texts. In contrast, our task formulation differs by not assigning polarities or defining aspects; instead, we classify entities into a rich set of roles inspired by storytelling elements.

Article-level framing Card et al. (2016) addressed article-level framing by developing a model that utilizes personas to infer framing dimensions defined in the Media Frames Corpus (Card et al., 2015). Their topic modeling approach identified 50 personas, but only a few, such as "refugee" and "immigrant," were found to be informative. Further research on news framing has predominantly focused on article-level analysis. Studies like Liu et al. (2019); Piskorski et al. (2023); Otmakhova et al. (2024); Pastorino et al. (2024) have explored various dimensions of framing in news articles. However, our approach centers on the framing of individual entities within the text, not on article-level framing.

3 The Entity Framing Task

Entity framing focuses on analyzing how a text portrays a specific entity through word choice and narrative structure. More concretely, given a news article and a list of entity mentions (i.e., entity mentions, along with their span offsets), we assign to each of them one or more roles based on the taxonomy shown in Figure 2. We developed this taxonomy specifically for this task and the roles were inspired by storytelling elements.

The taxonomy includes 22 archetypes, or fine-grained roles nested under three main frames: *protagonist*, *antagonist*, and *innocent*. The role an entity plays in a given article may differ from one context in that article to another depending on the portrayal. See Figure 1 for a complete, annotated example from the corpus.

Entity framing can be formalized mathematically as follows. Let R be a tree structure representing the taxonomy of roles. Let S be a string of length $|S|$ characters with the content of the full article. The goal of entity framing is to learn a function

$$f : (S, [i, j]) \rightarrow \{r_1, r_2, \ldots, r_k\} \subseteq R \quad (1)$$

where $0 \leq i < j \leq |S|$ and $\{r_1, r_2, \ldots, r_k\}$ is the set of roles assigned to the span $[i, j]$.

PROTAGONIST

Guardian: Heroes or guardians who protect values or communities, ensuring safety and upholding justice.
Martyr: Individuals who sacrifice their well-being, or even their lives, for a greater good or cause.
Peacemaker: Individuals who advocate for harmony, resolving conflicts and bringing about peace.
Rebel: Revolutionaries who challenge the status quo and fight for significant change or liberation.
Underdog: Entities who, despite a disadvantaged position, strive against greater forces and obstacles.
Virtuous: Individuals portrayed as righteous, fair, and upholding high moral standards.

ANTAGONIST

Instigator: Those who initiate conflict and provoke violence or unrest.
Conspirator: Individuals involved in plots and covert activities to undermine or deceive others.
Tyrant: Leaders who abuse their power, ruling unjustly and oppressing others.
Foreign Adversary: Entities from other nations creating geopolitical tension and acting against national interests.
Traitor: Individuals who betray a cause or country, seen as disloyal and treacherous.
Spy: Individuals engaged in espionage, gathering and transmitting sensitive information.
Saboteur: Those who deliberately damage or obstruct systems to cause disruption.
Corrupt: Individuals or entities engaging in unethical or illegal activities for personal gain.
Incompetent: Entities causing harm through ignorance, lack of skill, or poor judgment.
Terrorist: Individuals who engage in violence and terror to further ideological ends.
Deceiver: Manipulators who twist the truth, spread misinformation, and undermine trust.
Bigot: Individuals accused of hostility or discrimination against specific groups.

INNOCENT

Forgotten: Marginalized groups who are overlooked and ignored by society.
Exploited: Individuals or groups used for others' gain, often without consent.
Victim: People suffering harm due to circumstances beyond their control.
Scapegoat: Entities unjustly blamed for problems or failures to divert attention.

Figure 2: The hierarchical taxonomy of roles for entity framing. A more comprehensive description accompanied with examples is provided in Appendix A.

4 Corpus Description

4.1 Domains

The articles used in the task cover the following domains: (1) *Ukraine-Russia War*, which includes articles about the war that started in February 2022 when Russia launched a full-scale invasion of Ukraine and began occupying parts of the country, and (2) *Climate Change*, which encompasses both climate change denial (characterized by rejecting, refusing to acknowledge, disputing, or fighting the scientific consensus on climate change), and climate change activism.

4.2 Article Selection

For each language, articles were primarily selected from links we obtained from a large-scale in-house news aggregation tool. We performed the first candidate article selection based on multilingual keyword-based filters and we perform several steps, which we follow by to enrich the selection to match the criteria discussed below. To select the articles, we followed these steps:

Initial Collection: Articles were scraped and filtered based on criteria such as word count (e.g., only articles exceeding 250 words were selected). For duplicate articles, the version with the higher number of words was preferred.

Filtering: Each article was manually reviewed to determine its relevance to the annotation task. The articles were categorized into four groups: Perfect Fit, Average Fit, Uncertain (requiring further validation by language coordinators), or Unfit (excluded from annotation). Only articles classified as *Perfect Fit* and *Average Fit* were considered for annotation. Afterwards, we used a zero-shot classifier with selected key phrases, and a persuasiveness score using the persuasion technique classifier as described in Nikolaidis et al. (2024). These scores were used to further enrich the selection with relevant articles.

Additional sources were also incorporated to ensure diversity of perspectives for two languages: for Hindi, we selected articles from mainstream and bias-specific outlets (e.g., NDTV, The Hindu, OpIndia), and for Portuguese, from newspapers and political websites (e.g., *O Diabo*, Esquerda, Folha Nacional) that had more controversial opinion texts about the relevant topics.

4.3 Annotation Process

Given that our corpus contains articles in five languages, we had separate annotation teams, one for each language, each led by a language coordinator. Each language team included 3 to 5 annotators with prior experience in linguistics, social science, international relations, or prior annotation work. The annotators studied our detailed guidelines, attended live demonstrations, and completed real-time annotation exercises. During weekly meetings, teams clarified any uncertainties, resolved conflicts, improved consistency, and revised the annotation guidelines.

Each article was annotated by two annotators. Designated curators, often language coordinators or experienced annotators, reviewed and consolidated all annotations. They resolved the discrepancies through discussions with the respective teams. Over time, as the annotation quality improved, the curators reduced the frequency of checks but continued to perform random quality checks to ensure that annotations were of high quality. We used the Inception tool (Klie et al., 2018) for annotating and curating the corpus. See Appendix F for more details.

The annotation guidelines (Appendix B) were refined during initial weekly meetings between language coordinators and annotators. From these guidelines, we emphasize key aspects of entity selection for annotation. We annotated traditional named entities, extending this to also include eponym-derived entities (e.g., *Putin supporters*) and toponym-derived entities (e.g., *Western countries*, as illustrated in Figure 1). Additionally, we focused on annotating entities that are central to the narrative conveyed by the article. For example, in Figure 1, entities such as *Ulan-Ude Aviation Plant* and *Rossiya 24 TV* were not annotated because they were not considered central to the narrative. For a detailed explanation of how centrality was defined, refer to the guidelines in Appendix B.

4.4 Inter-Annotator Agreement

To assess the inter-annotator agreement (IAA), we used Krippendorff's alpha. We compared the annotations at the span level: they were approximately matched if they shared at least 50% of their length, to account for minor differences. The IAA values are shown in Table 2. The results indicate a moderate agreement (above 0.45), which we consider acceptable due to the span-based nature of the task.

Lang.	#DOC	#PAR	#SEN	#WORD	#CHAR	AVG_p	AVG_s	AVG_w	AVG_c	#ENT	#ANN	AVG_e	AVG_a
BG	274	2.7K	5.0K	104K	584K	9.8	18.5	380.6	2129.8	656 (179)	742	2.4	2.7
EN	229	3.7K	5.3K	131K	705K	16.2	23.1	571.3	3080.7	775 (413)	843	3.4	3.7
HI	377	3.8K	8.7K	200K	947K	10.2	23.0	530.1	2513.0	2,609 (724)	3,030	6.9	8.0
PT	337	3.5K	5.4K	150K	822K	10.5	15.0	445.0	2439.2	1,365 (440)	1,438	4.1	4.3
RU	161	0.7K	2.0K	42K	257K	4.5	12.7	261.1	1599.1	451 (265)	477	2.8	3.0
Total	1,378	14.5K	26.4K	627K	3,316K	10.5	19.2	455.0	2406.3	5,856 (1,962)	6,530	4.2	4.7

Table 1: Corpus statistics showing total number of documents (#DOC), paragraphs (#PAR), sentences (#SEN), words (#WORD), and characters (#CHAR) by language. The averages (AVG_p), (AVG_s), (AVG_w), and (AVG_c) refer to the average number of paragraphs, sentences, words, and characters per document, respectively. The table also shows the total number of annotated entity mentions (#ENT) accompanied with unique counts in parentheses, the total number of annotations (#ANN), the average number of entity mentions per document (AVG_e), and the average number of annotations per document (AVG_a).

Lang.	EN	RU	BG	PT	HI	All
α	.460	.436	.733	.467	.461	.558

Table 2: Inter-annotator agreement computed using Krippendorff's α.

Freq.	1	2-5	5-10	10-20	20-50	50-500
count (%)	1513 (74)	374 (18)	76 (4)	46 (2)	22 (1)	14 (1)

Table 3: Proportion of entities of a given frequency.

We can see that the IAA is similar across the languages, except for Bulgarian, for which it is notably higher (0.73), which can be explained by the low count of entities in Table 1.

