7-2023

SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup

Jakub Piskorski
Polish Academy of Sciences

Nicolas Stefanovitch
European Commission Joint Research Centre

Giovanni Da San Martino
Università degli Studi di Padova

Preslav Nakov
Mohamed Bin Zayed University of Artificial Intelligence

Recommended Citation

J. Piskorski et al., "SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup," *17th International Workshop on Semantic Evaluation, SemEval 2023 - Proceedings of the Workshop*, pp. 2343 - 2361, Jul 2023.
The definitive version is available at https://doi.org/10.18653/v1/2023.semeval-1.317

SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup

Jakub Piskorski[1], **Nicolas Stefanovitch**[2], **Giovanni Da San Martino**[3], **Preslav Nakov**[4]
[1]Institute of Computer Science, Polish Academy of Science, Poland jpiskorski@gmail.com
[2]European Commission Joint Research Centre, Italy nicolas.stefanovitch@ec.europa.eu
[3]Department of Mathematics, University of Padova, Italy dasan@math.unipd.it
[4]Mohamed bin Zayed University of Artificial Intelligence, UAE preslav.nakov@mbzuai.ac.ae

Abstract

We describe SemEval-2023 task 3 on *Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multilingual Setup*: the dataset, the task organization process, the evaluation setup, the results, and the participating systems. The task focused on news articles in nine languages (six known to the participants upfront: *English*, *French*, *German*, *Italian*, *Polish*, and *Russian*), and three additional ones revealed to the participants at the testing phase: *Spanish*, *Greek*, and *Georgian*). The task featured three subtasks: (1) determining the genre of the article (opinion, reporting, or satire), (2) identifying one or more frames used in an article from a pool of 14 generic frames, and (3) identify the persuasion techniques used in each paragraph of the article, using a taxonomy of 23 persuasion techniques. This was a very popular task: a total of 181 teams registered to participate, and 41 eventually made an official submission on the test set.

1 Introduction

The widespread use of Internet and the advances in Internet-related technologies paved the way to easily create direct communication channels between information producers and consumers, potentially leaving the latter exposed to manipulative, propagandistic, and deceptive content. Given the potentially huge audience that can be reached through online channels, major public events and debates revolving around relevant topics could be influenced as a result. Therefore, there is an ever-growing need to develop automated tools supporting experts in analysing the news ecosystem and identifying large-scale manipulation attempts, and facilitating the study of how different events, global topics, and policies are being embraced by media in various countries, in order to carry out cross-country analysis and to gather knowledge on the ways how media informs public opinion, i.e., what aspects are being highlighted and linked to a topic, what pros and cons are mentioned, the way opinions are conveyed, and what rhetorical devices, i.e., logical fallacies and appeal to emotions, are used to support flawed argumentation, potentially leading to manipulation.

To foster research in this direction, there have been several shared tasks asking to detect the use of specific propaganda techniques in text, as well as the specific span of each instance. This includes the *NLP4IF-2019 shared task on Fine-Grained Propaganda Detection* (Da San Martino et al., 2019), *SemEval-2020 task 11 on Detection of Persuasion Techniques in News Articles* (Da San Martino et al., 2020a), *SemEval-2021 task 6 on Detection of Persuasion Techniques in Texts and Images* (Dimitrov et al., 2021b), and *WANLP-2022 Shared Task on Propaganda Detection in Arabic* (Alam et al., 2022).

Our task is an extension of the above ones and introduces several novelties. First, it is multilingual, covering nine languages. Second, it adds additional dimensions for better news understanding, i.e., framing and news genre detection. Finally, our taxonomy of persuasion techniques is an extension compared to previous inventories, and it contains 23 techniques organised in a two-tier hierarchy.

2 The Tasks

The shared task comprises three subtasks:

Subtask 1 (ST1): News Genre Categorization. Given a news article, determine whether: (a) it is an *opinion* piece, (b) aims at objective news *reporting*, or (c) is *satirical*.[1] This is a multi-class classification task at the article level.

Subtask 2 (ST2): Framing Detection. Given a news article, identify one or more frames used in

[1]A satirical piece is a factually incorrect article, with the intent not to deceive, but rather to call out, ridicule, or expose behaviours considered 'bad'. It deliberately exposes real-world individuals, organisations and events to ridicule.

the article from a pool of 14 generic framing dimensions (introduced in Card et al. (2015)): *Economic*, *Capacity and resources*, *Morality*, *Fairness and equality*, *Legality, constitutionality and jurisprudence*, *Policy prescription and evaluation*, *Crime and punishment*, *Security and defense*, *Health and safety*, *Quality of life*, *Cultural identity*, *Public opinion*, *Political*, *External regulation and reputation*. This is a multi-class multi-label classification task at the article level.

