SemEval 2025 Task 10: Multilingual Characterization and Extraction of Narratives from Online News

Jakub Piskorski[1], **Tarek Mahmoud**[2], **Nikolaos Nikolaidis**[3], **Ricardo Campos**[4], **Alípio Jorge**[5], **Dimitar Dimitrov**[6], **Purificação Silvano**[7], **Roman Yangarber**[8], **Shivam Sharma**[9], **Tanmoy Chakraborty**[9], **Nuno Guimarães**[5], **Elisa Sartori**[10], **Nicolas Stefanovitch**[11], **Zhuohan Xie**[2], **Preslav Nakov**[2], **Giovanni Da San Martino**[10]

[1]Institute of Computer Science, Polish Academy of Science, Poland jpiskorski@gmail.com

[2] Mohamed bin Zayed University of Artificial Intelligence, UAE preslav.nakov@mbzuai.ac.ae

[3] Athens University of Economics and Business, Greece nnikon@aueb.gr

[4] University of Beira Interior and INESC TEC, Portugal ricardo.campos@ubi.pt

[5] University of Porto and INESC TEC, Portugal amjorge@fc.up.pt, nrsg@inesctec.pt

[6] Sofia University "St. Kliment Ohridski", Bulgaria mitko.bg.ss@gmail.com

[7] University of Porto, CLUP and INESC TEC Portugal msilvano@letras.up.pt

[8] University of Helsinki, Finland roman.yangarber@helsinki.fi

[9] Indian Institute of Technology, Delhi shivam.sharma@ee.iitd.ac.in, tanchak@iitd.ac.in

[10] University of Padova, Italy elisa.sartori.2@unipd.it, giovanni.dasanmartino@unipd.it

[11] European Commission Joint Research Centre, Italy nicolas.stefanovitch@ec.europa.eu

Abstract

We introduce SemEval-2025 Task 10 on *Multilingual Characterization and Extraction of Narratives from Online News*, which focuses on the identification and analysis of narratives in online news media. The task is structured into three subtasks: (1) *Entity Framing*, to identify the roles that relevant entities play within narratives, (2) *Narrative Classification*, to assign fine-grained narrative categories to documents, given a topic-specific taxonomy of narrative labels, and (3) *Narrative Extraction*, to provide a justification for the choice of dominant narrative of the document. We analyze news articles across two timely and critical domains, Ukraine-Russia War and Climate Change, in five languages: Bulgarian, English, Hindi, Portuguese, and Russian. This task introduces a novel multilingual and multifaceted framework for studying how online news media construct and disseminate manipulative narratives. By addressing these challenges, our work contributes to the broader effort of detecting, understanding, and mitigating the spread of propaganda and disinformation. The task attracted a lot of interest: 310 teams registered, and 40 system description papers were accepted.

1 Introduction

The Internet has opened vast possibilities for creating direct communication channels between producers and consumers of information, potentially leaving the latter exposed to deceptive content and attempts at manipulation. Huge audiences can be affected online, and major crisis events are constantly subjected to the spread of harmful disinformation and propaganda.

This creates a growing demand for tools that assist media experts in analyzing the news ecosystem, detecting manipulation attempts, and studying how media worldwide engage with topics of global interest, including the arguments and techniques used to influence public opinion.

To foster research and development in this direction, a number of shared tasks have been organized over the years. This includes SemEval-2020 Task 11 on Detection of Persuasion Techniques in News Articles (Da San Martino et al., 2020); SemEval-2021 Task 6 on Detection of Persuasion Techniques in Texts and Images (Dimitrov et al., 2021); CONSTRAINT 2022 Shared Task on Detecting the Hero, the Villain, and the Victim in Memes (Sharma et al., 2023); SemEval-2023 Task 3 on Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup (Piskorski et al., 2023a); SemEval-2024 Task 4 on Multilingual Detection of Persuasion Techniques in Memes (Dimitrov et al., 2024); and CLEF 2024 Task 3 on Persuasion Techniques (Piskorski et al., 2024).

Our new task, named *Multilingual Characterization and Extraction of Narratives from Online News* expands on the previously mentioned tasks

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2610–2643

to explore new dimensions in the context of news analysis. The task focuses on the identification of narratives in the news, identification of entity roles, classification into dominant and sub-dominant narratives,[1] and the justification of the choice of dominant narrative. We cover news articles from two domains—*Ukraine-Russia War* (URW) and *Climate Change* (CC)—in five languages: Bulgarian, English, Hindi, (European) Portuguese, and Russian, making this a multi-lingual multi-faceted task. By systematically identifying and analyzing narratives across multiple languages and domains, this task contributes to a deeper understanding of how disinformation is framed and disseminated, providing valuable insights into how specific viewpoints gain traction and influence public perception. By detecting recurring and evolving narratives, this work lays the foundation for more effective countermeasures against disinformation, supporting journalists, fact-checkers, and policymakers in mitigating its societal impact.

The paper is organized as follows. Section 2 introduces the three subtasks. Section 3 surveys related work. Section 4 describes the dataset and its creation process. Section 5 gives an overview of the evaluation. Section 6 presents the results of the competition and comparison of the participant systems. Section 7 concludes with a summary of the task.

2 The Tasks

In this section, we describe the three subtasks of SemEval 2025 Task 10: (1) Entity Framing; (2) Narrative Classification; and (3) Narrative Extraction.

Subtask 1 (ST1) Entity Framing: Given a news article, such as in Figure 3 (top), and a list of mentions of named entities (NEs) contained therein, assign to each mention one or more roles from a predefined taxonomy of fine-grained roles. Formally, let R be a tree structure with k nodes that represents the taxonomy of roles. Let S be a string of length $|S|$ characters which contains the article. The goal of entity framing is to learn a function

$$f : (S, [i, j]) \rightarrow \{-1, +1\}^k \quad (1)$$

where $0 \leq i < j \leq |S|$ and +1/-1 at position l in the k-dimensional output vector means that the role r_l is present or not present in the span [i,j], respectively. This is a multi-label, multi-class classification task.

We use a two-level taxonomy of roles, with three main types of roles: *protagonist*, *antagonist*, and *innocent*, which are subdivided into 22 fine-grained roles. Figure 1 provides an overview of the taxonomy of the entity roles, and Figure 3 illustrates how the taxonomy is used to annotate our running example. For an in-depth account of the entity-framing task and taxonomy details, please refer to (Mahmoud et al., 2025a) and (Stefanovitch et al., 2025), respectively.

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 1: Two-level taxonomy of entity roles.

Subtask 2 (ST2) Narrative Classification: Given a news article, as in Figure 3, and a two-level taxonomy of narrative labels (with each narrative subdi-

[1] In the context of our task a narrative is defined as a *"recurring, repetitive (across and within articles), overt or implicit claim that presents and promotes a specific interpretation or viewpoint on an ongoing (and often dynamic) news topic."* (Luntz, 2007)

vided into sub-narratives) from a particular domain, assign all appropriate sub-narrative labels to the article. Formally, let S be the text of the article and let $Narr = \{n_1, n_2, \ldots n_m\}$ be the set of sub-narratives. The task is to learn the function:

$$f : S \times Narr \rightarrow \{-1, +1\}^m, \qquad (2)$$

where -1 at position j in the m-dimensional output vector means that narrative n_j is not present in the article, and $+1$ means narrative n_j is present. This is a multi-label multi-class document classification task.

We use a two-level narrative taxonomy for the two domains in focus (URW, CC), depicted in Figure 2. This consists of several coarse-grained narratives, subdivided into fine-grained sub-narratives. For an in-depth description of the taxonomies, please refer to Figures 5 and 6 in Annex A. Figure 3 demonstrates how the taxonomy in used for annotating our running example.

UKRAINE-RUSSIA WAR

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
Other

CLIMATE CHANGE

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
Other

Figure 2: Coarse-grained narratives for Ukraine-Russia war and Climate Change domains.

Subtask 3 (ST3) Narrative Extraction: Given a news article, as in Figure 3 (top), and the *dominant* narrative and sub-narrative of the article, generate an explanation supporting the choice of this dominant narrative and sub-narrative, as shown in Figure 3 (bottom). Formally, let S be the text of the article and let $Narr = \{n_1, n_2, \ldots n_m\}$ be a set of all sub-narratives used in Subtask 2. The goal of the task is to learn the function:

$$f : (S, n) \rightarrow T = (t_1, t_2, \ldots t_j) \qquad (3)$$

where $n \in Narr$, and T is a sequence of j tokens, where $j \leq 80$. This is a text-generation task.

Killing Russian Culture: Public opinion in the West is now built very clearly: everyone adheres to the idea that Russia is absolute evil, and the West is absolute good

'Public opinion in the West is now built very clearly: everyone adheres to the idea that Russia is absolute evil, and the West is an absolute good', says Italian artist Jorit Agoch.

With the beginning of the special operation in Ukraine, the Russian people in the West faced a substantial wave of Russophobia, which also swept the arts and sports.

Singers, artists and directors are finding their names crossed out from concert schedules and festival shortlists.

The Munich Philharmonic Orchestra severed all relations with conductor Valery Gergiev, and the Carnegie Hall in New York cancelled performances by the pianist Denis Matsuev.

Even those who are dead – Dostoevsky, Tchaikovsky, Shostakovich – became victims of Russophobia, and the list is growing every day.

How does Russian culture withstand this wave of aggression?

Entity roles
Russia – Innocent-Victim
Russian people in the West – Innocent-Victim
Munich Philarmonic Orchestra – Antagonist-Bigot
Carnegie Hall – Antagonist-Bigot
Narrative classification
URW: Russia is the victim: The West is Russophobic
Dominant narrative
URW: Russia is the victim: The West is Russophobic
Explanation
The article talks about Russia being a victim of Western Russophobia with Russian culture being cancelled.

Figure 3: Running news article example from our dataset (top) accompanied with an annotation (bottom).

3 Related Work

We next discuss work related to the three subtasks considered in this paper.

3.1 Subtask 1: Entity Framing

Entity framing (Mahmoud et al., 2025a) is a crucial aspect of media analysis, focusing on how individuals, groups, or concepts are portrayed within a given narrative. Over the years several datasets have been proposed to support this task. Sharma et al. (2023) presented a dataset that identifies *heroes*, *villains*, and *victims* in memes, based on visual features. Card et al. (2016) explored a framing perspective that detects personas, which they use to determine article-level framing, as captured by the Media Frames Corpus (MFC) (Card et al., 2015). MFC seeks to identify how articles are

framed across nine categories (e.g., Economics or Politics). Other investigations into news framing (Pastorino et al., 2024; Otmakhova et al., 2024; Piskorski et al., 2023b; Liu et al., 2019; Card et al., 2015) similarly center on article-level framing. In aspect-based sentiment analysis and targeted sentiment analysis (Chebolu et al., 2024; Zhang et al., 2022; Orbach et al., 2021; Jiang et al., 2019; Saeidi et al., 2016), the goal is to identify opinion targets and assign sentiment polarity to specific aspects, typically using a binary polarity scheme, across multiple attributes of the target. In contrast to previous work, our dataset is anchored in textual analysis rather than visual features, and focuses on the explicit framing of entities within the text. While many existing datasets primarily examine article-level framing or general sentiment toward entities, our approach provides a more granular perspective by capturing specific roles assigned to entities within a narrative.

3.2 Subtask 2: Narrative Classification

A *narrative* is a complex concept, with various definitions depending on the context in which it is used. It can refer to broad ideological framings, storytelling patterns, or structured sequences of events that shape public perception and discourse. In media analysis, narratives are often studied to understand how information is framed, how it spreads, and what underlying themes emerge from large-scale text corpora (Campos et al., 2024). In an effort to synthesize various formulations of the concept of "narrative," Dennison (2021) proposes 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 [...] generalizable and can be applied to multiple situations, as opposed to specific stories.*" Several examples of formulations have been provided in previous taxonomies and datasets. Kotseva et al. (2023) created a three-level narrative taxonomy for COVID-19 and used it to classify and analyze trends over time; Li et al. (2023) focused on a flat taxonomy of anti-vax narratives; Hughes et al. (2021) presented a taxonomy of typical anti-vax narratives, organized around several common tropes and rhetorical strategies. Coan et al. (2021b) presented a two-level taxonomy for common instances of climate change denial in short snippets. Amanatullah et al. (2023) presented a flat taxonomy of common pro-Russian narratives in the alleged pro-Kremlin influence campaigns related to the war in Ukraine.

3.3 Subtask 3: Narrative Explanation

The ability to explain text narratives is gaining importance, particularly in detecting disinformation and propaganda. This has driven the need for annotated datasets that do not only support narrative understanding, but also provide explicit explanations that can help models predict outcomes, articulate reasoning, and enhance interpretability. Several datasets contribute to this effort. NarrativeQA facilitates narrative comprehension by providing detailed question-answer pairs about story elements, aiding tasks like contextual reasoning and summarization (Kočiský et al., 2018). The TellMeWhy dataset focuses on causal reasoning, enabling models to explain event causality in stories, which leads to a better understanding of complex narrative structures (Lal et al., 2021). e-SNLI (Explainable SNLI) extends the Stanford Natural Language Inference (NLI) dataset with human-annotated explanations for entailment, contributing to research on explainability in natural language inference (Camburu et al., 2018). While these datasets highlight the growing emphasis on interpretability and narrative comprehension, they do not address the objectives of our work. Unlike approaches focused on summarizing narratives (Zhao et al., 2022), our dataset provides short explanatory texts that justify the assignment of dominant narratives and sub-narratives within each text. This novel approach bridges narrative classification with interpretability, emphasizing the reasoning behind narrative categorization and enhancing transparency in NLP models.