4.5 Corpus Analysis

4.5.1 Statistics

Table 1 provides overall statistics about the corpus, including a breakdown per language, the number of annotated entity mentions, the number of unique entities, as well as the number of annotations. Figure 4 displays the distribution of the main and fine-grained roles in the corpus. We can see that there is a significant imbalance between the fine-grained roles, while the distribution of the main roles is relatively balanced. Notably, within the *innocent* category, the majority of the instances, 83.6%, are labeled as *victim*, with fewer occurrences of *exploited*, *forgotten*, and *scapegoat* roles. More details about the proportions of main roles and fine-grained roles across different languages can be found in Figure 8 in Appendix D.

Table 3 shows the number and proportion of entities within a given frequency range based on exact string matching. We can see that 74% of the entities were annotated only once, while 14 entities were annotated more than 50 times. These numbers only reflect surface string, not accounting for grammatical and name variants, or languages.
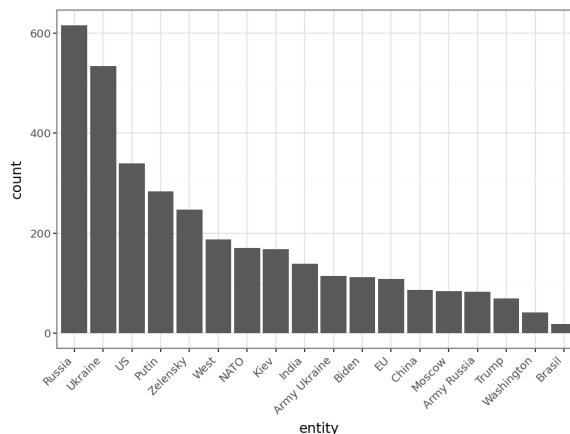

Figure 3: Top entities' counts after multilingual linking was manually performed to link surface string to unique identifiers.

In Figure 3, we match the most frequent entities accounting for these differences and present detailed statistics at the level of entities. The entities selected are all the ones for which at least one surface string has a count of at least 10 in any language.

4.5.2 Co-occurrence of Roles

In our definition of entity framing, entity mentions can be assigned one or more fine-grained roles. Our corpus contains on average 1.12 annotations per entity mention. Out of the 1,378 articles, 353 articles contain 638 instances where at least one entity mention has multiple annotations. For this set of articles, the average and the maximum are 2.1 and 3 annotations per entity mention, respectively. We further observe that roles such as *peacemaker* frequently accompany *guardian*. Similarly, entities portrayed as *scapegoats* are often framed as being *exploited*. More details about the co-occurrence matrix for entity mentions with multiple annotations are shown in Figures 5 and 6 in Appendix D.
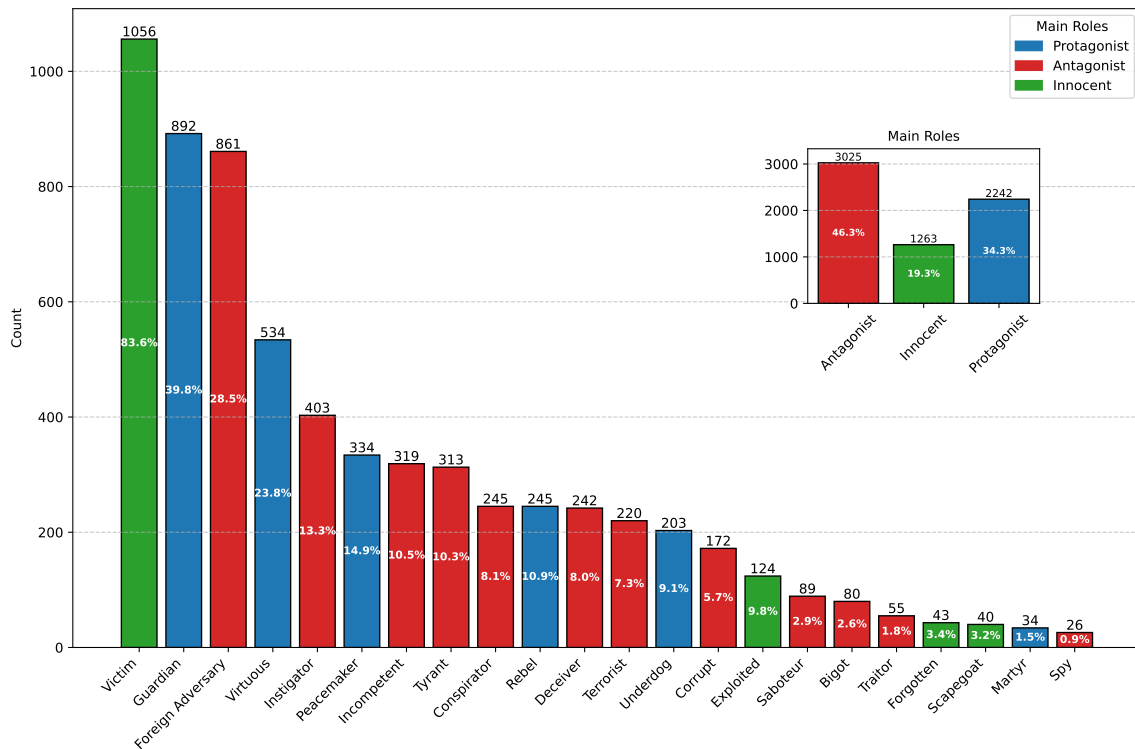
Figure 4: Distribution of fine-grained roles color-coded according to the main role. For the fine-grained roles, the percentages inside the bars indicate the proportion of each fine-grained role relative to its corresponding main role category. The counts on top of the bars show the total occurrences of each fine-grained role. In the mini histogram, the percentages inside the bars reflect the distribution of the main roles, with the counts displayed above the bars.

4.5.3 Analysis of Role Transitions

We analyze role transitions of an entity within the same article, and observed that out of 1,378 articles, only 99 contained occurrences with main role changes for the same entity, and 261 contained fine-grained role changes.

While the most common main role transitions are between protagonist and antagonist roles, fine role transitions are more nuanced. Examples include sequences like **victim** to **underdog** to **virtuous** and transitions from **rebel, and virtuous** to **exploited and victim**. We study the transition sequences for both main and fine-grained roles, and present the ten most frequent transitions sequences in Figure 7 in Appendix D.

Our annotation guidelines account for such instances. While such cases are relatively infrequent, they highlight the complexity of entity framing and the importance of capturing role transitions within our taxonomy. This analysis also reinforces the value of our dataset for studying such nuanced challenges.

5 Experiments

We experimented with classifying entities into main roles and fine-grained roles. As we performed the entity framing annotations at the span level, we framed the problem as a multi-class multi-label classification task. Given an article, an entity mention, and the span offsets, the goal was to predict the framing of that entity mention. We provide two sets of baselines and experiments to benchmark state-of-the-art models, as well as to assess the complexity of the entity framing task. The first set evaluates fine-tuning multilingual transformer models in various settings, while the second set explores hierarchical zero-shot learning using LLMs. To prevent data leakage, we created train/dev/test splits based on entire articles rather than individual entity-mention annotations. The details of these splits are provided in Table 4.

5.1 Fine-Tuning Pre-trained Multilingual Transformers

For the first set of experiments, we designed a series of experiments to address the following important aspects:

	BG	EN	HI	PT	RU	All
Train	165 (389)	133 (440)	203 (1347)	206 (833)	89 (252)	796 (3261)
Dev	94 (237)	69 (245)	139 (983)	100 (417)	44 (114)	446 (1996)
Test	15 (30)	27 (90)	35 (279)	31 (115)	28 (85)	136 (599)
Total	274 (656)	229 (775)	377 (2609)	337 (1365)	161 (451)	1378 (5856)

Table 4: Distribution of articles and entity mentions by language and split. The number of entity mentions is shown in parentheses

- **Granularity of context**–predicting role labels for entity mentions using the full document or narrowing the model's focus to only look at the pertinent paragraph or sentence containing the entity of interest.

- **Multilingual** comparison of the performance in the monolingual setting versus the multilingual setting trained on data in all five languages: Bulgarian, English, Hindi, European Portuguese, and Russian.

For both granularity-level classification and multilingual performance, we made predictions at two levels: main role (3 labels) and fine-grained role (22 labels). We fine-tuned the multilingual pre-trained transformer XLM-R (Conneau et al., 2020) and adapted its final layers for our tasks, applying $softmax$ for multiclass classification and $sigmoid$ for multi-label classification. To handle spans within potentially long documents, we addressed the 512-token limitation of XLM-R by narrowing the context to the paragraph or the sentence where the entity appeared. We constructed the input text using the following format:

```
input = entity mention + [SEP] + title
+ [SEP] + context.
```

In this setup, [SEP] is the separator token, and the context can vary based on the granularity level, ranging from the full text to just the paragraph or sentence containing the entity mention. We placed the entity mention first, followed by the title and context, to maintain consistent positional encodings for the entity mention across different inputs. We used Stanza (Qi et al., 2020) for sentence splitting for all languages.

To further support span-level multi-label classification, we modified the output layer to include a $sigmoid$ activation and optimized the model using *Binary Cross-Entropy loss*. This setup allowed the model to predict multiple overlapping roles for a given entity span. See Appendix C for more details.