Subtask 3 (ST3): Persuasion Techniques Detection. Given a news article, identify the persuasion techniques used in each paragraph of the article. The pool of persuasion techniques consists of 23 techniques, and is an extension of the taxonomy introduced in Da San Martino et al. (2019); Dimitrov et al. (2021b)[2]. This is a multi-class multi-label classification task at the paragraph level. The persuasion techniques are grouped into six main categories:

Attack on reputation: The argument does not address the topic, but rather targets the participant (personality, experience, deeds) in order to question and/or to undermine their credibility. The object of the argumentation can also refer to a group of individuals, an organization, an object, or an activity.

Justification: The argument is made of two parts, a statement and an explanation or an appeal, where the latter is used to justify and/or to support the statement.

Simplification: The argument excessively simplifies a problem, usually regarding the cause, the consequence or the existence of choices.

Distraction: The argument takes the focus away from the main topic or argument to distract the reader.

Call: The text is not an argument, but an encouragement to act or to think in a particular way.

Manipulative wording: the text is not an argument per se, but uses specific language, which contains words or phrases that are either non-neutral, confusing, exaggerating, loaded, etc., in order to impact the reader emotionally.

Figure 1 gives an overview of the two-tier persuasion techniques taxonomy.

3 Related Work

This section discusses prior work related to each of the subtasks of the shared task.

[2]the second paper has five additional techniques with respect to the previous one

ATTACK ON REPUTATION

Name Calling or Labelling [AR:NCL]: a form of argument in which loaded labels are directed at an individual, group, object or activity, typically in an insulting or demeaning way, but also using labels the target audience finds desirable.
Guilt by Association [AR:GA]: attacking the opponent or an activity by associating it with another group, activity, or concept that has sharp negative connotations for the target audience.
Casting Doubt [AR:D]: questioning the character or the personal attributes of someone or something in order to question their general credibility or quality.
Appeal to Hypocrisy [AR:AH]: the target of the technique is attacked based on their reputation by charging them with hypocrisy/inconsistency.
Questioning the Reputation [AR:QR]: the target is attacked by making strong negative claims about it, focusing specially on undermining its character and moral stature rather than relying on an argument about the topic.

JUSTIFICATION

Flag Waiving [J:FW]: justifying an idea by exhaling the pride of a group or highlighting the benefits for that specific group.
Appeal to Authority [J:AA]: a weight is given to an argument, an idea or information by simply stating that a particular entity considered as an authority is the source of the information.
Appeal to Popularity [J:AP]: a weight is given to an argument or idea by justifying it on the basis that allegedly "everybody" (or the large majority) agrees with it or "nobody" disagrees with it.
Appeal to Values [J:AV]: a weight is given to an idea by linking it to values seen by the target audience as positive.
Appeal to Fear, Prejudice [J:AF]: promotes or rejects an idea through the repulsion or fear of the audience towards this idea.

DISTRACTION

Strawman [D:SM]: consists in making an impression of refuting an argument of the opponent's proposition, whereas the real subject of the argument was not addressed or refuted, but instead was replaced with a false one.
Red Herring [D:RH]: consists in diverting the attention of the audience from the main topic being discussed, by introducing another topic, which is irrelevant.
Whataboutism [D:W]: a technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

SIMPLIFICATION

Causal Oversimplification [S:CaO]: assuming a single cause or reason when there are actually multiple causes for an issue.
False Dilemma or No Choice [S:FDNC]: a logical fallacy that presents only two options or sides when there are many options or sides. In extreme, the author tells the audience exactly what actions to take, eliminating any other possible choices.
Consequential Oversimplification [S:CoO]: is an assertion one is making of some "first" event/action leading to a domino-like chain of events that have some significant negative (positive) effects and consequences that appear to be ludicrous or unwarranted or with each step in the chain more and more improbable.

CALL

Slogans [C:S]: a brief and striking phrase, often acting like an emotional appeal, that may include labeling and stereotyping.
Conversation Killer [A:CK]: words or phrases that discourage critical thought and meaningful discussion about a given topic.
Appeal to Time [C:AT]: the argument is centred around the idea that time has come for a particular action.

MANIPULATIVE WORDING

Loaded Language [MW:LL]: use of specific words and phrases with strong emotional implications (either positive or negative) to influence and convince the audience that an argument is valid.
Obfuscation, Intentional Vagueness, Confusion [MW:OVC]: use of words that are deliberately not clear, vague, or ambiguous so that the audience may have its own interpretations.
Exaggeration or Minimisation [MW:EM]: consists of either representing something in an excessive manner or making something seem less important or smaller than it really is.
Repetition [MW:R]: the speaker uses the same phrase repeatedly with the hope that the repetition will lead to persuade the audience.