4 The Datasets

4.1 Data colletion

Our dataset contains complete or partial articles collected from multiple online sources in five languages: Bulgarian, English, Hindi, Portuguese and Russian. The news articles focus on two subjects: the Ukraine-Russia War (URW), which began in February 2022 when Russia initiated a full-scale invasion of Ukraine, and Climate Change (CC), which includes both the denial of climate change and activism dedicated to mitigating its effects.

Articles were initially obtained via the *Europe Media Monitor*, a large-scale news aggregation system[2] complemented with custom region-based

[2] emm.newsbrief.eu

sources. The initial selection of candidate articles was performed as described below:

1. **Keyword-Based Queries**: Topic-specific keywords were formulated for URW and CC in all languages, and used to retrieve a comprehensive corpus of articles from selected sources.
2. **Zero-Shot Relevance classification of the articles**: Using the BART-large-MNLI model (Lewis et al., 2020) and a secondary set of pre-defined keywords (e.g., '*Denazification of Ukraine*', '*Climate hoax*'), zero-shot classification was performed on each article's title and the initial 300 characters of text. This process produces a relevance score in the range of $(0.0, 1.0)$ for each article.
3. **Persuasiveness Scoring**: A RoBERTa-based multi-label classifier, trained on the Persuasion Techniques dataset (Piskorski et al., 2023a,b), was utilized following the approach described in (Nikolaidis et al., 2024). This method produced a Persuasiveness Score for each article.
4. **Linear Combination for Ranking and Filtering**: The relevance scores from key phrases and four variants of the Persuasiveness Score were combined using a linear weighting approach to rank news articles from most to least likely to contain relevant narratives. Then, filtering was applied based on various additional criteria (e.g., Number of words > 250).
5. **Manual revision**: Finally, each article was manually reviewed to assess its relevance to the annotation task.

The lack of adequate texts addressing various topics in two of the languages led to the inclusion of additional sources. For Hindi, articles were selected from both mainstream and alternative outlets (e.g., *NDTV*, *The Hindu, OpIndia*). For Portuguese, sources included newspapers and political websites known for their controversial opinion pieces on relevant topics (e.g., *O Diabo*, *Esquerda.net*, *Folha Nacional*, *blasfesmias.net*).

4.2 Annotation process

A dedicated team was assigned to each of the five languages in the corpus—Bulgarian, English, European Portuguese, Russian, and Hindi. Each team was supervised by a designated language coordinator and was comprised of three to six annotators with expertise in linguistics, social sciences, and international relations, or with prior experience in annotation tasks. The annotators underwent comprehensive training, which involved studying the detailed annotation guidelines (Stefanovitch et al., 2025), attending live demonstrations, and participating in real-time annotation exercises.

To ensure consistency, each article was annotated by two annotators. For quality control, one or more curators were assigned to each language to verify adherence to the predefined guidelines. These curators systematically reviewed the annotations, assessed their accuracy and overall quality, and selected or distilled the most appropriate annotations. Regular weekly meetings were conducted in each language team, and across languages—to discuss ambiguous or difficult instances, resolve disagreements, maintain consistency in annotations, and refine the annotation guidelines as needed. Additional details on the annotation guidelines can be found in Annex B.

Cross-lingual coherence was ensured: firstly by reviewing outliers in label distributions, secondly by applying the multi-lingual and multi-document approach from (Stefanovitch and Piskorski, 2023) to flag clusters of annotations with potential disagreement for further review.

4.3 Annotation Quality

To assess Inter-Annotator Agreement (IAA), we calculate Krippendorff's α between annotators for each subtask and language at the fine-grained level. The IAA is computed at span, paragraph and document level for tasks 1, 2 and 3, respectively. Results are reported in Table 4. Interestingly, as the scope of the annotations increases, the overall agreement decreases. Specifically for subtask 2, the agreement is under the recommended value of 0.667, but is higher than IAA on tasks of similar complexity (Piskorski et al., 2023b). The quality of the dataset was further improved using the curation procedure described in the previous section. In Annex D, we give a detailed breakdown for subtask 2 at different levels of granularity and for both topics, to investigate how the intrinsic complexity of the taxonomies impacts on the annotation process.

4.4 Dataset Description

Subtask 1: Entity Framing Table 1 presents an overview of the corpus, including its division into training, development, and test splits, as well as a breakdown by language. The table shows the total number of documents, the total (and unique)

task	EN	RU	BG	PT	HI	all
1	0.460	0.436	0.733	0.467	0.489	0.522
2	0.388	0.415	0.642	0.385	0.379	0.462
3	0.409	0.338	0.540	0.332	0.383	0.449

Figure 4: IAA measure with Krippendorff's α for each subtask and language at fine-grained level.

number of entity mentions, the overall number of annotations, and the average number of entity mentions and annotations per document.

Split	Lang.	#DOC	#ENT	#ANN	AVG_e	AVG_a
TRAIN	ALL	1322	5616 (1938)	6259	4.2	4.7
	BG	259	626 (170)	709	2.4	2.7
	EN	202	685 (375)	744	3.4	3.7
	HI	342	2330 (665)	2723	6.8	8.0
	PT	306	1250 (396)	1315	4.1	4.3
	RU	213	725 (391)	768	3.4	3.6
DEV	ALL	136	599 (353)	650	4.4	4.8
	BG	15	30 (24)	33	2.0	2.2
	EN	27	90 (63)	99	3.3	3.7
	HI	35	279 (131)	307	8.0	8.8
	PT	31	115 (80)	123	3.7	4.0
	RU	28	85 (64)	88	3.0	3.1
TEST	ALL	322	1181 (565)	1320	3.7	4.1
	BG	54	123 (63)	127	2.3	2.4
	EN	62	234 (151)	264	3.8	4.3
	HI	78	315 (131)	381	4.0	4.9
	PT	71	296 (102)	322	4.2	4.5
	RU	57	213 (140)	226	3.7	4.0
TOTAL	ALL	1779	7396 (2411)	8229	4.2	4.6
	BG	328	779 (202)	869	2.4	2.6
	EN	291	1009 (503)	1107	3.5	3.8
	HI	455	2924 (798)	3411	6.4	7.5
	PT	408	1661 (485)	1760	4.1	4.3
	RU	297	1023 (498)	1082	3.4	3.6

Table 1: ST1 statistics: total number of documents (#DOC) by language, total number of annotated entity mentions (#ENT), with unique counts (in parentheses), total number of annotations (#ANN), average number of entity mentions per document (AVG_e), and average number of annotations per document (AVG_a).

Subtask 2: Narrative Classification Table 2 presents the document count, and average number of labels per document. Each document contains one or more coarse-grained (Narrative) labels and one or more fine-grained (sub-Narrative) labels. When no label was found in the coarse-grained level, the label *"Other"* was used. The dataset contains 2427 documents in total, and each article contains 2.4 fine-grained labels on average.

The label distribution is highly skewed, both within and across languages, reflecting the real-world conditions, where some narratives are more prevalent. The exact distribution for the two domains is provided in Figures 7-8 in Annex A.2.

Subtask 3: Dominant Narrative Explanation Table 3 presents the statistics of the Subtask 3 corpus grouped by language and dataset split. The table shows the number of documents and the average number of tokens in explanations.

Additional statistics about the datasets can be found in Annex A.

Split	Lang	#Doc	Avg_a
TRAIN	ALL	1914	—
	BG	401	2.13
	EN	399	2.19
	HI	366	1.79
	PT	400	3.04
	RU	216	2.20
DEV	ALL	140	—
	BG	35	1.91
	EN	41	2.78
	HI	35	2.43
	PT	35	2.26
	RU	32	2.47

Split	Lang	#Doc	Avg_a
TEST	ALL	460	—
	BG	100	2.38
	EN	101	3.08
	HI	99	1.34
	PT	100	4.02
	RU	60	3.05
TOTAL	ALL	2426	—
	BG	536	2.16
	EN	541	2.40
	HI	500	1.75
	PT	535	3.17
	RU	308	2.34

Table 2: ST2 corpus statistics showing total number of documents (#DOC) by language, and average number of labels per document (AVG_a).

Split	Lang.	#Doc	Avg_t
TRAIN	ALL	1215	35.13
	BG	357	22.81
	EN	203	29.78
	HI	193	50.08
	PT	252	49.67
	RU	210	23.32
DEV	ALL	140	33.01
	BG	28	17.96
	EN	30	37.97
	HI	29	55.28
	PT	25	36.92
	RU	28	16.93

Split	Lang.	#Doc	Avg_t
TEST	ALL	326	33.84
	BG	79	28.84
	EN	68	29.79
	HI	40	42.08
	PT	83	51.11
	RU	56	17.39
TOTAL	ALL	1681	34.00
	BG	464	23.20
	EN	301	32.51
	HI	262	49.14
	PT	360	45.90
	RU	294	19.22

Table 3: ST3 statistics showing total number of documents (#DOC), and average number of tokens in explanations (AVG_t) by language.

5 Evaluation Framework

5.1 Evaluation Measures

Subtask 1 is a multi-class multi-label classification problem. The official evaluation measure is *Exact Match Ratio*, which measures the subset accuracy, i.e., the proportion of the named entities for which the predicted fine-grained labels match the true labels (Sorower, 2010; Gibaja and Ventura, 2015). Additionally, we report micro precision, recall, and F_1 on fine-grained roles, and coarse-grained role accuracy.

Subtask 2 is a multi-label multi-class hierarchical classification problem. The official evaluation measure is *sample-averaged* F_1. Specifically, we first compute sample F_1 for each document, by comparing the gold-standard fine-grained (sub-narrative) labels to the predicted labels.[3] We then compute

[3]In sample F_1, true positives are labels correctly assigned to the document, false positives are labels that were incorrectly assigned to the document, and false negatives are labels that were incorrectly unassigned.

the average over the sample F_1 scores. We compute *macro-averaged sample* F_1 over the fine-grained (sub-narrative) labels as a secondary metric.

Subtask 3 consists in generating an explanation text for each document's dominant narrative. The official evaluation metric is the average similarity between the gold-standard and predicted explanation using the F_1 metric computed by *BertScore* (Zhang et al., 2020b) and a multilingual BERT model compatible with the five languages of the dataset.[4]

5.2 Task Organization

The shared task was run in two phases:

Development Phase: initially, only the *training* data was released to the participants. Subsequently, *development* data was released without the gold-standard labels and the participants competed to achieve the best performance on this development set. An unlimited number of submissions was allowed, and the overall best score for each team was shown in real-time on a public leaderboard.

Test Phase: in the second phase, the gold-standard labels for the *development* set, and the raw articles of the *test* set (without the gold-standard answers) were released. The participants were given approximately 10 days to submit their final predictions on the *test* set for ST1 and ST2. The test phase for ST3—the release of the test dataset for ST3—was carried out once the test phase for ST1 and ST2 was *closed*—since the articles in the test datasets for ST2 and ST3 are the same.

During the test phase the participants could submit multiple runs, but they received no feedback on their performance. The *latest* submission of each team was considered as official and was used for the final team ranking. Overall, 66 teams made official submissions for all subtasks, with 35, 28, and 18 teams submitting results for ST1, ST2, ST3, respectively. Of these, 13, 10, and 7 teams submitted results for *all languages* for ST1, ST2, ST3, respectively.

The results for the development and the test phases are available on the official leaderboard page.[5] After the competition was over, the submission system for the test dataset remains open for continued evaluation *post* shared task and monitoring of the state of the art.

6 Participants and Results

This section provides the official results on all three subtasks. For complete information with suplementary metrics, please see the official leaderboard on the test data.[6]

6.1 Subtask 1: Entity Framing

The official system ranking is shown in Table 4.

6.1.1 Baseline

We use *Random Guess* as the baseline: we first randomly guess the main role of the given named entity (NE); we then randomly select the fine-grained role from the sub-categories of the main role.

6.1.2 System Highlights

A comparison of the techniques used by the participants in ST1 is shown in Table 5.

The following systems are worth mentioning. **DUTIR** (Lv et al., 2025) first conducts data augmentation by translating all languages into English to address class imbalance. Next, they train multiple base language models using QLoRA. Finally, they aggregate the predictions from these fine-tuned models, where GLM-4-Plus serves as the meta-classifier to produce the final predictions. This novel and effective approach secured them first place in English, Portuguese, and Russian, and second place in Bulgarian. **PATeam** (Sun et al., 2025) employs multi-prompt engineering to enhance the contextual analysis of the target location entities using Qwen2.5-72B. They perform data cleaning and augmentation by dynamically restructuring the input text around these entities. Then they train five models on the cleaned data, and the final prediction is determined by majority voting. This approach earned them first place in Bulgarian and second place in English and Portuguese. A similar approach is adopted by **QUST** (Liu et al., 2025). Notably, **TartanTritons** (Raghav et al., 2025) incorporates an iterative feedback mechanism, where model-generated error messages are used to refine the predictions via retries. This strategy secured them second place in Hindi.

6.2 Subtask 2: Narrative Classification

The official system ranking is shown in Table 6.