5.2 Hierarchical Zero-Shot Learning with LLMs

We experimented with two prompting approaches: *single-step* and *hierarchical multi-step*. The former aimed to predict both the main role and the fine-grained role simultaneously within a single prompt. It assumed that both tasks can be handled together, relying on the model's ability to process them in one go. On the other hand, the multi-step approach separated the prediction into two distinct stages. First, the main role was predicted, and based on that output, the fine-grained role was predicted in a second step, using the information from the initial prediction to refine the second task. This stepwise process involved an intermediate prediction, which allowed the model to focus on each task individually. See Appendix C, and E for more details on experimental settings and prompt structure.

5.3 Results

We report standard evaluation metrics, including *micro-average precision*, *recall*, and $F1$ score, along with the *macro-average* $F1$ score for fine-grained roles to address class imbalance. We further provide *accuracy* and *balanced accuracy* for predictions on the main role granularity.

Table 5 shows the performance of XLM-R across different context granularities (document, paragraph, and sentence) and training configurations (main roles vs. fine-grained roles). For models trained and evaluated on main roles, paragraph-level contexts perform best, followed closely by sentence-level contexts, while document-level contexts perform the worst. When models trained on fine-grained roles are evaluated on main roles, paragraph-level contexts again yield the best performance, with document and sentence-level contexts slightly behind. This indicates that training on fine-grained roles provides an advantage. For models trained and evaluated on fine-grained roles, sentence-level contexts perform best, followed by paragraph-level contexts, with document-level contexts showing the weakest performance. These results highlight that context granularity significantly impacts performance, with localized contexts outperforming document-level contexts for both main role and fine-grained role classification tasks.

Table 6 offers interesting insights into the performance of XLM-R when fine-tuned on monolingual and multilingual data for multi-label fine-grained role classification at the paragraph level.

Train	Context	Main Role		Fine Grained Role			
		Accuracy	Balanced Accuracy	P	R	Micro F1	Macro F1
	DOC	.6010	.5904	–	–	–	–
M	PAR	.7379	.7385	–	–	–	–
	SEN	.7179	.7123	–	–	–	–
	DOC	.7229	.7235	.3495	.4446	.3913	.2306
F	PAR	**.7529**	**.7553**	.3649	.4985	.4213	.2392
	SEN	.7496	.7503	**.4195**	**.4492**	**.4339**	**.2529**

Table 5: Performance of entity framing across different granularity settings using XLM-R trained on the full multilingual dataset. Models are trained and evaluated on texts with varying context sizes: full document (DOC), paragraph (PAR), or sentence (SEN) containing the entity mention. The results cover models trained on main roles (M), fine-grained roles (F), and evaluated on either main roles, fine-grained roles, or both.

(a) Monolingual setting

Lang.	P	R	Micro F1	Macro F1
EN	.1032	.1313	.1156	.0435
BG	.1056	.5758	.1784	.0505
HI	.3424	.4495	.3887	.1740
PT	.6124	.6423	.6270	.1505
RU	.1077	.5227	.1786	.0437

(b) Multilingual setting

Lang.	P	R	Micro F1	Macro F1
All	.3649	.4985	.4213	.2392
EN	.1854	.2828	**.2240**	**.1327**
BG	.3030	.3030	**.3030**	**.1349**
HI	.3234	.4951	**.3912**	**.2043**
PT	.6259	.7480	**.6815**	**.2040**
RU	.4831	.4886	**.4859**	**.2364**

Table 6: Results for multi-label fine-grained role classification with XLM-R trained on monolingual and multilingual data and evaluated at the paragraph level.

The monolingual setting exhibits varying performance across languages, with the highest scores achieved for Portuguese, while English, Bulgarian, and Russian show notably lower performance; the model's performance on Hindi is moderate. These differences stem from the quantity and quality of training data, linguistic variations, and the complexity of entity mentions across languages. The consistently low Macro F1 scores across all languages indicate difficulty in predicting rare roles. In contrast, the multilingual setting consistently outperforms the monolingual setting, demonstrating its ability to better capture diverse fine-grained roles through cross-lingual transfer.

Table 7 consolidates the results from fine-tuning XLM-R and hierarchical zero-shot learning, showing that the multi-step approach achieves slightly better performance for main role prediction compared to the single-step approach. This can be attributed to the structured decomposition of the task.

By explicitly isolating the main role prediction into a separate dedicated step, the model can focus on high-level role identification without the distraction of fine-grained details. While this multi-step approach enhances performance for main role prediction, it underperforms compared to the single-step approach for predicting fine-grained roles. We hypothesize two potential reasons for this discrepancy:

1. **Error Propagation:** In the multi-step approach, the model first predicts the main role and then proceeds to predict fine-grained roles. Errors introduced during the main role prediction step can propagate to subsequent steps, thereby reducing the overall accuracy of fine-grained predictions.

2. **Loss of Joint Context:** The single-step approach enables the model to reason jointly about both main roles and fine-grained roles, allowing it to better capture dependencies between role labels. This integrated reasoning leads to more precise and consistent fine-grained predictions. Another finding is that the multi-step approach is significantly more cost-effective than the single-step approach. This efficiency may stem from token efficiency, as multi-step prompts are designed to be more concise, with each step addressing a specific sub-task (e.g., main role followed by fine-grained role). Consequently, this results in fewer tokens per prompt.

Table 7 additionally compares the performance of zero-shot approaches and XLM-R, both using similarly sized input contexts. We can see that XLM-R outperforms zero-shot methods on all evaluation measures except for Macro F1, where it shows the lowest performance among all approaches we experimented with.

Method	Lang.	Main Role		Fine Grained Role				Cost (USD)
		Accuracy	Balanced Accuracy	P	R	Micro F1	Macro F1	
Single-Step LLM Prompting	EN	.8346	.6756	.2692	.4632	.3405	.2171	0.7989
	BG	.8065	.7380	.3725	.5588	.4471	.3481	0.2751
	HI	.6327	.6247	.2753	.4000	.3262	.2196	2.4696
	PT	.7812	.7455	.5167	.6643	.5813	.2891	1.0200
	RU	.7558	.6719	.3939	.5843	.4706	.4644	0.7587
	All	.7030	.6957	.3211	.4726	.3824	.3103	5.3223
Multi-Step LLM Prompting	EN	.8031	.6799	.2887	.4118	.3394	.2383	0.5130
	BG	.8031	.6799	.4318	.5588	.4872	.3601	0.5130
	HI	.6367	.6284	.2676	.2868	.2769	.1771	1.4581
	PT	.8125	.7882	.3895	.2643	.3149	.2498	0.5634
	RU	.7442	.6680	.4118	.4719	.4398	.3774	0.4769
	All	.7053	.7017	.3051	.3294	.3168	.2765	**3.1852**
XLM-R	EN	0.6889	0.5276	.1854	.2828	.2240	.1327	–
	BG	0.7333	0.5791	.3030	.3030	.3030	.1349	–
	HI	0.7025	0.7046	.3234	.4951	.3912	.2043	–
	PT	0.8957	0.8840	.6259	.7480	.6815	.2040	–
	RU	0.8000	0.7604	.4831	.4886	.4859	.2364	–
	All	**0.7529**	**0.7553**	**.3649**	**.4985**	**.4213**	.2392	–

Table 7: Consolidated results comparing fine-tuning XLM-R and zero-shot learning with GPT-4o. The table shows performance and cost comparisons between single-step and multi-step LLM prompting approaches, where the highest scores between these two approaches across all languages are underlined. The top results across all three methods and languages are highlighted in **bold**.

We hypothesize that this discrepancy arises because XLM-R had only seen a limited number of training instances for the rare roles, which makes it very hard for the model to learn to detect these categories. In contrast, zero-shot approaches do not rely on training data and thus are not constrained by this limitation.

This hypothesis is based on observable patterns in the dataset and model performance. Figure 4 illustrates the fine-grained role distribution, revealing a significant imbalance, where we can see that certain roles are severely underrepresented. This imbalance intuitively contributes to the observed discrepancy between Micro and Macro F1 scores, as XLM-R's fine-tuning relies on sufficient training instances for rare roles, which are limited in this dataset. This observation is consistent with previous findings in NLP (Henning et al., 2023), where class imbalance is known to affect the performance for low-frequency classes.

In contrast, zero-shot approaches do not depend on labeled training data, which makes them less sensitive to class imbalance, which can explain their comparatively higher Macro F1 scores. While factors such as model architecture or prompt design could also influence performance, dataset imbalance remains a key explanatory factor supported by our analysis. Thus, while XLM-R achieves a higher Micro F1 score, the higher Macro F1 score that we observe for zero-shot approaches offers an intuitive explanation tied to the distribution of roles in the dataset.

6 Conclusion and Future Work

We presented a novel multilingual and hierarchical dataset for characterizing entity framing and role portrayal in news articles. The dataset introduces a unique taxonomy inspired by storytelling elements, featuring 22 fine-grained archetypes nested within three main categories: *protagonist*, *antagonist*, and *innocent*. The dataset covers 1,378 recent news articles in five languages (Bulgarian, English, Hindi, European Portuguese, and Russian), spanning two domains: the Ukraine–Russia War and Climate Change. Over 5,800 entity mentions have been thoroughly annotated with role labels, capturing nuanced portrayals. We evaluated the dataset using fine-tuned state-of-the-art multilingual transformers and explored hierarchical zero-shot learning at the document, paragraph, and sentence level. Our experiments highlighted the potential of multilingual representations and hierarchical approaches for entity-framing tasks. We release the dataset to the community freely for research purposes, and we hope it will enable future research.