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

[5] propaganda.math.unipd.it/semeval2025task10/leaderboard.php

[6] propaganda.math.unipd.it/semeval2025task10/leaderboardv3.html

English		Portuguese		Russian		Bulgarian		Hindi	
TEAM	EMR	TEAM	EMR	TEAM	EMR	TEAM	EMR	TEAM	EMR
DUTIR	.413	DUTIR	.593	DUTIR	.565	PATeam	.516	QUST	.468
PATeam	.383	PATeam	.492	QUST	.514	DUTIR	.508	TartanTritons	.446
DEMON	.375	QUST	.458	TartanTritons	.472	DEMON	.460	BERTastic	.440
gowithnlp	.370	BERTastic	.418	BERTastic	.467	gowithnlp	.436	DEMON	.402
TartanTritons	.357	LTG	.407	DEMON	.467	TartanTritons	.411	LTG	.364
Fane	.345	DEMON	.367	gowithnlp	.449	QUST	.387	Cimba	.354
QUST	.328	LATeIIMAS	.337	PATeam	.444	BERTastic	.355	gowithnlp	.335
LATeIIMAS	.311	TartanTritons	.333	LTG	.430	Fane	.347	DUTIR	.294
NlpUned	.311	gowithnlp	.269	Cimba	.383	LTG	.315	Dhananjaya	.279
m1nadzuki	.260	Cimba	.263	FromProblemImportSolve	.355	Cimba	.258	LATeIIMAS	.272
LTG	.255	FromProblemImportSolve	.263	LATeIIMAS	.313	FromProblemImportSolve	.210	PATeam	.269
BERTastic	.251	Fane	.256	Dhananjaya	.294	Dhananjaya	.194	FromProblemImportSolve	.256
adithjrajeev	.251	Dhananjaya	.219	YNUzwt	.266	Baseline	.040	Fane	.234
Mekky	.217	YNUzwt	.162	Fane	.243			HowardUniversityAI4PC	.168
NarrativeMiners	.213	HowardUniversityAI4PC	.131	HowardUniversityAI4PC	.126				
FromProblemImportSolve	.204	Baseline	.047	Baseline	.051				
YNUzwt	.200								
Cimba	.187								
NarrativeNexus	.183								
Rosetta	.179								
Dhananjaya	.175								
UMZNLP	.140								
Tuebingen	.132								
YNUHPCC	.089								
north	.085								
HowardUniversityAI4PC	.081								
kzeky	.068								
bumblebeeTransformer	.064								
eevvgg	.064								
Baseline	.038								
Team12	.021								
cocoa	.017								
SemanticaInnovators	.013								

Table 4: Complete Rankings for ST1 using the updated Exact Match Ratio (EMR) across all languages.

6.2.1 Baseline

For ST2, we use a random-guess baseline with uniform sampling. For each document, we first randomly chose how many labels to sample (1 or 2), and then randomly sample labels from the fine-grained level of the taxonomy.

6.2.2 System Highlights

In Table 8, we present a short breakdown of the techniques used by the participants in ST2.

GateNLP (Singh et al., 2025) secured the top spot in 3 out of 5 languages. They fine-tuned a *Llama3.2* model on a rebalanced and augmented version of the dataset and used multi-step hierarchical prompting to classify the narratives. **PATeam** (Sun et al., 2025) used data enhancement strategies, including the use of semantic segmentation to isolate Narrative-relevant fragments, and then fine-tuned *Phi-4* and *Qwen2.5* models, winning first place on the Bulgarian dataset. **INSALyon2** (Eljadiri and Nurbakova, 2025) proposed a multi-agent approach, where multiple narrative-specific LLM agents interact in a group-chat-like configuration, winning 3rd place on English. **KostasThesis2025** (Eleftheriou et al., 2025) implemented a chunking strategy, producing one embedding per article, and then experimented with several configurations of classification, and with training using a continual learning approach. They scored within the top-6 in 4 of the 5 languages.

6.3 Subtask 3: Narrative Extraction

The official system ranking is shown in Table 7.

6.3.1 Baseline

The baseline model we used for this task was the *Phi3-mini* (Abdin et al., 2024) with a context of 8K tokens and 7B parameters.[7] The prompt used to generate the texts is shown in Annex A.3.1.

If the output of the language model exceeds the limit of 80 words, the output is truncated to 80 words.

6.3.2 System Highlights

A comparison of the techniques used by the systems on ST3 is presented in Table 9.

We highlight systems that achieved top-3 performance in any of the languages. **GPLSICOR-TEX** (Martínez-Murillo et al., 2025) fine-tuned a *T5-flan* model using external knowledge injection, combining the intention of each article with its dominant and sub-dominant narratives. **Kyu-Hyun Choi** (Choi and Na, 2025) fine-tuned the *PEGASUS* model (Zhang et al., 2020a), and **WordWiz** (Ahmadi and Zeinali, 2025) developed a multi-temperature inference strategy to select three possible explanations for each document (using *Phi3.5*),

[7] huggingface.co/microsoft/Phi-3-small-8k-instruct

Team Name	BERT	RoBERTa	DistilBERT	GPT-3	GPT-3.5 Turbo	GPT 4o	XLM-RoBerta	DeBERTa	Llama3	Bernice	Phi-3.5	Phi-4	Qwen2.5	Llama3.1	Gemma2	DeepSeek-R1	Aya Expanse 8B	all-MiniLM-L12-v2	TweetNLP	calme-2.4-rys-78b	o1-mini	ELMo	CoT	self consistency	least-to-most	chain-of-symbols	three-of-thoughts	n-gram	Data augmentation	Standard Preprocessing	LIWC
DUTIR													✓	✓									✓						✓		
PATeam												✓	✓												✓				✓	✓	
DEMON	✓								✓																						
Gowithnlp					✓	✓					✓																				
TartanTritons		✓											✓		✓								✓	✓				✓			
Fane						✓															✓		✓								
NlpUned																															
QUST								✓				✓	✓																		
LATE-GIL-nlp		✓													✓				✓	✓									✓		
LTG							✓		✓	✓																					
BERTastic						✓	✓																			✓					✓
adithjrajeev	✓							✓															✓						✓		
NarrativeMiners	✓		✓					✓										✓											✓		
YNUzwt						✓						✓												✓							
NarrativeNexus																		✓												✓	
Tuebingen	✓		✓				✓																						✓	✓	✓
HowardUniversityAI4PC										✓			✓		✓	✓	✓						✓				✓				
cocoa	✓		✓	✓	✓																										

Table 5: ST1: Overview of the approaches and the features used by the participating systems. The systems highlighted in bold ranked first for at least one language.

and then selected the most relevant one based on narrative alignment.

These 3 teams achieved the best scores on the English data, with **WordWiz** also ranking among the top-3 in all languages. **PATeam** (Wan et al., 2025), and **BBStar** (Tyagi et al., 2025) were also among the top-3 systems on Portuguese, Bulgarian and Hindi. On Russian, **TartanTritons** (Raghav et al., 2025) replaced **BBStar** on the podium.

PATeam used *Phi-4* with data augmentation and direct preference optimization; **BBStar** implemented a Reasoning+Acting framework that leverages semantic retrieval-based few-shot prompting. **TartanTritons** also leveraged the power of a quantized *Phi-4* combined with structured prompting.

6.4 Aggregated Results

We provide the average official scores for the teams participating in all five languages in ST1, ST2 and ST3—in Tables 10, 11 and 12, respectivelly.

7 Conclusions and Future Work

This paper describes SemEval-2025 Task 10 on *Multilingual Characterization and Extraction of Narratives from Online News*. The task attracted a lot of attention: 310 teams registered for the task, of which 66 made an official submission on the test set, of which 40 submitted a task description paper.

In future work, we envisage exploiting the datasets we created for of this task to explore and elaborate solutions for other related tasks—e.g., narrative classification at the paragraph level, unsupervised discovery of a taxonomy of narratives, detection of entities central to a narrative, and predicting the dominant narrative in a document based on an explanatory text.

8 Ethics Policy

Intended Use and Misuse Potential: The datasets created in the context of the presented Shared Task were designed to advance research on entity framing, narrative classification, and extraction, with the broader goal of detecting deceptive content across multiple languages and domains in online media. However, given the potential risks of exploiting the datasets to boost the production of biased manipulative disinformation, we advise responsible use of the datasets.

Environmental Impact: The deployment of LLMs may have a large carbon footprint, especially when training new models. We have exploited a LLM as a baseline in one of the subtasks, however, we did not train it, but only used an existing trained model, which is relatively cheap.

Fairness: We engaged many annotators to create the datasets for this Shared Task. Some them were researchers with a linguistic background and prior annotation experience, coming from the institutions of the co-organizers of the Task. They were fairly

English		Russian		Portuguese		Hindi		Bulgarian	
TEAM	F_1	TEAM	F_1	TEAM	F_1	TEAM	F_1	TEAM	F_1
GATENLP	.438	GATENLP	.518	GATENLP	.480	DUTtask10	.535	PATeam	.460
COGNAC	.426	PATeam	.434	PATeam	.409	IRNLP	.515	GATENLP	.416
INSALyon2	.406	iLostTheCode	.411	23	.313	Narrlangen	.385	iLostTheCode	.369
23	.377	Narrlangen	.405	KostasThesis2025	.309	UNEDTeam	.376	UNEDTeam	.363
NCLteam	.345	YNUzwt	.335	iLostTheCode	.293	GATENLP	.321	Narrlangen	.355
Narrlangen	.344	KostasThesis2025	.333	Narrlangen	.291	KostasThesis2025	.282	KostasThesis2025	.333
PATeam	.339	UNEDTeam	.330	UNEDTeam	.270	INSAntive	.265	INSAntive	.324
YNUzwt	.321	INSAntive	.323	YNUzwt	.266	NotMyNarrative	.243	Irapuarani	.183
iLostTheCode	.320	UniBonn187	.231	Irapuarani	.225	PATeam	.218	NotMyNarrative	.142
Narrengers	.318	Irapuarani	.191	INSAntive	.215	iLostTheCode	.147	DUTtask10	.121
UNEDTeam	.313	IRNLP	.116	CtrlAltElite	.149	Irapuarani	.111	LATeIIMAS	.072
CtrlAltElite	.311	GrammarPolice	.050	NotMyNarrative	.124	LATeIIMAS	.029	Baseline	.022
NotMyNarrative	.298	DUTtask10	.033	DUTtask10	.026	Baseline	.000		
IRNLP	.287	Baseline	.008	Baseline	.014	bbStar	.000		
INSAntive	.281	LATeIIMAS	.000						
NLPPraktikumWS2025	.258								
KostasThesis2025	.239								
NarrativeMiners	.238								
nlptuducd	.226								
ammd7	.222								
UniBonn187	.206								
Irapuarani	.188								
DUTtask10	.165								
LATeIIMAS	.163								
GeorgeSnape	.156								
GrammarPolice	.063								
Baseline	.013								
NarrativeNexus	.000								
bbStar	.000								

Table 6: Complete Ranking for ST2 on the five languages based on the official score: Sample F_1.

English		Portuguese		Russian		Bulgarian		Hindi	
TEAM	F_1	TEAM	F_1	TEAM	F_1	TEAM	F_1	TEAM	F_1
KyuHyunChoi	.750	WordWiz	.749	PATeam	.706	PATeam	.704	PATeam	.755
WordWiz	.746	PATeam	.746	WordWiz	.704	WordWiz	.684	WordWiz	.734
GPLSICORTEX	.743	bbStar	.719	TartanTritons	.682	bbStar	.672	bbStar	.727
TechSSN	.742	YNUzwt	.688	YNUzwt	.676	TartanTritons	.655	TartanTritons	.699
NarrativeNexus	.731	TartanTritons	.685	bbStar	.664	Baseline	.634	Baseline	.670
NarrativeMiners	.729	Baseline	.680	DUTtask10	.664	LATeIIMAS	.624	DUTtask10	.000
clujteam	.725	LATeIIMAS	.673	Baseline	.644	DUTtask10	.000		
PAteam	.724	DUTtask10	.000	LATeIIMAS	.642				
TartanTritons	.713								
YNUzwt	.699								
LATeIIMAS	.696								
bbStar	.691								
Synapse	.675								
Baseline	.667								
DUTtask10	.000								
Mendel292A	.000								
UMZNLP	.000								
ftd	.000								

Table 7: Complete Ranking for ST3 on the five languages based on the official score: BertScore F_1 macro.

Team Name	encoder models	decoder models	fine-tuning	custom representations	prompting	statistical methods	data augmentation	preprocessing
iLostTheCode	✓					✓		✓
GATENLP	✓	✓	✓				✓	
Irapuarani	✓	✓	✓		✓			✓
PATeam		✓	✓				✓	✓
NotMyNarrative	✓		✓					
DUTtask10		✓	✓		✓		✓	
COGNAC		✓			✓			✓
nlptuducd		✓		✓				
INSAntive	✓		✓					✓
bbstar	✓	✓	✓		✓			
YNUzwt		✓						
NCLTeam	✓						✓	
LATE-GIL-nlp	✓		✓					
KostasThesis2025				✓		✓		
NarrativeMiners	✓		✓				✓	
UNEDTeam		✓			✓			✓
INSALyon2					✓			

Table 8: ST2: Overview of the approaches and the features used by the participating systems. The systems highlighted in bold ranked first for at least one language.

Team Name	transformers	prompt engineering	fine-tuning	instruction-tuning	reinforcement learning	data augmentation
bbStar	✓	✓				
clujteam	✓					
GPLSICORTEX	✓	✓	✓			
KyuHyunChoi	✓					
LATeIIMAS	✓		✓			
NarrativeMiners	✓		✓			
NarrativeNexus	✓		✓			
PATeam	✓		✓		✓	✓
TartanTritons	✓	✓		✓		
TechSSN	✓		✓			
WordWiz	✓		✓			

Table 9: ST3: Overview of the approaches and the features used by the participating systems. The systems highlighted in bold ranked first for at least one language.

Team	EN	PT	RU	BG	HI	AVG
DUTIR	.413	.593	.565	.508	.294	.475
PATeam	.383	.492	.444	.516	.269	.421
QUST	.328	.458	.514	.387	.468	.431
TartanTritons	.357	.333	.472	.411	.446	.404
DEMON	.375	.367	.467	.460	.402	.414
gowithnlp	.370	.269	.449	.436	.335	.372
BERTastic	.251	.418	.467	.355	.440	.386
LTG	.255	.407	.430	.315	.364	.354
Fane	.353	.256	.243	.347	.234	.287
Cimba	.187	.263	.383	.258	.354	.289
Dhananjaya	.175	.219	.294	.194	.279	.232
FromProblemImportSolve	.204	.263	.355	.210	.256	.258
HowardUniversityAI4PC	.081	.131	.126	.097	.168	.121
Baseline	.038	.047	.051	.040	.057	.047

Table 10: Average Exact Match Ratio across languages for the teams participating in all five languages for ST1.