In future work, we plan to extend the annotations to additional languages and explore other sources, such as social media posts. This expansion aims to provide a broader understanding of role portrayal across diverse linguistic and contextual settings. We also intend to perform a more in-depth dataset analysis to examine formulations such as central entity identification. We further plan to evaluate the potential biases in the dataset aiming to ensure robustness for downstream tasks and end-user applications.

Limitations

Corpus Our dataset focuses on two domains: the Ukraine–Russia War and Climate Change, and covers news articles in five languages (Bulgarian, English, Hindi, Portuguese, and Russian). While these domains and languages provide a diverse foundation for entity framing analysis, the corpus contains 1,378 articles and it should not be considered representative of all news coverage or media landscapes in any specific country. Moreover, the dataset is not balanced with respect to topics, entities, or languages. Furthermore, human annotation of entity framing and role portrayal inevitably is subjective and annotators may subconsciously have biases that could influence quality. Despite providing detailed annotation guidelines and conducting quality control measures such as double annotation and adjudication through the curation process, some level of subjectivity may remain in the dataset.

Baseline Models Our reported experiments utilize state-of-the-art baselines covering a range of fine-tuned multilingual transformer models and hierarchical zero-shot learning with LLMs. However, we have not yet explored alternative architectures or advanced techniques such as few-shot, instruction-based evaluation, or multitask learning. Future work could investigate these approaches to improve model efficiency and performance.

Additionally, our zero-shot learning experiments rely on OpenAI's GPT-4o, a closed-source model that is subject to changes over time and may be deprecated in the future. This dependency may impact the reproducibility and interpretability. To address these challenges, future research should prioritize improving open-source models to ensure greater accessibility, transparency, and reproducibility.

Ethics and Broader Impact

Biases Our dataset aims to capture a balanced range of perspectives. While our goal is to incorporate diverse news sources and viewpoints, achieving perfect balance is not always possible. Consequently, inherent biases in the original media sources may be present in the annotations. To reduce unwanted annotation biases, the corpus is annotated with clear instructions to annotators to focus strictly on the framing of entities, setting aside their personal opinions. All annotations are performed by subject-matter experts, and we did not use crowd-sourcing.

Intended Use and Misuse Potential The primary goal of this corpus is to facilitate research on entity framing, role portrayal, and media analysis. These tools can help researchers, journalists, and the general public identify framing patterns and biases in news content. However, there is a risk that the corpus could be misused for malicious purposes, such as manipulating news narratives. We urge users to employ this resource responsibly and remain aware of potential ethical risks associated with its misuse.

Environmental Impact The use of LLMs requires substantial computational power, contributing to carbon emissions. Even though we used LLMs in a zero-shot in-context-learning setting rather than training models from scratch, the LLMs still rely on GPUs for inference, which has an environmental impact.

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

Acknowledgments

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.

Mohaddeseh Bastan, Mahnaz Koupaee, Youngseo Son, Richard Sicoli, and Niranjan Balasubramanian. 2020. Author's sentiment prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 604–615, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kelly Bergstrand and James M. Jasper. 2018. Villains, victims, and heroes in character theory and affect control theory. *Social Psychology Quarterly*, 81(3):228–247.

Ted Brader. 2006. *Campaigning for hearts and minds: how emotional appeals in political ads work*. Studies in communication, media, and public opinion. University of Chicago Press, Chicago.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2024. OATS: A challenge dataset for opinion aspect target sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12336–12347, Torino, Italia. ELRA and ICCL.

Travis Coan, Constantine Boussalis, John Cook, and Mirjam Odile Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, 11.

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.

Lea Frermann, Jiatong Li, Shima Khanehzar, and Gosia Mikolajczak. 2023. Conflicts, villains, resolutions: Towards models of narrative media framing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8712–8732, Toronto, Canada. Association for Computational Linguistics.

Kimberly Gross and Lisa D'Ambrosio. 2004. Framing emotional response. *Political Psychology*, 25(1):1–29.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia. Association for Computational Linguistics.

Shanto Iyengar. 1991. *Is anyone responsible? how television frames political issues*. American politics and political economy series. University of Chicago Press, Chicago.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.

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

Alapan Kuila and Sudeshna Sarkar. 2024. Deciphering political entity sentiment in news with large language models: Zero-shot and few-shot strategies. In *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. Emotion and Decision Making. *Annual Review of Psychology*, 66(1):799–823.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Seyed Mahed Mousavi, Gabriel Roccabruna, Aniruddha Tammewar, Steve Azzolin, and Giuseppe Riccardi. 2022. Can emotion carriers explain automatic sentiment prediction? a study on personal narratives. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 62–70, Dublin, Ireland. Association for Computational Linguistics.

Robin L. Nabi. 2003. Exploring the Framing Effects of Emotion: Do Discrete Emotions Differentially Influence Information Accessibility, Information Seeking, and Policy Preference? *Communication Research*, 30(2):224–247.

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.

Matan Orbach, Orith Toledo-Ronen, Artem Spector, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. YASO: A targeted sentiment analysis evaluation dataset for open-domain reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9154–9173, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15407–15428. Association for Computational Linguistics.

Valeria Pastorino, Jasivan A. Sivakumar, and Nafise Sadat Moosavi. 2024. Decoding news narratives: A critical analysis of large language models in framing detection. *Preprint*, arXiv:2402.11621.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023. 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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2022. Entity-level sentiment analysis (ELSA): An

exploratory task survey. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6773–6783, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556, Osaka, Japan. The COLING 2016 Organizing Committee.

Anne Schneider and Helen Ingram. 1993. Social Construction of Target Populations: Implications for Politics and Policy. *American Political Science Review*, 87(2):334–347.

HA Semetko and PM Valkenburg. 2000. Framing european politics: a content analysis of press and television news. *Journal of Communication*, 50(2):93–109.

Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2149–2163, Dubrovnik, Croatia. Association for Computational Linguistics.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13(7).

Yixuan Tang, Yi Yang, Allen Huang, Andy Tam, and Justin Tang. 2023. FinEntity: Entity-level sentiment classification for financial texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15465–15471, Singapore. Association for Computational Linguistics.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2023. On the relationship between frames and emotionality in text. In *Northern European Journal of Language Technology, Volume 9*, Linköping, Sweden. Linköping University Electronic Press.

Drew Westen. 2008. *The political brain: the role of emotion in deciding the fate of the nation*, nachdr. edition. PublicAffairs, New York, NY.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era

of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Trans. on Knowl. and Data Eng.*, 35(11):11019–11038.

A Detailed Taxonomy with Examples

Below, we definer the entries in our taxonomy, and we give an example for each category. Any references to "URW" and "CC" below denote the Ukraine-Russia War, and the Climate Change domains, respectively.

A.1 Protagonist

Guardian: Heroes or guardians who protect values or communities, ensuring safety and upholding justice. They often take on roles such as law enforcement officers, soldiers, or community leaders (e.g., climate change advocacy community leaders).
Example: Police officers protecting citizens during a crisis, firefighters saving lives during a disaster, community leaders standing against crime or leaders standing up for action to address climate change.

Martyr: Martyrs or saviors who sacrifice their well-being, or even their lives, for a greater good or cause. These individuals are often celebrated for their selflessness and dedication. This is mostly in politics, not in CC.
Example: Civil rights leaders like Martin Luther King Jr., who was assassinated while fighting for equality, or journalists who risk their lives to report on corruption and injustice.

Peacemaker: Individuals who advocate for harmony, working tirelessly to resolve conflicts and bring about peace. They often engage in diplomacy, negotiations, and mediation. This is mostly in politics, not in CC.
Example: Nelson Mandela's efforts to reconcile South Africa post-apartheid, or diplomats working to broker peace deals between conflicting nations.

Rebel: Rebels, revolutionaries, or freedom fighters who challenge the status quo and fight for significant change or liberation from oppression. They are often seen as champions of justice and freedom.
Example: Leaders of independence movements like Mahatma Gandhi in India, or modern-day activists fighting for democratic reforms in authoritarian regimes. In CC domain, this includes characters such as Greta Thunberg, or persons who, for instance, chain themselves to trees to prevent deforestation.

Underdog: Entities who are considered unlikely to succeed due to their disadvantaged position but strive against greater forces and obstacles. Their stories often inspire others.

Example: Grassroots political candidates overcoming well-funded incumbents, or small nations standing up to larger, more powerful countries. In CC, this could included NEs portrayed as underfunded organizations that are framed as showing promise to make positive impact on CC.

Virtuous: Individuals portrayed as virtuous, righteous, or noble, who are seen as fair, just, and upholding high moral standards. They are often role models and figures of integrity.
Example: Judges known for their fairness, or politicians with a reputation for honesty and ethical behavior. In CC, this includes leaders standing up for environmental ethical values to protect planet Earth, or activists pushing for environmental sustainability.

A.2 Antagonist

Instigator: Individuals or groups initiating conflict, often seen as the primary cause of tension and discord. They may provoke violence or unrest.
Example: Politicians using inflammatory rhetoric to incite violence, or groups instigating protests to destabilize governments. In CC, this could also include Greta Thunberg or activists chaining themselves to trees. In the previous example, they were portrayed in positive light as rebels. However, they could just as well be framed in a negative light if they are being portrayed as troublemakers and instigators of problems, and in such a scenario, they would also take the sub-role of Sabateur.

Conspirator: Those involved in plots and secret plans, often working behind the scenes to undermine or deceive others. They engage in covert activities to achieve their goals.
Example: Figures involved in political scandals or espionage, such as Watergate conspirators or modern cyber espionage cases. In CC, this could manifest as persons or organizations conspiring to bypass environmental regulations for profit.