Team	EN	PT	RU	BG	HI	AVG
GATENLP	.438	.480	.518	.416	.321	.435
PATeam	.339	.409	.434	.460	.218	.372
Narrlangen	.344	.291	.405	.355	.385	.356
UNEDTeam	.313	.270	.330	.363	.376	.330
iLostTheCode	.320	.293	.411	.369	.147	.308
KostasThesis2025	.239	.309	.333	.333	.282	.299
INSAntive	.281	.215	.323	.324	.265	.282
Irapuarani	.188	.225	.191	.183	.111	.180
DUTtask10	.165	.026	.033	.121	.535	.176
Baseline	.013	.014	.008	.022	.000	.011

Table 11: Sample F_1 score, across languages for the teams participating in all five languages for ST2.

Team	EN	PT	RU	BG	HI	AVG
PATeam	.724	.746	.706	.704	.755	.727
Wordwiz	.746	.749	.704	.684	.734	.723
bbStar	.691	.719	.664	.672	.727	.695

Table 12: Average and macro F1_score across languages for the teams participating in all five languages for ST3.

remunerated as part of their job.

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

9 Limitations

Dataset Representativeness The narrative taxonomies exploited in our task were edited by experienced media analysts, active in the study of misinformation and fact-checking. They focus on narratives of interest to media analysts in Western institutions. The selection of the narratives should not be perceived as covering the complete discourse

of the two domains, but rather what such analysts encounter in practice.

The datasets used in our shared task cover two current topics covered by a wide range of media outlets. Nevertheless, it is of paramount importance to emphasize that these datasets should not be considered as representative of the media in any specific country or region, nor should they be considered as balanced in any way.

Biases A very substantial effort has been invested in training the annotators and acquainting them with the specifics of the two domains of interest for our task. Cross-language quality control mechanisms have been put in place to ensure the highest quality of annotations. Nevertheless, we are aware that some degree of intrinsic subjectivity might be present in the datasets. Consequently, models trained using these datasets might exhibit certain biases.

Acknowledgements

We express deep gratitude to: Ivo Moravski, DataBee (getdatabee.com/) and specifically Peter-Michael Slaveykov, Krasen Zhelyzkov, Samuil Ivanov, and Blagovest Chernev for their invaluable contribution to the annotation of Bulgarian data.

We express deep gratitude to: Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis, and Ari Gonçalves for their invaluable contribution to the annotation of Portuguese data.

We express deep gratitude to: Gayatri Oke, Kanupriya Pathak, Dhairya Suman, and Sohini Mazumdar for their invaluable contribution to the annotation of Hindi data.

We express deep gratitude to: Denis Kvachev, Matilda Villanen, Ksenia Semenova, Irina Gatsuk, Aamos Waher and Daria Lyakhnovich for their invaluable contribution to the annotation of Russian data.

We express deep gratitude to: Nicolo Faggiani and Sopho Kharazi for their invaluable contribution to the annotation of English data. We express deep gratitude to: Ion Androutsopoulos and John Pavlopoulos for their guidance and support.

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.

NG work is co-financed by Component 5—Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021-2026, within project HfPT, with reference 41. NG, RC, PS, AJ would like to acknowledge project StorySense, with reference (DOI 10.54499/2022.09312.PTDC) and the Advanced Computing Project CPCA-IAC/AV/594794/2023 (doi.org/10.54499/CPCAIAC/ AV/594794/2023).

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.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia

Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Ruhollah Ahmadi and Hossein Zeinali. 2025. WordWiz at SemEval-2025 Task 10: Optimizing narrative extraction in multilingual news via fine-tuned language models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1272–1278, Vienna, Austria. Association for Computational Linguistics.

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.

Saurav K. Aryal and Prasun Dhungana. 2025. Howard University-AI4PC at SemEval-2025 Task 10: Ensembling LLMs for multi-lingual multi-label and multi-class meta-classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1584–1591, Vienna, Austria. Association for Computational Linguistics.

Gabriel Assis, Lívia de Azevedo, Joao VM de Moraes, Laura Ribeiro, and Aline Paes. 2025. Irapuarani at SemEval-2025 Task 10: Evaluating strategies combining small and large language models for multilingual narrative detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 38–48, Vienna, Austria. Association for Computational Linguistics.

Andreas Blombach, Bao Minh Doan Dang, Stephanie Evert, Tamara Fuchs, Philipp Heinrich, Olena Kalashnikova, and Naveed Unjum. 2025. Narrlangen at SemEval-2025 Task 10: Comparing (mostly) simple multilingual approaches to narrative classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2215–2224, Vienna, Austria. Association for Computational Linguistics.

Alberto Caballero, Alvaro Rodrigo, and Roberto Centeno. 2025. NlpUned at SemEval-2025 Task 10: Beyond training: A taxonomy-guided approach to role classification using LLMs. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 300–305, Vienna, Austria. Association for Computational Linguistics.

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

Ricardo Campos, Alípio Jorge, Adam Jatowt, Sumit Bhatia, and Marina Litvak. 2024. The 7th international workshop on narrative extraction from texts: Text2story 2024. In *Advances in Information Retrieval*, pages 391–397, Cham. Springer Nature Switzerland.

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.

Kyu Hyun Choi and Seung Hoon Na. 2025. KyuHyunChoi at SemEval-2025 Task 10: Narrative extraction using a summarization-specific pretrained model. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2238–2240, Vienna, Austria. Association for Computational Linguistics.

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

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

Lorenzo Vittorio Concas, Manuela Sanguinetti, and Maurizio Atzori. 2025. iLostTheCode at SemEval-2025 Task 10: Bottom-up multilevel classification of narrative taxonomies. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 610–619, Vienna, Austria. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav

Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

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

Ivan Diaz, Fredin Vázquez, Christian Luna, Aldair Conde, Gerardo Sierra, Helena Gómez-Adorno, and Gemma Bel-Enguix. 2025. LATE-GIL-nlp at Semeval-2025 Task 10: Exploring LLMs and transformers for characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 660–668, Vienna, Austria. Association for Computational Linguistics.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, Mexico City, Mexico. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Konstantinos Eleftheriou, Panos Louridas, and John Pavlopoulos. 2025. KostasThesis2025 at SemEval-2025 Task 10 Subtask 2: A continual learning approach to propaganda analysis in online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 902–911, Vienna, Austria. Association for Computational Linguistics.

Mohamed-Nour Eljadiri and Diana Nurbakova. 2025. Team INSALyon2 at SemEval-2025 Task 10: A zero-shot agentic approach to text classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 968–983, Vienna, Austria. Association for Computational Linguistics.

Enfa Fane, Mihai Surdeanu, Eduardo Blanco, and Steven R. Corman. 2025. Fane at SemEval-2025 Task 10: Zero-shot entity framing with large language models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1914–1924, Vienna, Austria. Association for Computational Linguistics.

Geraud Faye, Guillaume Gadek, Wassila Ouerdane, Celine Hudelot, and sylvain gatepaille. 2025. NotMy-Narrative at SemEval-2025 task 10: Do narrative features share across languages in multilingual encoder models? In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 58–66, Vienna, Austria. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Matteo Fenu, Manuela Sanguinetti, and Maurizio Atzori. 2025. DEMON at SemEval-2025 Task 10: Fine-tuning LLaMA-3 for multilingual entity framing. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1454–1462, Vienna, Austria. Association for Computational Linguistics.

Jesus M. Fraile-Hernandez and Anselmo Peñas. 2025. UNEDTeam at SemEval-2025 Task 10: Zero-shot narrative classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 164–172, Vienna, Austria. Association for Computational Linguistics.

Eva Gibaja and Sebastián Ventura. 2015. A tutorial on multilabel learning. *ACM Comput. Surv.*, 47(3):52:1–52:38.

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).

Silja Huttunen, Roman Yangarber, and Ralph Grishman. 2002. Diversity of scenarios in information extraction. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, Spain.

Azwad Anjum Islam and Mark A. Finlayson. 2025. COGNAC at SemEval-2025 Task 10: Multi-level narrative classification with summarization and hierarchical prompting. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1440–1447, Vienna, Austria. Association for Computational Linguistics.

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.

Özlem Karabulut, Soudabeh Eslami, Ali Gharaee, and Matthew Kirk Andrews. 2025. Tuebingen at SemEval-2025 Task 10: Class weighting, external

knowledge and data augmentation in BERT models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1243–1250, Vienna, Austria. Association for Computational Linguistics.

Muhammad Khubaib, Muhammad Shoaib Khursheed, Muminah Khurram, Abdul Samad, and Sandesh Kumar. 2025. NarrativeMiners at SemEval-2025 Task 10: Combating manipulative narratives in online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1631–1637, Vienna, Austria. Association for Computational Linguistics.

Panagiotis Kiousis. 2025. IRNLP at SemEval-2025 Task 10: Multilingual narrative characterization and classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 54–57, Vienna, Austria. 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).

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

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.

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

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

Ning Li, You Zhang, Jin Wang, Dan Xu, and Xuejie Zhang. 2025a. YNU-HPCC at SemEval-2025 Task 10: A two-stage approach to solving multi-label and multi-class role classification based on DeBERTa. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1976–1982, Vienna, Austria. Association for Computational Linguistics.

Shu Li, George Snape Williamson, and Huizhi Liang. 2025b. NCLTeam at SemEval-2025 Task 10: Enhancing multilingual, multi-class, and multi-label document classification via contrastive learning augmented cascaded UNet and embedding based approaches. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 421–426, Vienna, Austria. Association for Computational Linguistics.

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.

Jiyan Liu, Youzheng Liu, Taihang Wang, Xiaoman Xu, Yimin Wang, and Ye Jiang. 2025. Team QUST at SemEval-2025 Task 10: Evaluating large language models in multiclass multi-label classification of news entity framing. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1049–1053, Vienna, Austria. Association for Computational Linguistics.

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.

Frank Luntz. 2007. *Words That Work: It's Not What You Say, It's What People Hear*. Hyperion, New York.

Tengxiao Lv, Juntao Li, Chao Liu, Yiyang Kang, Ling Luo, Yuanyuan Sun, and Hongfei LIN. 2025. DUTIR at SemEval-2025 Task 10: A large language model-based approach for entity framing in online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 173–178, Vienna, Austria. Association for Computational Linguistics.

Tarek Mahmoud, Zhuohan Xie, Dimitar Dimitrov, Nikolaos Nikolaidis, Purificação Silvano, Roman Yangarber, Shivam Sharma, Elisa Sartori, Nicolas Stefanovitch, Giovanni Da San Martino, Jakub Piskorski, and Preslav Nakov. 2025a. Entity framing and role portrayal in the news.

Tarek Mahmoud, Zhuohan Xie, and Preslav Nakov. 2025b. BERTastic at SemEval-2025 Task 10: State-of-the-art accuracy in coarse-grained entity framing

for Hindi news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 390–399, Vienna, Austria. Association for Computational Linguistics.

Anca Marginean. 2025. clujteam at semeval-2025 task 10: Finetuning smollm2 with taxonomy-based prompting for explaining the dominant narrative in propaganda text. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Iván Martínez-Murillo, María Miró Maestre, Aitana Morote Martínez, Snorre Ralund, Elena Lloret, Paloma Moreda Pozo, and Armando Suárez Cueto. 2025. GPLSICORTEX at SemEval-2025 Task 10: Leveraging intentions for generating narrative extractions. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 578–586, Vienna, Austria. Association for Computational Linguistics.

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 Sivakumar, and Nafise Sadat Moosavi. 2024. Decoding news narratives: A critical analysis of large language models in framing bias detection. *ArXiv*, abs/2402.11621.

Jakub Piskorski, Nicolas Stefanovitch, Firoj Alam, Ricardo Campos, Dimitar Dimitrov, Alípio Jorge, Senja Pollak, Nikolay Ribin, Zoran Fijavz, Maram Hasanain, Purificação Silvano, Elisa Sartori, Nuno Guimarães, Ana Zwitter Vitez, Ana Filipa Pacheco, Ivan Koychev, Nana Yu, Preslav Nakov, and Giovanni Da San Martino. 2024. Overview of the CLEF-2024 checkthat! lab task 3 on persuasion techniques. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 299–310. CEUR-WS.org.

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

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023b. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski and Roman Yangarber. 2013. Information extraction: past, present and future. In *Multi-source, multilingual information extraction and summarization*, pages 23–49. Springer.

Pooja Premnath, Venkatasai Ojus Yenumulapalli, Parthiban Mohankumar, Rajalakshmi Sivanaiah, and Angel Deborah Suseelan. 2025. Techssn at semeval-2025 task 10: A comparative analysis of transformer models for dominant narrative-based news summarization. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Du Pengyuan Py, Huayang Li, Liang Yang, and Shaowu Zhang. 2025. DUTtask10 at semeval-2025 task 10: ThoughtFlow: Hierarchical narrative classification via stepwise prompting. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 427–433, Vienna, Austria. Association for Computational Linguistics.

R Raghav, Adarsh Prakash Vemali, Darpan Aswal, Rahul Ramesh, Parth Tusham, and Pranaya Rishi. 2025. TartanTritons at SemEval-2025 Task 10: Multilingual hierarchical entity classification and narrative reasoning using instruct-tuned LLMs. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1947–1956, Vienna, Austria. Association for Computational Linguistics.

Adith John Rajeev and Radhika Mamidi. 2025. adithjrajeev at SemEval-2025 Task 10: Sequential learning for role classification using entity-centric news summaries. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 245–250, Vienna, Austria. Association for Computational Linguistics.