Tyrant: Tyrants and corrupt officials who abuse their power, ruling unjustly and oppressing those under their control. They are often characterized by their authoritarian rule and exploitation.
Example: Dictators like Kim Jong-un in North Korea, or corrupt officials embezzling public funds and suppressing dissent.

Foreign Adversary: Entities from other nations or regions creating geopolitical tension and acting against the interests of another country. They are often depicted as threats to national security. This is mostly in politics, not in CC.

Example: Rival nations involved in espionage or military confrontations, such as the Cold War adversaries, or countries accused of election interference. In CC, foreign adversaries could include portrayal of how other countries are not adhering to CC policies (e.g., China refuses to adhere to CC policies resulting in 20% increase in CO2 emissions.

Traitor: Individuals who betray a cause or country, often seen as disloyal and treacherous. Their actions are viewed as a significant breach of trust. This is mostly in politics, not in CC.
Example: Whistleblowers revealing sensitive information for personal gain, or soldiers defecting to enemy forces. Note that if whistleblowers are portrayed in a positive light, their role would be Virtuous. This could equally apply to both politics and CC.

Spy: Spies or double agents accused of espionage, gathering and transmitting sensitive information to a rival or enemy. They operate in secrecy and deception. This is mostly in politics, not in CC.
Example: Historical figures like Aldrich Ames, who spied for the Soviet Union, or contemporary cases of corporate espionage.

Saboteur: Saboteurs who deliberately damage or obstruct systems, processes, or organizations to cause disruption or failure. They aim to weaken or destroy targets from within.
Example: Insiders tampering with critical infrastructure, or activists sabotaging industrial operations.

Corrupt: Individuals or entities that engage in unethical or illegal activities for personal gain, prioritizing profit or power over ethics. This includes corrupt politicians, business leaders, and officials.
Example: Companies involved in environmental pollution, executives engaged in massive financial fraud, or politicians accepting bribes and engaging in graft.

Incompetent: Entities causing harm through ignorance, lack of skill, or incompetence. This includes people committing foolish acts or making poor decisions due to lack of understanding or expertise. Their actions, often unintentional, result in significant negative consequences.
Example: Leaders making reckless policy decisions without proper understanding, officials mishandling crisis responses, or managers whose poor judgment leads to organizational failures.

Terrorist: Terrorists, mercenaries, insurgents, fanatics, or extremists engaging in violence and terror to further ideological ends, often targeting civilians. They are viewed as significant threats to peace and security. This is mostly in politics, not in CC.
Example: Groups like ISIS or Al-Qaeda carrying out attacks, or lone-wolf terrorists committing acts of violence.

Deceiver: Deceivers, manipulators, or propagandists who twist the truth, spread misinformation, and manipulate public perception for their own benefit. They undermine trust and truth.
Example: Politicians spreading false information for political gain, or media engaging in propaganda.

Bigot: Individuals accused of hostility or discrimination against specific groups. This includes entities committing acts falling under racism, sexism, homophobia, antisemitism, islamophobia, or any kind of hate speech. This is mostly in politics, not in CC.

A.3 Innocent

Forgotten: Marginalized or overlooked groups who are often ignored by society and do not receive the attention or support they need. This includes refugees, who face systemic neglect and exclusion.
Example: Indigenous populations facing ongoing discrimination; homeless individuals struggling without adequate support; refugees fleeing conflict or persecution.

Exploited: Individuals or groups used for others' gain, often without their consent and with significant detriment to their well-being. They are often victims of labor exploitation, trafficking, or economic manipulation.
Example: Workers in sweatshops; victims of human trafficking; communities suffering from corporate exploitation of natural resources.

Victim: People cast as victims due to circumstances beyond their control, specifically in two categories: (1) victims of physical harm, including natural disasters, acts of war, terrorism, mugging, physical assault, ... etc., and (2) victims of economic harm, such as sanctions, blockades, and boycotts. Their experiences evoke sympathy and calls for justice, focusing on either physical or economic suffering.
Example: Victims of natural disasters, such as hurricanes or earthquakes; individuals affected by violent crimes. Victims of economic blockades, sanctions, or boycotts.

Scapegoat: Entities blamed unjustly for problems or failures, often to divert attention from the real causes or culprits. They are made to bear the brunt of criticism and punishment without just cause.
Example: Minority groups blamed for economic problems; political opponents, accused of provoking national strife, without evidence.

B Annotation Guidelines

These guidelines aim to prepare the annotators and avoid human biases before starting the annotation:

- The annotators should get acquainted with the two domains covered by the tasks; for instance, (Coan et al., 2021) and (Amanatullah et al., 2023) provide a good coverage of the CC and URW domains, respectively,

- The annotators' opinions on the topics and sympathies towards key entities mentioned in the articles are irrelevant and should by no means impact the annotation process and their choices,

- The annotators should not exploit any external knowledge bases for the purpose of annotating documents.

Our guidelines for annotating and curating the entity framing corpus are as follows. Any references to "URW" and "CC" below denote the Ukraine-Russia War, and the Climate Change domains, respectively.

1. The entities of interest are understood in a broad sense to include both traditional named entities (such as persons, organizations, and locations) and *toponym-derived* entities. Toponym-derived entities are phrases that indicate a group or collective identity based on a place or affiliation, including, but not limited to:

   - Political, military, or social groups defined by their association with a location or entity, e.g., "Trump supporters," or "residents of Ukraine."
   - Entities denoting a geographic or organizational affiliation, such as "Russian forces" or "European officials."

2. Annotators are provided with a number of news articles and are expected to assign roles to named entities that are **central** to the article's story, according to the taxonomy of roles that was provided earlier.

3. Annotators are provided with a detailed taxonomy that includes definitions and examples.

4. The title of an article should not be annotated. The title of the article is the first block of text that appears in the annotation platform *INCEpTION*.

5. Only named entities that are central to the narrative of the article should be annotated. Unnamed entities (i.e., nominal entity mentions such as "migrants") should not be annotated.

   For more details on what qualifies as a named entity, in addition to the definition of the broader sense of named entities given above in these guidelines, the annotators should also examine the NER annotation guidelines in www.universalner.org/guidelines/.

6. Annotators pick one or more fine-grained roles for the named entities they believe are central to the article's story.

7. Entity mentions can be assigned fine-grained roles from more than one main role. However, during curation, we will not include these instances in the current version of the corpus, even though we annotate them.

8. Named entities that are not central to the story should not be annotated.

   The determination of how central a named entity is in an article is admittedly subjective. To reduce bias, such determination should be based on the careful reading of the article.

9. As a general rule, annotators should annotate only the first mention of each entity where it is clear that this entity has the specific role(s). There is no need to annotate subsequent mentions of this entity with the same role, but annotating more mentions with the same surface form and role is not a mistake; it is simply not required.

   This rule also extends to surface mentions of the same entity. For example, "Putin" and "Vladimir Putin" are both surface mentions of the same entity, so only the first occurrence would be annotated.

On the other hand, while entities such as "Moscow", "Russia", and "Putin" are closely related, they are not surface forms of the same entity, and are considered to be distinct separate entities.

10. If the above results in more than one mention of the same entity with the same role, the curator does not need to remove all of these additional mentions. We keep all of them.

11. Should an entity that was previously annotated with a certain role appear in a different context with different roles, the first mention where the roles changed should be annotated.

    The above rule is repeated as many times as the entity changes roles across mentions. For example, if an entity, let's say NATO, appears 20 times in an article, the first 10 mentions show NATO as a Guardian and a Virtuous entity. The 11th-15th mentions portray NATO as a Foreign Adversary, and the 16th-20th mentions portray NATO as Exploited, then we need only 3 annotations in total to account for the 3 different roles that NATO was portrayed as. These 3 annotations should all be the first mentions where NATO assumed each distinct role (i.e., mention 1, mention 11, and mention 16 should be annotated).

12. If different surface forms for the same named entity (e.g., NATO vs. North Atlantic Treaty Organization) appear in the article, it is sufficient to annotate only one of the surface forms.

13. If the above results in multiple surface forms of the same entity being annotated, the curator does not need to remove all of these additional mentions. We keep all of them.

14. There is no "Other" label in the taxonomy, as mentions without a discernible role in relation to the taxonomy are simply not assigned any role.

15. The curator may see conflicting annotations in the curation mode and can resolve the conflict, and then the remaining non-conflicting roles can be checked and adapted accordingly.

C   Experimental Settings

All fine-tuning experiments were conducted on a single NVIDIA RTX 4090 GPU with 24 GB of memory. We fine-tuned XLM-R (XLM-RoBERTa) in a single run, using a fixed random seed to ensure reproducibility. When the input context was at the sentence granularity, we performed sentence splitting using Stanza pipelines for each one of our five languages. For XLM-R, default settings were applied, with the following configurations:

- Model: XLM-R$_{base}$ (125M parameters)

- Learning Rate: 2e-5

- Batch Size: 8

- Epochs: 20 (with early stopping of 3 based on validation loss)

- Random Seed: 42

- Weight Decay: 0.01

To optimize performance, the sigmoid thresholds for fine-grained role predictions were tuned on the validation set. These optimized thresholds were then applied to generate predictions on the test set.

For the zero-shot experiments, we used OpenAI's GPT-4o (gpt-4o-2024-11-20) with a temperature setting of 0.2 to produce more conservative responses. To ensure the outputs conformed to our defined data types, we employed OpenAI's Structured Outputs API, which returned results in the expected JSON format.
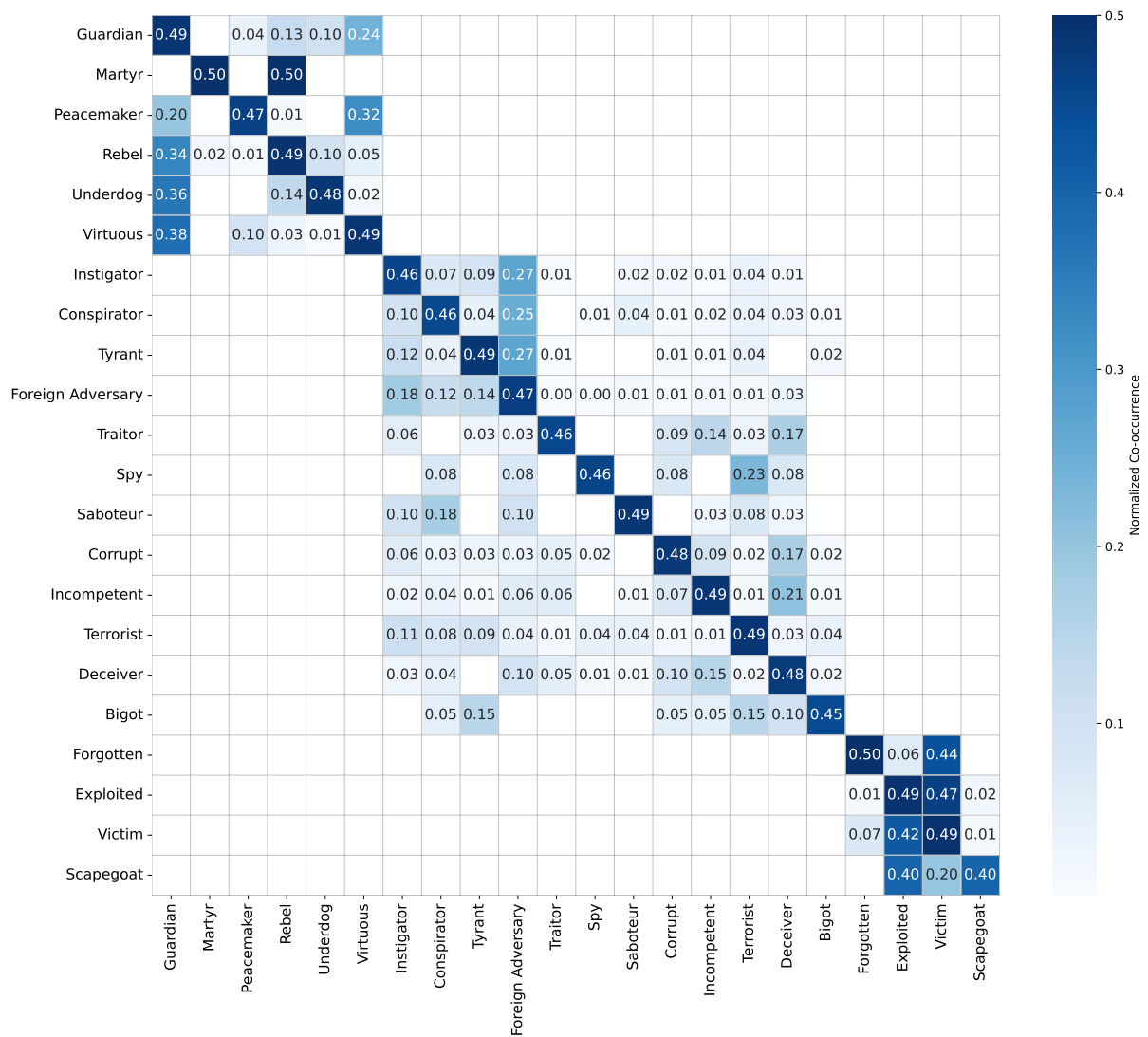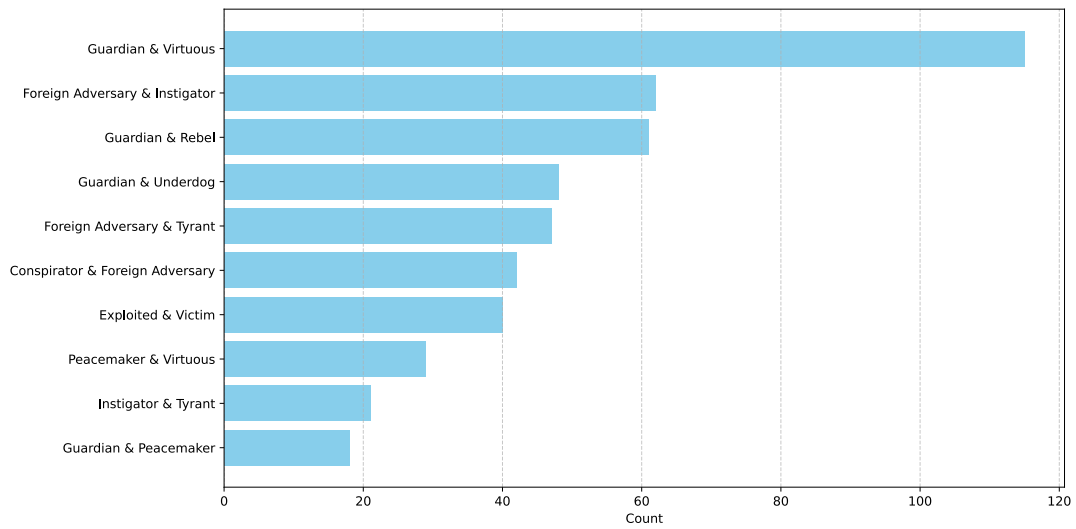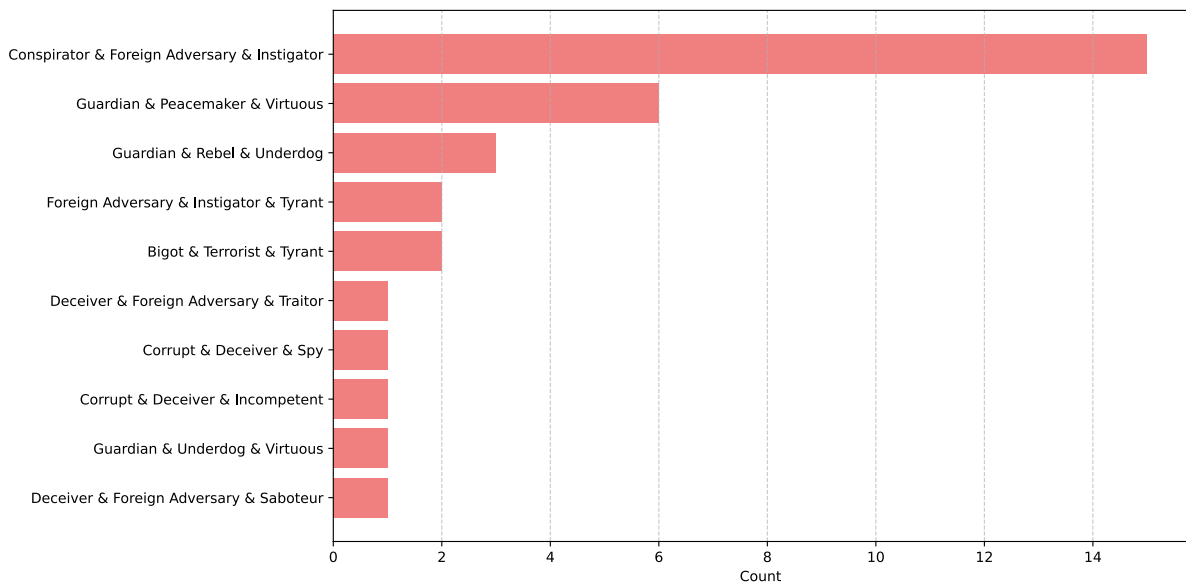
D Dataset Statistics
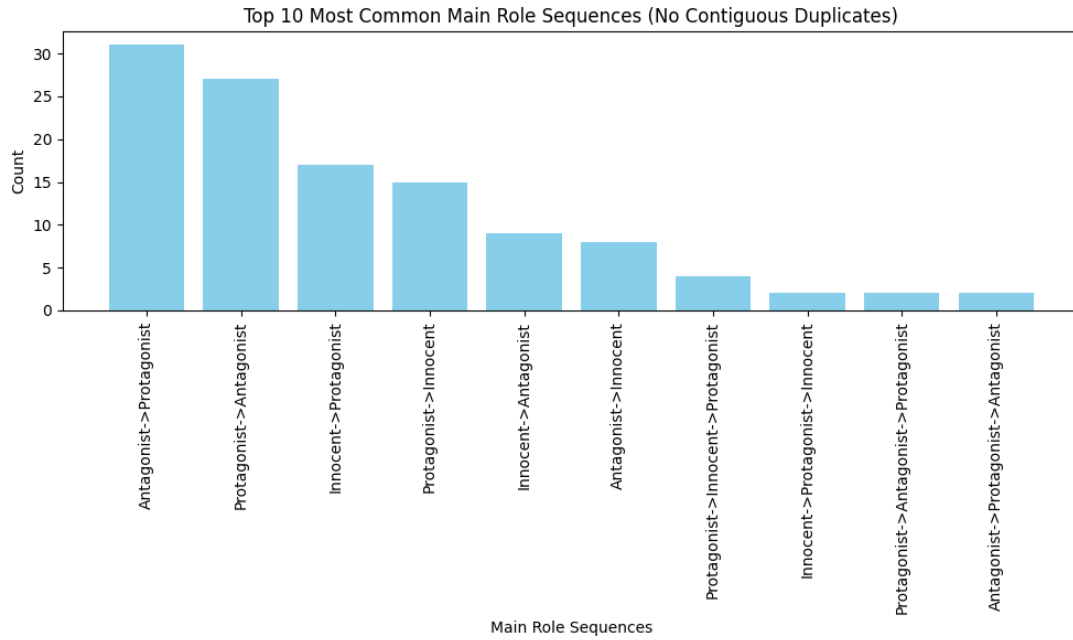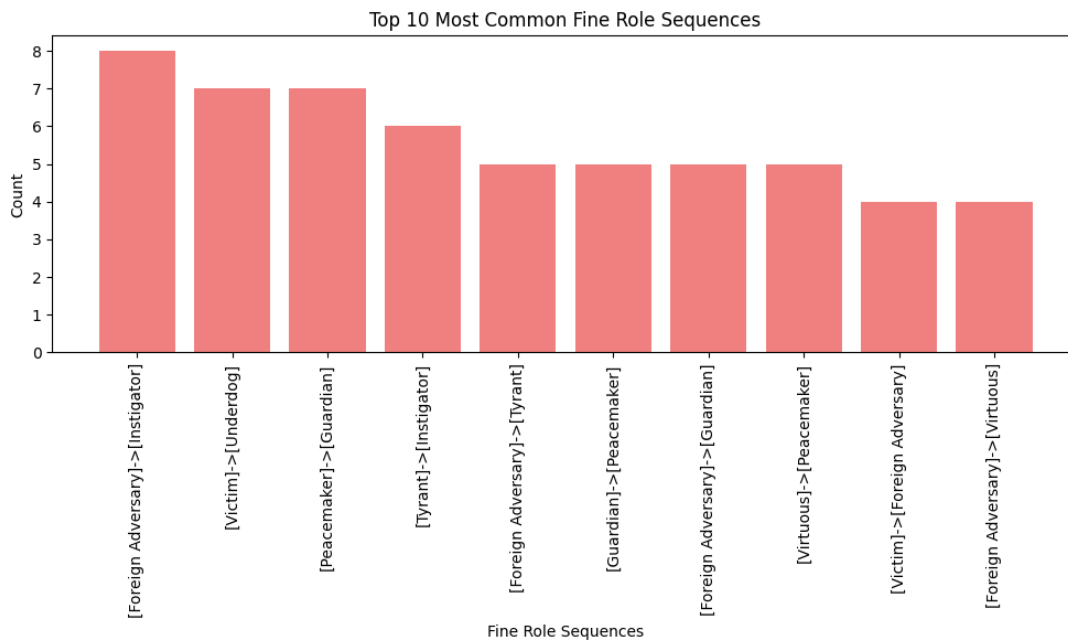