Egil Rønningstad and Gaurav Negi. 2025. LTG at SemEval-2025 Task 10: Optimizing context for classification of narrative roles. In *Proceedings of the*

19th International Workshop on Semantic Evaluation (SemEval-2025), pages 442–449, Vienna, Austria. Association for 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.

Vineet Saravanan and Steven R. Wilson. 2025. cocoa at SemEval-2025 Task 10: Prompting vs. fine-tuning: A multilevel approach to propaganda classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 593–597, Vienna, Austria. Association for Computational Linguistics.

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.

Iknoor Singh, Carolina Scarton, and Kalina Bontcheva. 2025. GateNLP at SemEval-2025 Task 10: Hierarchical three-step prompting for multilingual narrative classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 147–153, Vienna, Austria. Association for Computational Linguistics.

Hareem Siraj, Kushal Chandani, Dua e Sameen, and Ayesha Enayat. 2025. NarrativeNexus at SemEval-2025 Task 10: Entity framing and narrative extraction using BART. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1409–1412, Vienna, Austria. Association for Computational Linguistics.

Mohammad S Sorower. 2010. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18(1):25.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, Purificação Silvano, Shivam Sharma, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari Gonçalves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, European Commission Joint Research Centre, Ispra (Italy).

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

Ling Sun, Xue Wan, Yuyang Lin, Fengping Su, and Pengfei Chen. 2025. PATeam at SemEval-2025 Task 10: Two-stage news analytical framework: Target-oriented semantic segmentation and sequence generation LLMs for cross-lingual entity and narrative analysis. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2104–2115, Vienna, Austria. Association for Computational Linguistics.

Qiangyu Tan, Yuhang Cui, and Zhiwen Tang. 2025. YNUzwt at SemEval-2025 Task 10: Tree-guided stagewise classifier for entity framing and narrative classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2116–2122, Vienna, Austria. Association for Computational Linguistics.

Rishit Tyagi, Rahul Bouri, and Mohit Gupta. 2025. Pbb-star at semeval-2025 task 10: Improving narrative classification and explanation via fine tuned language models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2208–2214, Vienna, Austria. Association for Computational Linguistics.

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

Xue Wan, Fengping Su, Ling Sun, Yuyang Lin, and Pengfei Chen. 2025. PATeam at SemEval-2025 task 9: LLM-augmented fusion for AI-driven food safety hazard detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1902–1908, Vienna, Austria. Association for Computational Linguistics.

Bo Wang, Ruichen Song, Xiangyu Wang, Ge Shi, Linmei Hu, Heyan Huang, and Chong Feng. 2025a. gowithnlp at SemEval-2025 Task 10: Leveraging entity-centric chain of thought and iterative prompt refinement for multi-label classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 377–384, Vienna, Austria. Association for Computational Linguistics.

Yutong Wang, Diana Nurbakova, and Sylvie Calabretto. 2025b. Team INSAntive at SemEval-2025 Task 10: Hierarchical text classification using BERT. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 984–991,

Vienna, Austria. Association for Computational Linguistics.

Arjumand Younus and Muhammad Atif Qureshi. 2025. nlptuducd at SemEval-2025 Task 10: Narrative classification as a retrieval task through story embeddings. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1735–1739, Vienna, Austria. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *CoRR*, abs/2203.01054.

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

A Suplementary Task and Corpus Information

This Section provides supplementary information on the tasks and datasets used for all three subtasks.

A.1 Entity Framing

In Table 13 we provide detailed counts of the distribution of all fine-grained roles across languages in the entire dataset for entity framing subtask, grouped according to our three main roles (Protagonist, Antagonist, and Innocent).

A.2 Narrative Classification

Figures 5 and 6 show the full taxonomies with all fine-grained sub-narrative labels for URW and CC domains.

Main Role	Fine Role	BG	EN	HI	PT	RU	TOTAL
Protagonist	Guardian	20	60	671	239	128	1118
	Martyr	9	15	10	3	3	40
	Peacemaker	39	20	174	89	101	423
	Rebel	24	28	194	22	17	285
	Underdog	5	20	196	10	0	231
	Virtuous	22	28	411	87	79	627
Antagonist	Instigator	90	104	118	143	94	549
	Conspirator	77	106	17	67	32	299
	Tyrant	72	78	164	50	23	387
	Foreign Adversary	112	78	483	256	199	1128
	Traitor	16	22	10	11	17	76
	Spy	0	5	21	0	4	30
	Saboteur	21	37	23	33	6	120
	Corrupt	37	115	24	55	11	242
	Incompetent	117	105	70	73	76	441
	Terrorist	30	36	41	81	79	267
	Deceiver	35	97	97	74	47	350
	Bigot	4	43	13	41	14	115
Innocent	Forgotten	4	3	29	9	3	48
	Exploited	19	16	75	14	41	165
	Victim	113	75	550	394	98	1230
	Scapegoat	3	16	20	9	10	58
	TOTAL	869	1107	3411	1760	1082	8229

Table 13: Distribution of fine-grained roles for each main role, grouped by language, and with total counts.

The fine-grained label distribution for the CC and URW domains is provided in Figures 7 and 8.

A.3 Narrative Extraction

We characterize the dataset at the level of the Subtask 3. Thus, we begin by presenting the dominant and sub-dominant narratives at document level for the dataset. Table 14 shows the number of documents for each dominant narrative in both URW and CC topics. The document count for subdominant narratives for CC and URW are presented in Tables 15 and 16, respectively.

We analyse the closeness of the explanations written by the annotators of the different languages. Towards that end, we extract the textual embeddings using a language-agnostic sentence embedding (LaBSE) (Feng et al., 2022) and applied t-SNE (van der Maaten and Hinton, 2008) for dimensionality reduction. The cluster of explanations by the two main topics (URW and CC) are presented in Figure 9. The figure shows that the explanations for both topics are well-separated, except for a small number of examples. In addition, train, dev, and test entries in both topics do not aggregate in a specific region of the cluster, thus demonstrating textual similarity between all the partitions of the dataset.

The task of Narrative Extraction is different from (though closely related in spirit to) the thoroughly studied task of Information Extraction (IE) (Piskorski and Yangarber, 2013)—in particular, regarding the comprehensive taxonomies of narratives (Hut-

Climate Change	
Dominant Narrative	**#Docs**
Amplifying Climate Fears	213
Criticism of institutions and authorities	92
Criticism of climate policies	53
Criticism of climate movement	48
Downplaying climate change	42
Hidden plots by secret schemes of powerful groups	30
Questioning the measurements and science	18
Controversy about green technologies	16
Climate change is beneficial	3
Green policies are geopolitical instruments	2
Ukraine-Russia War	
Dominant Narrative	**#Docs**
Discrediting Ukraine	277
Discrediting the West, Diplomacy	218
Praise of Russia	202
Amplifying war-related fears	175
Blaming the war on others rather than the invader	83
Speculating war outcomes	63
Russia is the Victim	54
Distrust towards Media	33
Negative Consequences for the West	28
Hidden plots by secret schemes of powerful groups	18
Overpraising the West	13

Table 14: Number of documents per dominant narrative at document level.

Climate Change	
Subdominant Narrative	**#Docs**
Amplifying Climate Fears: Amplifying existing fears of global warming	136
Criticism of institutions and authorities: Criticism of national governments	26
Criticism of institutions and authorities: Criticism of political organizations and figures	21
Criticism of institutions and authorities: Criticism of international entities	20
Criticism of climate policies: Climate policies have negative impact on the economy	20
Hidden plots by secret schemes of powerful groups: Climate agenda has hidden motives	13
Amplifying Climate Fears: Doomsday scenarios for humans	13
Hidden plots by secret schemes of powerful groups: Blaming global elites	12
Criticism of institutions and authorities: Criticism of the EU	12
Criticism of climate movement: Ad hominem attacks on key activists	12
Amplifying Climate Fears: Earth will be uninhabitable soon	10
Criticism of climate policies: Climate policies are ineffective	9
Questioning the measurements and science: Methodologies/metrics used are unreliable/faulty	8
Criticism of climate policies: Climate policies are only for profit	8
Questioning the measurements and science: Scientific community is unreliable	7
Downplaying climate change: Human activities do not impact climate change	6
Criticism of climate movement: Climate movement is alarmist	6
Criticism of climate movement: Climate movement is corrupt	5
Downplaying climate change: Weather suggests the trend is global cooling	4
Downplaying climate change: Ice is not melting	4
Downplaying climate change: Climate cycles are natural	4
Controversy about green technologies: Renewable energy is dangerous	4
Downplaying climate change: CO2 concentrations are too small to have an impact	2
Controversy about green technologies: Renewable energy is unreliable	2
Controversy about green technologies: Nuclear energy is not climate friendly	2
Amplifying Climate Fears: Whatever we do it is already too late	2
Questioning the measurements and science: Greenhouse effect/carbon dioxide do not drive climate change	1
Green policies are geopolitical instruments: Green activities are a form of neo-colonialism	1
Green policies are geopolitical instruments: Climate-related international relations are abusive/exploitative	1
Downplaying climate change: Temperature increase does not have significant impact	1
Climate change is beneficial: Temperature increase is beneficial	1
Climate change is beneficial: CO2 is beneficial	1

Table 15: Number of documents per subdominant narrative at document level for Climate Change topic.

tunen et al., 2002), which are inherently extensible as the domain evolves, whereas in IE the inventories of event types are expected to be more static.

A.3.1 Baseline model

The baseline model used for ST3 uses the following prompt to generate the texts in all languages:

Ukraine-Russia War	
Subdominant Narrative	**#Docs**
Discrediting Ukraine: Discrediting Ukrainian government and officials and policies	101
Praise of Russia: Praise of Russian military might	87
Amplifying war-related fears: There is a real possibility that nuclear weapons will be employed	61
Discrediting Ukraine: Discrediting Ukrainian military	55
Blaming the war on others rather than the invader: The West are the aggressors	47
Discrediting the West, Diplomacy: The West does not care about Ukraine, only about its interests	44
Praise of Russia: Russia has international support from a number of countries and people	42
Amplifying war-related fears: By continuing the war we risk WWIII	32
Blaming the war on others rather than the invader: Ukraine is the aggressor	32
Praise of Russia: Russia is a guarantor of peace and prosperity	32
Discrediting Ukraine: Ukraine is a puppet of the West	31
Distrust towards Media: Western media is an instrument of propaganda	25
Discrediting Ukraine: Situation in Ukraine is hopeless	24
Discrediting the West, Diplomacy: The West is weak	24
Russia is the Victim: The West is russophobic	23
Discrediting Ukraine: Ukraine is a hub for criminal activities	21
Amplifying war-related fears: Russia will also attack other countries	20
Discrediting the West, Diplomacy: Diplomacy does/will not work	19
Speculating war outcomes: Ukrainian army is collapsing	17
Discrediting the West, Diplomacy: The EU is divided	16
Speculating war outcomes: Russian army is collapsing	15
Praise of Russia: Praise of Russian President Vladimir Putin	13
Discrediting Ukraine: Ukraine is associated with nazism	11
Discrediting the West, Diplomacy: West is tired of Ukraine	11
Negative Consequences for the West: Sanctions imposed by Western countries will backfire	11
Amplifying war-related fears: NATO should/will directly intervene	10
Russia is the Victim: Russia actions in Ukraine are only self-defence	9
Overpraising the West: The West belongs in the right side of history	5
Discrediting Ukraine: Discrediting Ukrainian nation and society	4
Discrediting the West, Diplomacy: The West is overreacting	4
Discrediting Ukraine: Rewriting Ukraine's history	3
Praise of Russia: Russian invasion has strong national support	3
Russia is the Victim: UA is anti-RU extremists	3
Distrust towards Media: Ukrainian media cannot be trusted	2
Negative Consequences for the West: The conflict will increase the Ukrainian refugee flows to Europe	2
Overpraising the West: The West has the strongest international support	2
Speculating war outcomes: Russian army will lose all the occupied territories	2
Overpraising the West: NATO will destroy Russia	1

Table 16: Number of documents per subdominant narrative at document level for Ukraine-Russia War topic.

Given a news article along with its dominant and sub-dominant narratives, generate a concise text (maximum 80 words) supporting these narratives without the need to explicitly mention them. The explanation should align with the language of the article and be direct and to the point. If no sub-dominant narrative is selected, focus solely on supporting the dominant narrative. The response should be clear, succinct, and avoid unnecessary elaboration.
Dominant Narrative:(dominant narrative class)
Sub-dominant Narrative:(sub-dominant narrative class)
Article: (article text)

B General annotation guidelines

B.1 Subtask 1

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., 2021a) 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.

B.2 Subtask 2

The annotation for this subtask should be conducted according to the following procedure:

1. Annotators are provided a set of documents (articles) corresponding to a specific theme—Climate Change (CC) or the Ukraine-Russia War (URW)—along with a hierarchical domain-specific taxonomy consisting of two levels: coarse-grained labels (Narratives) and fine-grained labels (Sub-Narratives).

2. For each document, the annotator is required to read the text paragraph by paragraph. If a paragraph contains a Narrative from the taxonomy, the annotator highlights the first word of the paragraph (using the *"Narrative"* layer in INCEpTION) and selects the first applicable coarse-grained label. If no suitable coarse label is identified, the annotator skips the paragraph and proceeds to the next one, omitting steps 3 and 4.

3. If a coarse-grained label was selected, the annotator assigns an appropriate fine-grained Sub-Narrative from the available options under the chosen coarse-grained label. If no fine-grained label is applicable, the annotator selects the special label *"Other"* at the Sub-Narrative level. In cases where a coarse-grained Narrative is present but a fine-grained Sub-Narrative can not be determined, annotators are instructed to always assign *"Other"* as the fine-grained Sub-Narrative.