Figure 5: Normalized co-occurrence of fine-grained roles.

(a)

(b)

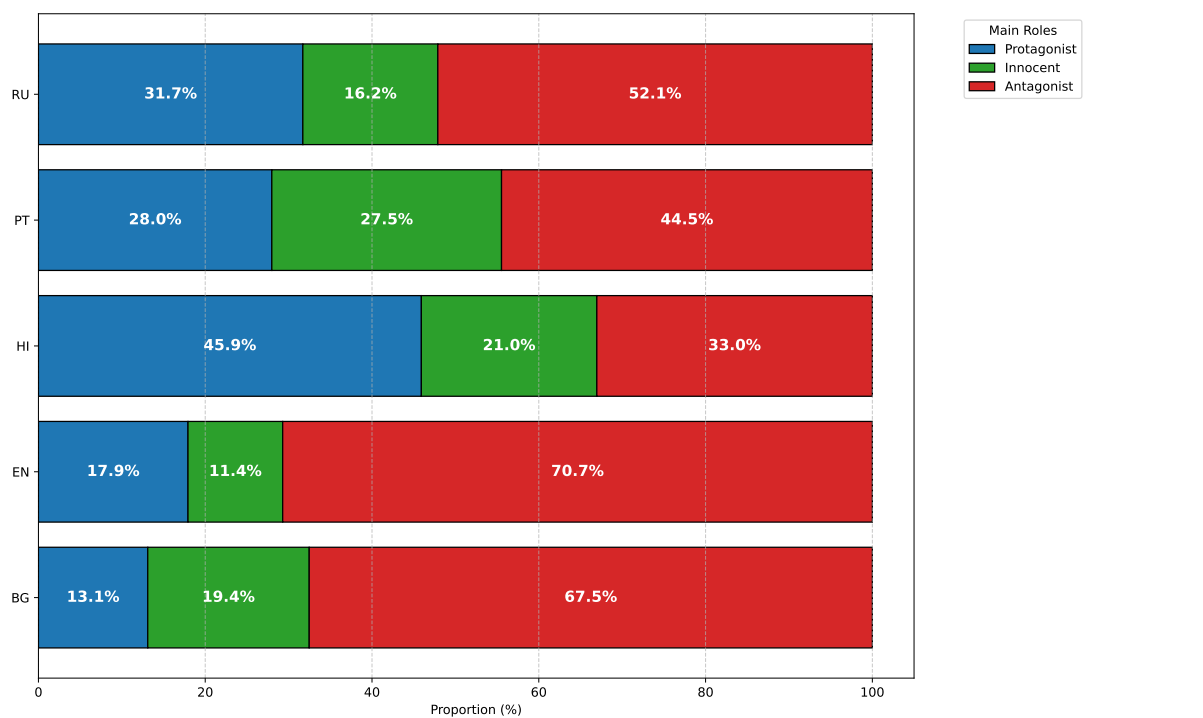Figure 6: The 10 most frequent co-occurring (a) pairs and (b) triplets of fine-grained roles.

Top 10 Most Common Main Role Sequences (No Contiguous Duplicates)
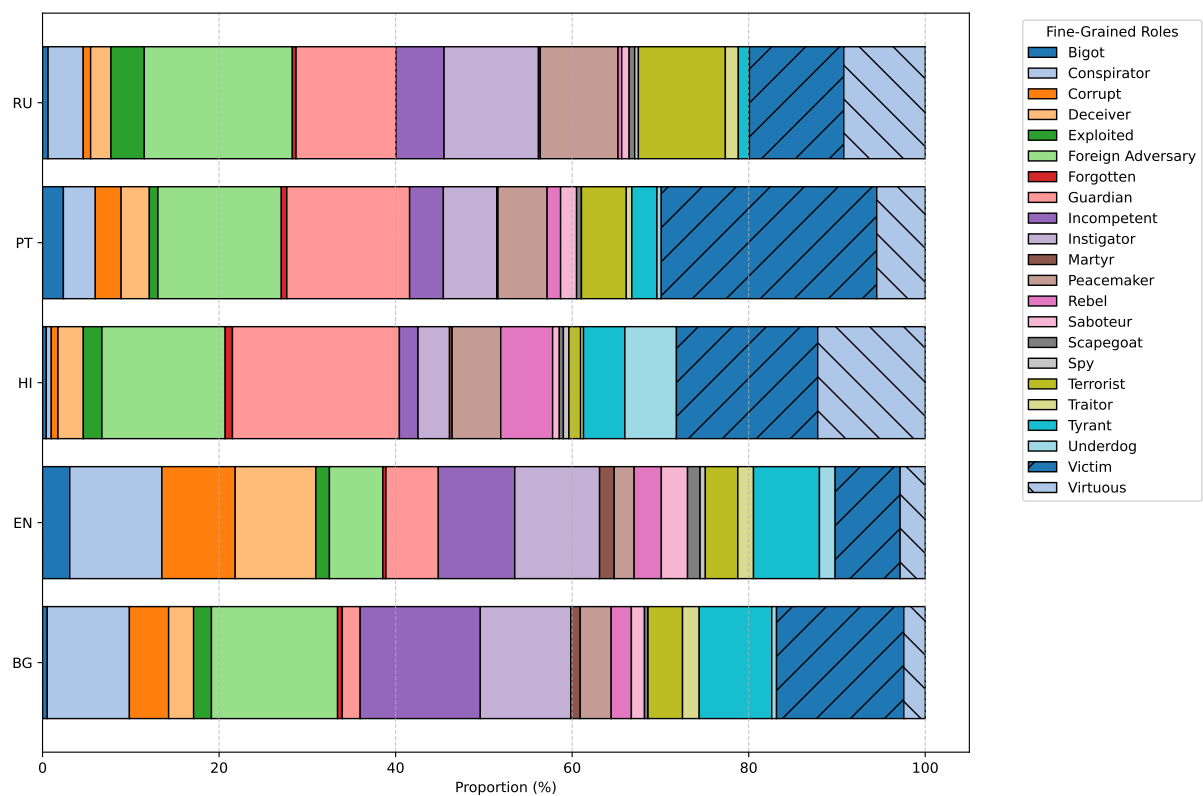
(a)

Top 10 Most Common Fine Role Sequences

(b)

Figure 7: The 10 most frequent (a) main role and (b) fine-grained role transition sequences.
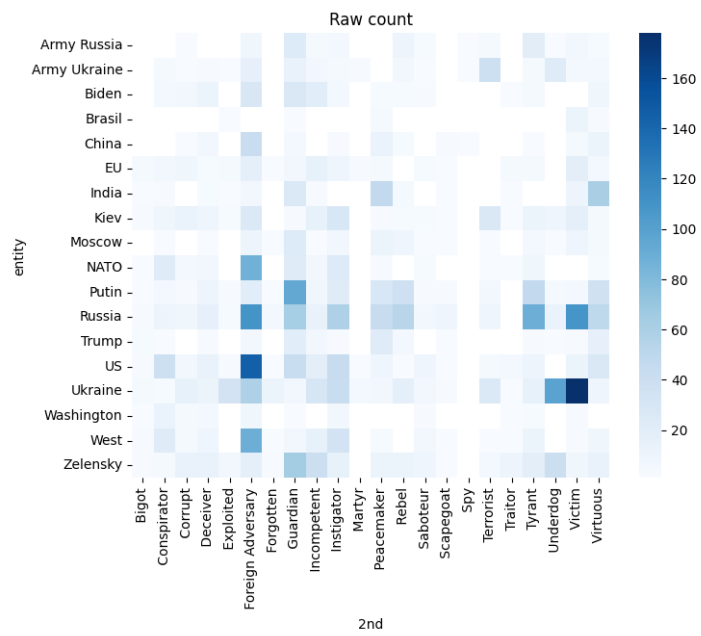
(a)

(b)

Figure 8: Proportions of (a) main roles and (b) fine-grained roles per language.