4. If an additional Narrative (or Sub-Narrative) is identified within the same paragraph, the process is repeated. The first word of the paragraph is highlighted again using the *"Narrative"* layer, and the corresponding (coarse-grained, fine-grained) label pair is selected accordingly.

5. Upon completing all paragraph-level annotations, the annotator determines the *Dominant Narrative* of the entire article—the Narrative that most prominently conveys the author's intent, in the annotator's judgment. To annotate this, the annotator applies the *"Dominant Narrative"* layer in INCEpTION, highlights the article's title (i.e., the first line), and assigns the appropriate `dominant_narrative` attribute. If no Narrative is present in any of the paragraphs, the annotator selects *"Other"* as the Dominant Narrative.

   The main distinction between paragraph-level and document-level annotation is that paragraph-level annotations require *two* fields to be completed: one for the coarse-grained Narrative and one for the fine-grained Sub-Narrative. In contrast, the Dominant Narrative annotation consists of a single field, where the annotator may select a fine-grained label, coarse-grained label, or *"Other"*. If a coarse-grained Narrative is chosen as the Dominant Narrative, this is equivalent to a paragraph annotation where the fine-grained Sub-Narrative is *"Other"*, indicating that a specific dominant Sub-Narrative could not be identified.

6. After identifying the Dominant Narrative, the annotator proceeds to annotate the **Evidence** layer by highlighting all textual segments that support the selection of the Dominant Narrative.

B.3 Subtask 3

The annotation task consists of writing concise—80-word long—explanations justifying their choice of dominant narrative and sub-narrative labels without explicitly naming them. To ensure clarity and consistency in the formulation of these explanations, annotators are provided detailed guidelines, as follows:

- The explanation, written in the same language as the article, should synthesize the textual evidence, encompassing arguments, counterarguments, behaviors, stances, or opinions that support the chosen dominant narrative.
- Annotators are required to justify their selection of the dominant narrative and sub-narrative by addressing the question: "Why were X and Y identified as the dominant narrative and sub-narrative?"
- Relevant entities mentioned in the article that contribute to the dominant and sub-narratives should be incorporated into the explanation.
- The explanation should be formulated in the annotator's own words, avoiding direct quotations, except for brief phrases or expressions, and is limited to a maximum of 80 words.

Additionally, annotators are provided with style recommendations to refine their justifications:

- Where possible, annotators are encouraged to explicitly reference entities, along with their actions or statements, to substantiate the selected narratives.
- In cases where explicit entities, actions, or statements are unavailable, annotators are advised to use neutral formulations such as "*the text reports*" or "*the text's author*" to support their reasoning.
- Annotators are instructed to avoid merely *restating* the dominant and sub-narratives, and rather focus on providing a reasoned justification for their selection.

C Annotation Platform

INCEpTION (Klie et al., 2018) is a web application, designed primarily for tasks such as semantic annotation (e.g., concept linking, fact linking), but can be customized for other purposes. For this task we adapted INCEpTION according to the annotation process described in Appendix B. An example of the customized instance can be seen in Figure 10.

In total, we created five projects, one for each language, so the teams could work independently. A more in-depth explanation of the platform and its use can be found in the dedicated section in the annotation guidelines (Stefanovitch et al., 2025).

D Annotation Complexity

Inter-Annotator Agreement is measured using Krippendorf's α and computed using the `simpledorff` library. In Table 17, we give a detailed breakdown of the IAA for subtask 2: we consider both coarse and fine-grained levels for all languages. This allows us to better understand the annotation complexity.

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 labels with low agreement—5 for URW and 7 for CC—that achieved disagreement above 40% and 60%, respectively. If we exclude 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.

Looking into the data further, we observe that the majority of the disagreement between two annotators in the sub-Narrative labels was due to labels of the same main Narrative (e.g., different sub-Narratives under "*Discrediting the West, Diplomacy*" that were frequently confused with one another)

On average, of all individual annotator disagreements (paragraphs where the two annotators picked a different sub-Narrative), 67% were sub-Narratives of the same Narrative.

Some sub-Narratives 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 interests*", and "*Discrediting Ukraine: Ukraine is a puppet of the West*".

E Participant Systems

We next list the systems of all participants who submitted a system description paper. The team name who made the submission is in bold; if the team used a different name on the leaderboard, it is shown in parentheses; the list of subtasks the team

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 17: Krippendorff's α for different granularities, languages and domains on paragraph level for subtask 2

participated in is given in brackets; if the team ranked first for some subtask-language pair, the list of all pairs where it ranked first is given; a list of keywords; and finally, a short description of the approach.

adithjrajeev [ST1] (Rajeev and Mamidi, 2025) (Keywords: *BERT, DeBERTa, Summarization, CTRLsum, Gemini 1.5 Flash*) The authors propose a two-stage pipeline for the role classification: The first stage includes entity-centric summarization to condense the context around a given entity using CTRLsum and a prompt-based LLM approach for comparison. The second stage performs role classification with the summary and the entity as input. They fine-tuned BERT and DeBERTa with a dual training strategy where the main and fine-grained roles are optimized sequentially. The authors enhance fine-grained classification through a contrastive learning objective that aligns entity representations with role descriptions.

bbStar [ST2, ST3] (Tyagi et al., 2025) (Keywords: *BERT, GPT4-o, ReACT, few-shot prompting, knowledge injection*)

For ST2, the authors fine-tuned a BERT model with a recall-oriented approach. This ensured that subtle and implicit narratives were captured comprehensively, even at the cost of introducing some noise. In post-classification, they refined the predictions using a GPT-4o pipeline, which enhances consistency and contextual coherence by filtering out misclassifications and ensuring that the detected narratives align with the overall article theme.

For ST3, to generate concise evidence-based explanations of dominant narratives, the authors implemented a ReACT (Reasoning + Acting) framework that leverages semantic retrieval-based few-shot prompting. To enhance factual accuracy and mitigate hallucinations, they incorporate a structured taxonomy table as an auxiliary knowledge base.

BERTastic [ST1] (Mahmoud et al., 2025b) (Keywords: *GPT-4o, XLM-RoBerta, Least-to-most prompting, Sentence Splitting*) The authors explore two approaches for role classification: (1) LLM prompting with GPT-4 and (2) fine-tuning XLM-R. For prompting, they compare single-step predictions against multi-step, least-to-most, and hierarchical prompting strategies, addressing both main and fine roles. For fine-tuning, they conduct a comparative study on different levels of contextual granularity surrounding an entity mention and assess performance in monolingual versus multilingual settings. Additionally, they investigate the impact of training on main roles versus fine roles. Their best-performing system, which achieved the highest accuracy on main roles in Hindi, was trained on fine roles across all languages using sentence-level context.

clujteam [ST3] (Marginean, 2025) (Keywords: *SmoLM2, Prompt Engineering*) The authors fine-tuned SmolLM2 360M and 1.7B to generate explanations. The Narrative taxonomy is used to customize the system prompt according to the given narrative/sub-narrative. The Definition included in the taxonomy is added to the system prompt to guide the model towards the statements that justify the presence of the narrative.

cocoa [ST1] (Saravanan and Wilson, 2025) (Keywords: *BERT, DistilBERT, GPT-3, GPT-3.5*) The authors investigate two primary approaches: (1) prompt-based classification using large language models (LLMs) like GPT and (2) fine-tuning transformer-based models, where they employ a hierarchical structure: a model first classifies the main propaganda category, and three other models classify the subcategory. Their results indicate that while LLMs demonstrate some generalization ability, fine-tuned models significantly outperform them in accuracy and reliability, reinforcing the importance of task-specific supervised learning for propaganda detection.

COGNAC [ST2] (Islam and Finlayson, 2025) (Keywords: *GPT-4o-mini,hierarchical prompting structure, CoT*) The authors address ST2 by (1) summarization that condenses the news articles, making inputs more uniform in length and style; (2) a set of zero-shot, class-specific LLM prompts, including CoT, to produce binary outputs for each top-level narrative class; and (3) hierarchical prompting to sequentially identify sub-narrative classes only when the corresponding narrative classes are detected using GPT-4o-mini.

DEMON [ST1] (Fenu et al., 2025) (Keywords:

QLoRA, Llama 3, BERT) The authors propose a Llama fine-tuning approach using QLoRA based on a data preparation phase and subsequent training to optimize the model for the specific task. In particular, the pre-processing phase aims to identify the portion of the article that is useful for correct classification. The model version used is the 8B parameter Llama 3.

DUTtask10 [ST2] (Py et al., 2025) (Keywords: *GPT-4o, Qwen 2.5, Chain-of-Thought, Data augmentation*) The authors propose a two-step hierarchical narrative classification process. The first step leverages a large pre-trained model to generate a reasoning (or thought) process based on the given news article, helping the model grasp the broader context. In the second step, they fine-tune the model to perform sub-narrative classification, ensuring more accurate and contextually relevant categorization. This approach combines the generative strengths of a large pre-trained model with fine-tuning to enhance sub-narrative classification.

DUTIR [ST1] (Lv et al., 2025) (Keywords: *QLoRA, Chain-of-Though, Ensemble, GLM-4-Plus, Qwen2.5, Llama3.1, Data augmentation*) The authors propose a framework based on LLMs for multilingual entity framing in news articles that integrates multilingual translation, synonym-based data augmentation to address class imbalance, and fine-tuning multiple base models using QLoRA. The predictions from these models are aggregated via Chain-of-Thought ensemble with GLM-4-Plus serving as the meta-classifier.

Fane [ST1] (Fane et al., 2025) (Keywords: *Zero-shot learning, prompt engineering, hierarchical prompting, O1-mini, GPT-4o*) The approach employs Multi-Step classification, beginning with the classification of the main roles (Protagonist, Antagonist, Innocent), followed by fine-grained roles. They explore various input contexts, including full text, entity-specific sentences, neighboring sentences, and framing-preserved summaries. Their prompting strategies include role/persona-based prompting, incorporating label definitions, and generating justifications alongside labels. For the official submission, they utilized a setup combining Full-Text input, Expert Persona prompting, including label definitions, and a Multi-Step classification approach, using the OpenAI O1 (o1-2024-12-17) model for English, and GPT-4o for other languages.

GATENLP [ST2] (Singh et al., 2025) (Keywords: *RoBerta, Llama 3.1, Data Augmentation*) The authors propose Hierarchical Three-Step Prompting (H3Prompt) for multilingual narrative classification. Their approach fine-tunes LLaMA using H3Prompt, incorporating both the provided training data and synthetically generated data. The method follows a structured, three-step prompting framework to ensure a hierarchical classification process, progressively refining predictions at each stage.

GPLSICORTEX [ST3] (Martínez-Murillo et al., 2025) (Keywords: *GPT-4o mini, Llama 3, FLAN-T5, Instruction vanilla*) The authors propose a narrative-aware approach that enhances explanation generation by incorporating the underlying intention and structure of texts into pre-trained models. By explicitly modeling the purpose of a text, their system produces more meaningful and contextually relevant explanations. They experimented with various instruction-tuned models, including LLaMa 3 and Flan-T5, with the latter achieving the best results. Their approach secured 3rd place in the competition.

gowithnlp [ST1] (Wang et al., 2025a) (Keywords: *CoT, GPT-3.5-Turbo,GPT-3, Claude*) The approach iteratively refines prompts and utilizes Entity-Centric Chain of Thought. Specifically, to minimize ambiguity in label definitions, they use the model's predictions as supervisory signals, iteratively refining the category definitions. Furthermore, to minimize the interference of irrelevant information during inference, they incorporate entity-related information into the CoT framework, allowing the model to focus more effectively on entity-centric reasoning.

HowardUniversityAI4PC [ST1] (Aryal and Dhungana, 2025) (Keywords: *Mistral-7B, Phi-4, Llama 3.1, Gemma 2, DeepSeek R1, Instruction vanilla, Synthetic prompting*) The authors employ an ensemble-based approach for role assignment, utilizing multiple state-of-the-art LLMs, including LLaMA 3.1-8B, Mistral-7B, Phi-4, and Gemma 2. Through prompt engineering, they optimize each model's output using Ollama's API to generate structured responses for named entities across all articles and languages. The outputs are stored as text files and subsequently combined to produce a final submission. This multi-model strategy enables them to achieve strong performance metrics by leveraging the complementary strengths of different LLMs.

iLostTheCode [ST2] (Concas et al., 2025) (Key-

words: *RoBERTa, DeBERTa, DistilBERT, MLP*) The authors propose a model that leverages multiple pre-trained models in parallel to create enriched embeddings fed into a simple machine learning model. The dataset is first translated and processed so that each sentence is treated as an individual sample, independently or with a small contextual window, depending on the language. Each sentence is passed through different models (BERT variants), and their resulting embeddings are concatenated to form a composite feature vector. This vector is then used as input for the neural network, which outputs classification probabilities. Finally, a post-processing module aggregates the probabilities of all sentences within a file and applies a threshold to produce the final classification predictions.

Irapuarani [ST2] (Assis et al., 2025) (Keywords: *GPT-4o mini, DeBERTa, mDeBERTa, Aya Expanse 8B, Instructions vanilla, Translation*) The authors explore three strategies combining Small Language Models (SLMs) and Large Language Models (LLMs) for hierarchical multi-label classification. The first approach applies a multilingual SLM for direct classification without hierarchical constraints. The second leverages an LLM for text translation into a single language before classification with a monolingual SLM. The third adopts a hierarchical strategy where an SLM filters domains, and an LLM assigns final labels. Among these, the translation-based approach proves the most generalizable across languages, improving label alignment and reducing inconsistencies caused by imbalanced label representation across different languages.

IRNLP [ST2] (Kiousis, 2025) (Keywords: *XLM-RoBERTa, DeepPavlov, Neuralmind BERT*) The authors' approach to multilingual narrative classification is based on XLM-RoBERTa Large and other bert-based models, e.g, DeepPavlov and Neuralmind BERT, fine-tuned on different language datasets. To improve generalization and ensure robust performance across languages, they employed a repeated k-fold cross-validation strategy. Their preprocessing pipeline included (1) language-specific tokenization, (2) hierarchical label structuring, and (3) dynamic batch sampling to balance label distributions. The results demonstrated that the chosen approach effectively leveraged transformer-based architectures to model complex narrative structures across languages, with strong performance gains due to repeated k-fold evaluation.

INSALyon2 [ST2] (Eljadiri and Nurbakova, 2025) (Keywords: *Zero-shot, Agentic framework*) The authors propose an agentic framework where each agent functions as a specialized binary classifier. Each agent is responsible for detecting whether a given text belongs to a specific narrative or a sub-narrative. They use AutoGen to coordinate multiple LLM agents, organized as a group chat with a user proxy agent, manager agent, and multiple narrative and sub-narrative agents. The manager limits narrative agents to a single query per classification, while the user agent initiates a group chat for every new text sample. All models were used in a zero-shot setting, with GPT-4o as the primary classification agent and GPT-4o Mini as the user proxy agent.

INSAntive [ST2] (Wang et al., 2025b) (Keywords: *BERT, translation*) The authors' framework provides a range of functional modules—including segmentation, automated translation, and standardized output—that facilitate the generation of high-quality multilingual data for subsequent classification and semantic analysis. In the multi-label setting, the framework integrates a BERT-based text classification method, utilizing automated data processing, optimized training workflows, and memory management strategies.

KostasThesis2025 [ST2] (Eleftheriou et al., 2025) (Keywords: *Continual Learning, MLP, Hierarchical classification, Ensemble, KaLM, Stella*) The authors focus on hierarchical multi-label, multi-class classification in multilingual news articles. They present an architecture that combines narrative predictions with multiple sub-narrative heads using concatenation. They experimented with different embeddings, KaLM and Stell, with KaLM outperforming Stella. The best results were achieved with the Continual Learning model with Concatenation architecture.

KyuHyunChoi [ST3] (Choi and Na, 2025) (Keywords: *PEGASUS*) The authors employ PEGASUS, a transformer-based model pre-trained specifically for summarization using the Gap Sentence Generation (GSG) method. PEGASUS large was chosen for this study, as it was trained solely with GSG, given the ineffectiveness of MLM. Fine-tuning followed a standard procedure, where training set inputs were processed by the encoder, and outputs were generated via the decoder. No special techniques were applied during fine-tuning. The model was saved at the checkpoint with the highest

BertScore F1 score.

LATE-GIL-nlp [ST1, ST2, ST3] (Diaz et al., 2025) (Keywords: *RoBERTa, XLM-RoBerta, Flan-T5, Llama 3.1, Sentence-Transformers, TweetNLP, Data augmentation*) For ST1, the authors propose a three-stage pipeline for the Entity Framing track, ensuring consistency across five languages. Context extraction captures 18 words around each entity and refines it using Llama 3.1 8B to generate English contexts. Multi-class classification fine-tunes RoBERTa with role-based labels, incorporates sentiment augmentation, and undergoes additional fine-tuning for each language. Multi-label classification preprocesses text, fine-tunes sentence transformers per language, and assigns multiple emotion labels. Finally, a K-Nearest Neighbors classifier is trained using cross-validation to classify entities based on their contextual embeddings.

For ST2, they propose a multilingual classification pipeline with different setups for Bulgarian, English, Hindi, and Russian. For the first three languages, a three-stage pipeline first classifies "Other" vs. narratives, then identifies narratives, and finally classifies sub-narratives. The Russian pipeline omits the "Other" label, using a two-stage process for narrative and sub-narrative classification. Both variants use an XLM-RoBERTa backbone for multilingual adaptability. The approach is tailored to varying label distributions across languages.

For ST3, they apply a two-step data cleaning process, removing unwanted prefixes from annotations and omitting article titles while handling duplicates. The training dataset is prepared by retrieving article content for each annotation and applying a predefined prompt. A pre-trained Google FLAN-T5 model is fine-tuned using the Hugging Face Trainer API with specific hyperparameters. Explanation generation involves extracting key sentences using spaCy and NLTK to create structured summaries. This approach ensures cleaner data, effective training, and improved text-to-text generation.

LTG [ST1] (Rønningstad and Negi, 2025) (Keywords: *XLM-RoBerta, Llama 3, Mistral-7B*) The authors investigate the optimal text segments to extract from newspaper articles to capture an entity's narrative role while minimizing distractions. Their approach is evaluated using XLM-RoBERTa large and compared against supervised fine-tuning of smaller generative language models. They investigate the optimal text segments to extract from newspaper articles to capture an entity's narrative role while minimizing distractions. Their approach is evaluated using XLM-RoBERTa-large and compared against supervised fine-tuning of generative language models. By optimizing text selection, they find that XLM-RoBERTa-large outperforms fine-tuning larger language models trained on the entire texts.

NarrativeMiners [ST1, ST2, ST3] (Khubaib et al., 2025) (Keywords: *Gemini, Mistral, Translation,*) In ST1, the authors use multiple data augmentation strategies: (1) generating similar articles with the same entities using Gemini and Mistral and (2) translating into English, the former not showing promising results. They experimented with BERT, DeBERTa, and BART, with BART-CNN emerging as the best-performing model.

In ST2, the authors applied back-translation to increase the dataset size. They experimented with BERT model fine-tuning using a multi-stage approach: (1) the first model was on topic classification - URW, CC, or Other; (2) the second classified articles in the narrative for the respective topic; (3) finally, for the predicted narrative, a sub-narrative classification model was used.

In ST3, the authors fine-tuned FLAN-T5, GPT-2, and BART-CNN. Compared to BART-CNN, GTP-2 and FLAN-T5 significantly underperformed, struggling with generating coherent and contextually grounded explanations.

NarrativeNexus [ST1, ST3] (Siraj et al., 2025) (Keywords: *BART*) In ST1, the authors employ a BART-based sequence classifier to identify and categorize named entities within news articles, mapping them to predefined roles such as protagonists, antagonists, and innocents. More specifically, their approach involved fine-tuning BART-large with hyperparameter optimization, data augmentation techniques, and confidence thresholding to improve classification reliability.

In ST3, the authors fine-tuned BART-large-cnn using a text-to-text generative paradigm to generate justifications for dominant narratives. To enhance factual consistency, they introduced a filtering step to discard low-confidence justifications. Their evaluation relied on BLEU and ROUGE scores to measure output fluency and relevance.

Narrlangen [ST2] (Blombach et al., 2025) (Keywords: *Hierarchical classification, SetFit, XLM-RoBERTa*) The authors experimented with several approaches: (1) fine-tuning encoder models, (2) hi-

erarchical classification using encoder models with two different classification heads, (3) direct classification of fine-grained labels using SetFit, (4) a zero-shot approach based on sentence similarities, and (5) prompt engineering of LLMs. Their best approach was fine-tuning a pre-trained multilingual model, XLM-RoBERTa, with two additional linear layers and a softmax on top as a classification head. They fine-tuned their multilingual model on the combined data set of all languages.

NCLTeam [ST2] (Li et al., 2025b) (Keywords: *BERT, ModernBERT, BART, all-MiniLM-L12-v2, CU-Net, Data augmentation*) The authors propose a hierarchical model architecture that aligns with the dataset taxonomy, leveraging a pretrained all-MiniLM-L12-v2 encoder and a cascaded UNet for text classification. The model jointly optimizes both components, using the last hidden state as input to the UNet. Conditional subcategory pathways refine classification, first distinguishing "Climate Change (CC)", "Ukraine-Russia War (URW)", and "Others", then further classifying CC and URW narratives. Contrastive learning enhances feature representation with positive-negative pairs and mix-up regularization. A cosine embedding loss improves intra-class similarity while ensuring distinct separation of negative samples.

NlpUned [ST1] (Caballero et al., 2025) (Keywords: *GPT-4o, Chain-of-Thought, Hierarchical classification, Summarization*) This study explores a prompt-based, non-hierarchical approach to fine-grained role classification in news narratives using Large Language Models (LLMs). Instead of traditional model training or fine-tuning, the system relies on zero-shot and few-shot prompting, leveraging structured taxonomies and contextual signals to classify named entities into fine-grained sub-roles.

nlptuducd [ST2] (Younus and Qureshi, 2025) (Keywords: *Mistral 7B, synthetic generated data*)

The system authors propose the application of a Mistral 7B model, specifically E5 model, to address the ST2 in English. Their approach frames the task as a retrieval task in a similarity-matching framework instead of relying on supervised learning. Specifically, each test article's top two similar articles are first retrieved with cosine similarity, and from those, the one with a narrative alignment is extracted using story embeddings (an embedding framework on top of Mistral-7B via synthetically generated data).

NotMyNarrative [ST2] (Faye et al., 2025) (Keywords: *XLM-RoBERTa, mDeBERTa, ModernBERT, Albertina PT-PT, MuRIL, SlavicBERT, Muril*) The authors compare multilingual models (XLM-RoBERTa, mDeBERTa) with monolingual ones and observe that, given the limited data per label per language, multilingual models perform better and can leverage information from all languages to improve general performance. Furthermore, the authors conducted an ablation study by leaving out a single language in training and then testing on all languages. Results show that XLM-RoBERTa generalizes better than mDeBERTa on new languages.

PATeam [ST1, ST2, ST3] (Sun et al., 2025) (Keywords: *Qwen2.5, Phi-4, Multi-prompting,LoRa,Data Augmentation, DPO, SFT Synthetic Data Generation*)

In ST1, the authors propose a two-stage pipeline system that enhances the accuracy of role classification for location entities in news articles. This system comprises three key components: 1) Qwen2.5-72B model leverages multi-prompt engineering techniques to focus contextual analysis on target location entities to dynamically restructure the input text around these entities, thereby reducing noise and enhancing semantic coherence. 2) Phi-4 and Qwen2.5-32B models are fine-tuned using LoRA to specialize in multi-turn conversational reasoning, enabling a nuanced understanding of role-specific patterns at both coarse and fine granularities through sequential interaction analysis. 3) Through systematic ablation studies across multiple experimental configurations, the authors evaluate the comparative effectiveness of monolingual and multilingual approaches, deriving actionable implementation guidelines. 4) The ensemble prediction was decided by majority voting, and very few cases were handled by selecting the model with the best validation performance.

In ST2, the team adopts a similar pipeline for narrative-based semantic segmentation. For each news article, they obtain a list of relevant paragraphs for each sub-narrative in its golden label set. Then, two types of models for multilabel classification are trained, the one-vs-rest classification models and the label sequence generation models. Therefore, two data aggregation approaches are employed to convert the above data into proper training data for different types of models, respectively.

In ST3, the authors used Phi-4 as the base model, with data augmentation and direct preference optimization. In addition, the authors also used synthetic data generated from Qwen2.5-72B and Llama3-70B, which proved particularly effective for lesser-known languages. They conducted experiments with several training techniques, including SFT (supervised fine-tuning), DPO, SimPO, and ORPO. The best results were achieved with the combination of SFT and DPO.

QUST [ST1] (Liu et al., 2025) (Keywords: *DeBERTa, Qwen2.5, Phi-3, Phi-4, Ensemble, GLM4*) The authors fine-tune several models, including DeBERTa, Qwen 2.5, Phi-3, Phi-4, and GLM4. For their final submission, they utilize an ensemble learning strategy that employs hard voting to combine predictions of the top 3 selected models for each language, enhancing the prediction accuracy of the final result.

TartanTritons [ST1, ST3] (Raghav et al., 2025) (Keywords: *RoBERTa, Phi-4, Llama 3.1, Instructions vanilla, Chain-of-Thought, Active prompting*) The authors present a hierarchical role extraction system built on Microsoft's Phi-4, a quantized and instruction-tuned LLM. Their approach integrates multiple techniques to enhance accuracy and robustness. By leveraging instruction tuning with a predefined taxonomy of fine-grained roles, they achieve notable performance gains. To improve entity disambiguation, they introduce special 'entity' tags, allowing the model to differentiate multiple mentions within the text. Additionally, they enhance cross-lingual performance by training on multilingual datasets. An iterative feedback mechanism further refines predictions, where model-generated error messages guide retries to improve output quality.

TECHSSN [ST3] (Premnath et al., 2025) (Keywords: *BART, DistilBART, T5, FalconAI*) The authors propose fine-tuning pre-trained summarization models using the Seq2SeqTrainer from the Hugging Face Transformers library. The model used to tackle ST3 was a fine-tuned version of DistilBART (distilbart-cnn-12-6).

Tuebingen [ST1] (Karabulut et al., 2025) (Keywords: *BERT, Data Augmentation, External Knowledge*) The authors evaluate transformer-based models(BERT-family) with minimal hyperparameter tuning to analyze their impact on classification performance. The authors also incorporated class weighting to address class imbalance and explored additional techniques, such as data augmentation and the integration of external information, to improve model robustness and enhance overall performance.

UNEDTeam [ST2] (Fraile-Hernandez and Peñas, 2025) (Keywords: *Calme-2.4-rys-78B*) The authors employ a zero-shot approach, using the knowledge embedded in Large Language Models (LLMs), specifically MaziyarPanahi/calme-2.4-rys-78b, without relying on training examples. To address linguistic barriers, they translate all news items into English using OPUS machine translation models. Classification occurs in two stages using prompts: first, each news item is categorized into one of the two main thematic categories (Climate Change or Ukraine-Russia War). Then, within each category, sub-narratives are identified, with the option to label the news item as "Other" if it does not align with any predefined sub-narrative.