Figure 9: Heathmap of the raw count of mention of entity and 2nd level role, where entities are defined and selected as in Figure 3
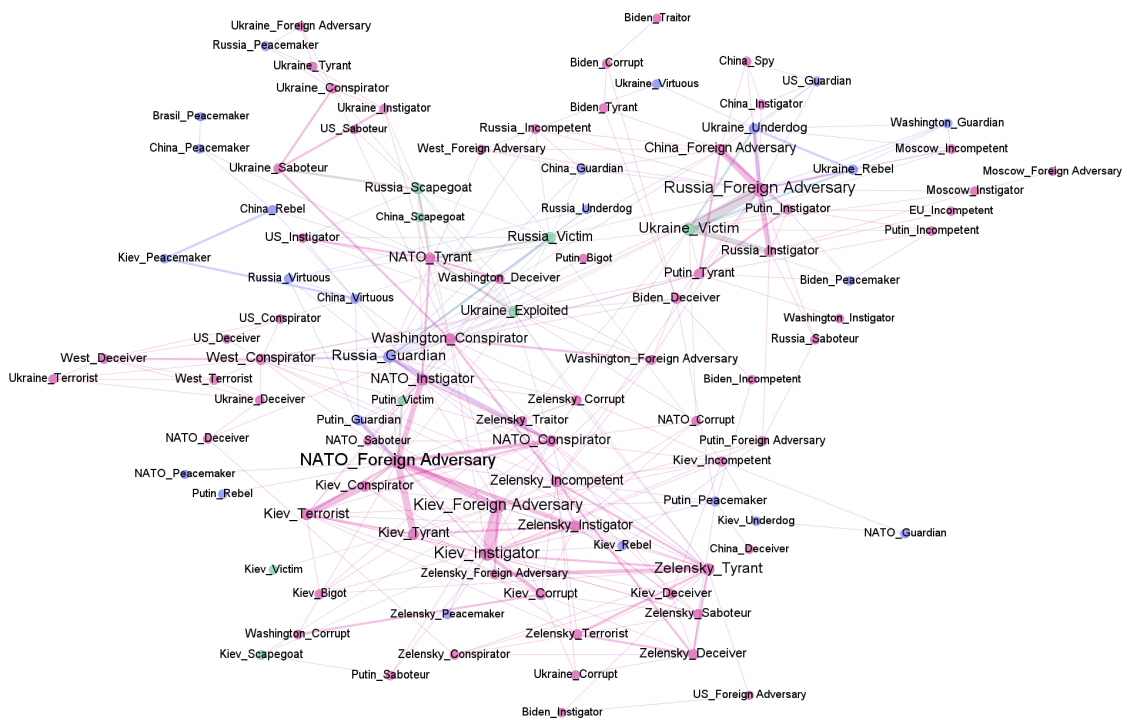
Figure 10: Graph of the top entities, nodes are a pair of entity and 2nd level role, node size is relative to the count of mentions, node color codes the 1st level role, there is and edge between two nodes, if they appear in the same document. This graph illustrate how group of entity+role pairs can be used to identify potential narratives.

E Prompts for Hierarchical Zero-Shot Experiments

```
You are an expert at identifying entity framing and role portrayal in news articles. Analyze the following entity
mention in context, and predict its main role and fine-grained role(s) from the taxonomy below.

Taxonomy: {detailed taxonomy with definitions and examples}

Context Around Entity: {context}

Entity Mention: {entity mention}

Task: Based on the provided context, assign to the entity mention at least one fine-grained role and
exactly one main role.

Return a JSON that has below attributes:
- main role: either one of Protagonist, Antagonist, or Innocent
- fine grained roles: a list of all your predicted fine-grained roles
```

Figure 11: **Single-Step Prompt Template.** The detailed taxonomy is the same one shown in A. The context is the text consisting of entity mention along with the 20 words before and after the entity mention.

```
First Step (LLM Call 1): Predict the Main Role

You are an expert at identifying entity framing and role portrayal in news articles. Analyze the following entity
mention in context, and predict its main role from the taxonomy below.

Taxonomy: {list of fine-grained roles per main role}

Context Around Entity: {context}

Entity Mention: {entity mention}

Task: Based on the provided context, assign to the entity mention exactly one main role.

Return a JSON that has this attribute:
- main role: either one of Protagonist, Antagonist, or Innocent


Second Step (LLM Call 2): Predict the Fine-Grained Role

You are an expert at identifying entity framing and role portrayal in news articles. This entity is
portrayed as a(n) {main role} and your task is to analyze the entity mention in context
and predict its fine-grained role(s) from the taxonomy below.

Taxonomy: {pertinent portion of the detailed taxonomy with definitions and examples}

Context Around Entity: {context}

Entity Mention: {entity mention}

Task: Based on the provided context, assign to the entity mention at least one fine-grained role.

Return a JSON that has this attribute:
- fine grained roles: a list of all your predicted fine-grained roles
```
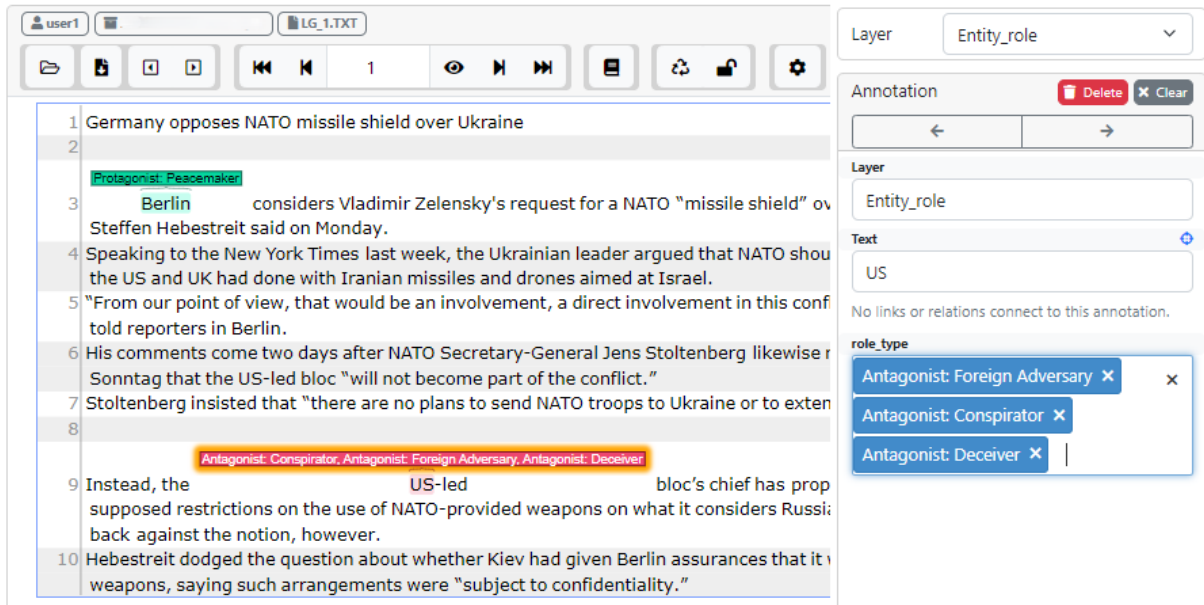
Figure 12: **Multi-Step Prompt Template.** In the first step, the taxonomy is only the tree structure of the taxonomy and does not include any definitions or examples. In the second step, the detailed taxonomy only includes the branch under the predicted main role in the first step. The context is as defined in Figure 11.

F Annotation Tool

We used the Inception (Klie et al., 2018) platform[2] to annotate our corpus because it has a rich set of features that extends beyond mere annotation to also include useful tools such as the ability to perform annotation adjudication through curation, monitoring the annotation progress, and calculating agreement between annotators. Inception allows to assign the following roles to users: annotator, curator, and manager.

To annotate a mention of an entity with a role, annotators should go to the part of the article where the entity is mentioned and select it. After selecting an entity mention, annotators can then assign roles as shown in 13.

Figure 13: Annotating entity framing using Inception.

---

[2]https://inception-project.github.io/