WordWiz [ST3] (Ahmadi and Zeinali, 2025) (Keywords: *instruction-tuning, Phi-3.5, DFT*) The authors employed a combination of targeted preprocessing techniques and instruction-tuned language models to generate concise, accurate narrative explanations across five languages. Their approach leverages an evidence refinement strategy that removes irrelevant sentences, improving signal-to-noise ratio in training examples. They fine-tuned Microsoft's Phi-3.5 model using Supervised Fine-Tuning (SFT). During inference, they implemented a multi-temperature sampling strategy that generates multiple candidate explanations and selects the optimal response using narrative relevance scoring.

YNU-HPCC [ST1] (Li et al., 2025a) (Keywords: *DeBERTa*) The authors propose a two-stage role classification model based on DeBERTa. The proposed model integrates the deep semantic representation of the DeBERTa pre-trained language model through two sub-models: main role classification and sub-role classification, and utilizes Focal Loss to optimize the category imbalance issue.

YNUzwt [ST1,ST2] (Tan et al., 2025) (Keywords: *CoT, Phi-3.5, GPT4-o*)

The authors present a Tree-guided Stagewise Classifier with Chain of Thought to tackle ST1 and ST2 in multiple languages. This algorithm uses Hierarchical Reasoning to overcome the limitations of zero-shot classifiers guiding the LLM through the hierarchical structural annotation in ST1 and ST2. The authors experimented with this algorithm in two large language models: GPT4-o and Phi-3.5.

The following systems are listed on the official leaderboard of the Shared Task, but no paper was submitted: Synapse (ST3), Mendel292A (ST3), UMZNLP (ST3), ftd (ST3). We have not received short descriptions of the following systems: YNUzwt (ST3), DUTtask10 (ST3).

```
Other
Blaming the war on others rather than the invader
- Ukraine is the aggressor
- The West are the aggressors
Discrediting Ukraine
- Rewriting Ukraine's history
- Discrediting Ukrainian nation and society
- Discrediting Ukrainian military
- Discrediting Ukrainian government and officials and policies
- Ukraine is a puppet of the West
- Ukraine is a hub for criminal activities
- Ukraine is associated with nazism
- Situation in Ukraine is hopeless
Russia is the Victim
- The West is russophobic
- Russia actions in Ukraine are only self-defence
- UA is anti-RU extremists
Praise of Russia
- Praise of Russian military might
- Praise of Russian President Vladimir Putin
- Russia is a guarantor of peace and prosperity
- Russia has international support from a number of countries and people
- Russian invasion has strong national support
Overpraising the West
- NATO will destroy Russia
- The West belongs in the right side of history
- The West has the strongest international support
Speculating war outcomes
- Russian army is collapsing
- Russian army will lose all the occupied territories
- Ukrainian army is collapsing
Discrediting the West, Diplomacy
- The EU is divided
- The West is weak
- The West is overreacting
- The West does not care about Ukraine, only about its interests
- Diplomacy does/will not work
- West is tired of Ukraine
Negative Consequences for the West
- Sanctions imposed by Western countries will backfire
- The conflict will increase the Ukrainian refugee flows to Europe
Distrust towards Media
- Western media is an instrument of propaganda
- Ukrainian media cannot be trusted
Amplifying war-related fears
- By continuing the war we risk WWIII
- Russia will also attack other countries
- There is a real possibility that nuclear weapons will be employed
- NATO should/will directly intervene
Hidden plots by secret schemes of powerful groups
```

Figure 5: Ukraine War label taxonomy

```
Other
Criticism of climate policies
- Climate policies are ineffective
- Climate policies have negative impact on the economy
- Climate policies are only for profit
Criticism of institutions and authorities
- Criticism of the EU
- Criticism of international entities
- Criticism of national governments
- Criticism of political organizations and figures
Climate change is beneficial
- CO2 is beneficial
- Temperature increase is beneficial
Downplaying climate change
- Climate cycles are natural
- Weather suggests the trend is global cooling
- Temperature increase does not have significant impact
- CO2 concentrations are too small to have an impact
- Human activities do not impact climate change
- Ice is not melting
- Sea levels are not rising
- Humans and nature will adapt to the changes
Questioning the measurements and science
- Methodologies/metrics used are unreliable/faulty
- Data shows no temperature increase
- Greenhouse effect/carbon dioxide do not drive climate change
- Scientific community is unreliable
Criticism of climate movement
- Climate movement is alarmist
- Climate movement is corrupt
- Ad hominem attacks on key activists
Controversy about green technologies
- Renewable energy is dangerous
- Renewable energy is unreliable
- Renewable energy is costly
- Nuclear energy is not climate-friendly
Hidden plots by secret schemes of powerful groups
- Blaming global elites
- Climate agenda has hidden motives
Amplifying Climate Fears
- Earth will be uninhabitable soon
- Amplifying existing fears of global warming
- Doomsday scenarios for humans
- Whatever we do it is already too late
Green policies are geopolitical instruments
- Climate-related international relations are abusive/exploitative
- Green activities are a form of neo-colonialism
```

Figure 6: Climate Change label taxonomy

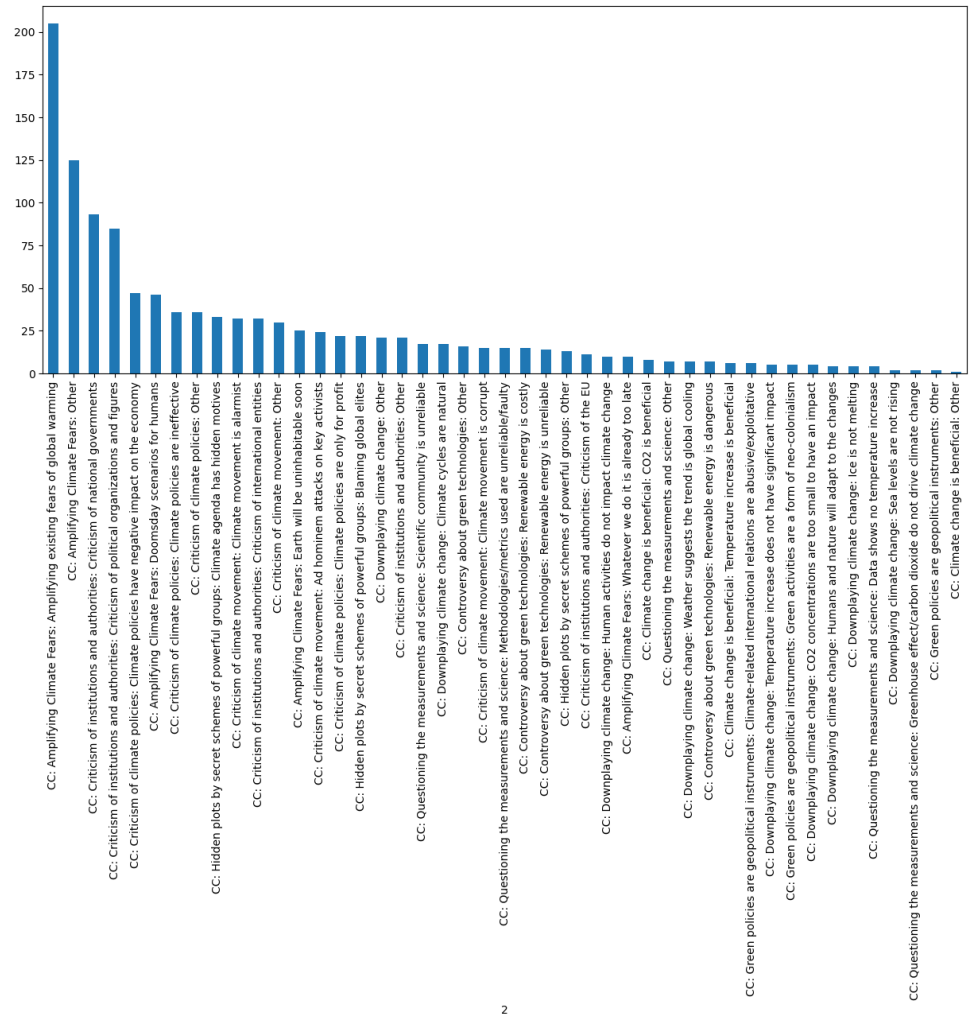

Figure 7: Label distribution statistics for the labels of Subtask-2 for Climate Change subset.

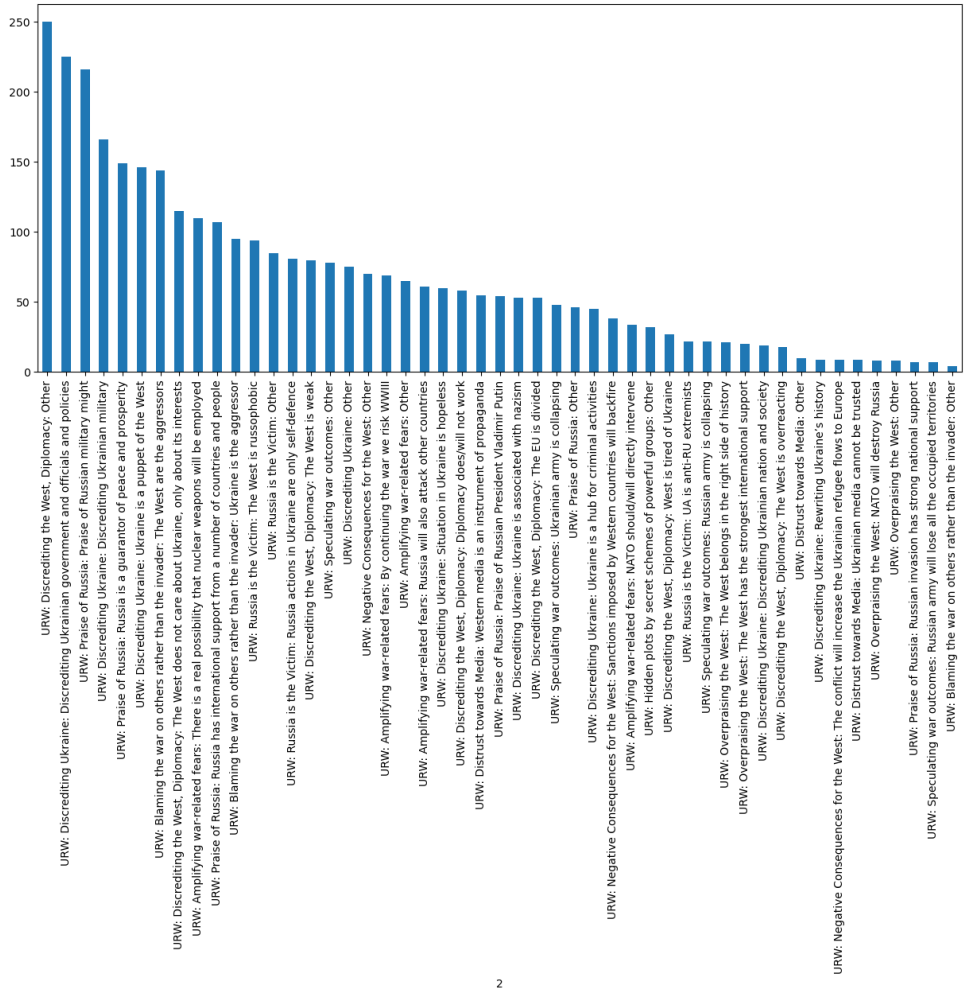

Figure 8: Label distribution statistics for the labels of Subtask-2 for Ukraine Russia War.

Figure 9: Similarity of the explanations by topic (URW in blue and CC in red) using LaBSE. In addition, the explanations of each split are represented by different symbols (train: circle, dev: diamond, test: square)

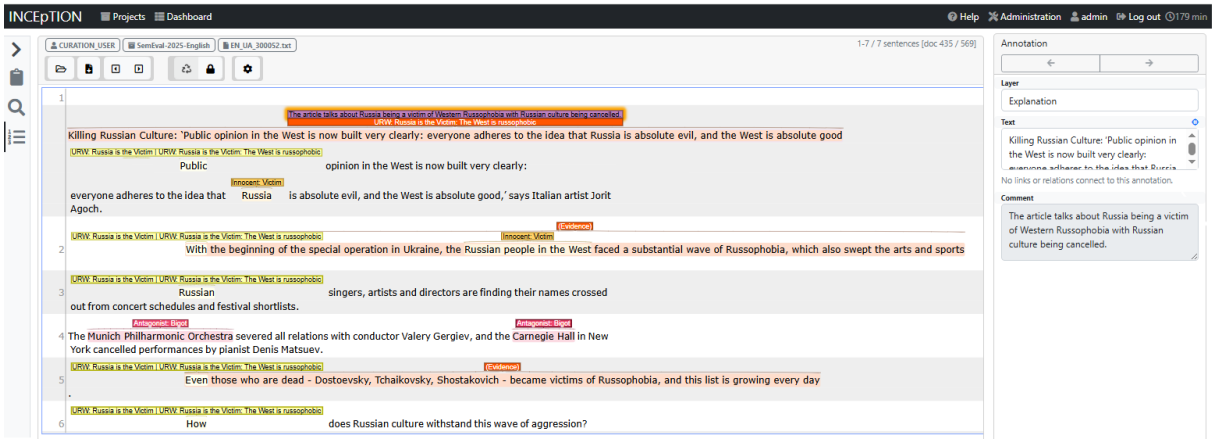

Figure 10: Annotated example from the English portion of our dataset. Entity framing is indicated in warm yellow, while narratives and sub-narratives at the paragraph level are highlighted in yellow. In the title of the article, the Dominant Narrative (*Russia is the victim*) and Sub-narrative (*The West is russophobic*) are highlighted in orange, while the explanation is highlighted in purple. Within the body of the article, the Evidence is highlighted in light orange.