Figure 1: Persuasion techniques taxonomy. The six coarse-grained techniques are subdivided into 23 fine-grained ones. An acronym for each technique is given in squared brackets.

3.1 News Genre Categorization

Rashkin et al. (2017) developed a corpus with document-level annotations into four classes (*trusted*, *satire*, *hoax*, and *propaganda*), annotated using distant supervision. Horne and Adali (2017) studied the relationship between fake news, real news, and satire with focus on style. They found that fake news is more similar to satire than to real news. Golbeck et al. (2018) developed a dataset of fake news and satire stories and analyzed and compared their thematic content. Satire was also one of the categories in the NELA-GT-2018 dataset (Nørregaard et al., 2019), as well as in its extended version NELA-GT-2019 (Gruppi et al., 2020).

The set up of our shared task is different, and focusing on distinguishing between objective news reporting vs. opinion piece vs. satire.

3.2 Framing Detection

Framing is a strategic device and a central concept in political communication for representing different salient aspects and perspectives for the purpose of conveying the latent meaning about an issue (Entman, 1993). It is important for news media as the same topics can be discussed from different perspectives. There has been work on automatically identifying media frames, including annotation schemes and datasets such as the Media Frames Corpus (Card et al., 2015), systems to automatically detect media frames (Liu et al., 2019; Zhang et al., 2019), large-scale automatic analysis of New York Times Articles (Kwak et al., 2020), and a semi-supervised approach to detecting frames in online news sources (Cheeks et al., 2020).

In our shared task, we adopt the frame inventory of the Media Frames Corpus.

3.3 Persuasion Techniques Detection

Work on persuasion detection overlaps to a large extent with work on propaganda detection, as there are many commonalities between the two.

Early work on propaganda detection focused on document-level analysis. Rashkin et al. (2017) predicted four classes (*trusted*, *satire*, *hoax*, and *propaganda*), labeled using distant supervision. Barrón-Cedeno et al. (2019) developed a corpus with two labels (i.e., *propaganda* vs. *non-propaganda*) and further investigated writing style and readability level. Their findings confirmed that using distant supervision, in conjunction with rich representations, might encourage the model to predict the source of the article, rather than to discriminate propaganda from non-propaganda.

An alternative line of research focused on detecting the use of specific propaganda techniques in text, e.g., Habernal et al. (2017, 2018) developed a corpus with 1.3k arguments annotated with five fallacies that directly relate to propaganda techniques. A more fine-grained analysis was done by Da San Martino et al. (2019), who developed a corpus of news articles annotated for 18 propaganda techniques, considering separately the task of technique spans detection and classification. They further tackled a sentence-level propaganda detection task, and proposed a multi-granular gated deep neural network. Subsequently, the Prta system was released (Da San Martino et al., 2020c), and improved models were proposed addressing the limitations of transformers (Chernyavskiy et al., 2021), or looking into interpretable propaganda detection (Yu et al., 2021). Finally, there is work addressing the detection of use of propaganda techniques in memes (Dimitrov et al., 2021a), the relationship between propaganda and coordination (Hristakieva et al., 2022), and work studying COVID-19 related propaganda in social media (Nakov et al., 2021a,b). See (Da San Martino et al., 2020b) for a survey on computational propaganda detection.

Several shared tasks on propaganda detecting in text were also organized. *SemEval-2020 task 11 on Detection of Persuasion Techniques in News Articles* (Da San Martino et al., 2020a) focused on news articles, and asked to detect the text spans where propaganda techniques are used, and to predict their type (14 techniques). Closely related to that is the *NLP4IF-2019 task on Fine-Grained Propaganda Detection* (Da San Martino et al., 2019), which asked to detect the spans of use in news articles of each of 18 propaganda techniques. The *SemEval-2021 task 6 on Detection of Persuasion Techniques in Texts and Images* focused on 22 propaganda techniques in memes (Dimitrov et al., 2021b), while WANLP'2022 shared task asked to detect the use of 20 propaganda techniques in Arabic tweets (Alam et al., 2022). Here, we extend and redesign the above annotation schemes.

4 The Dataset

This section provides a brief description of the dataset, whereas detailed guidelines, definitions and examples are provided in a separate technical report (Piskorski et al., 2023).

We collected articles in nine languages: English, French, German, Georgian, Greek, Italian, Polish, Russian, and Spanish published in the period between 2020 and mid-2022, and revolving around various globally discussed topics, including the COVID-19 pandemic, abortion-related legislation, migration, Russo-Ukrainian war, some local events such as parliamentary elections, etc. We considered both mainstream media and "alternative" media sources that could potentially spread mis-/disinformation. For the former, we used various news aggregation engines, e.g., Google News[3] and Europe Media Monitor[4], etc., which cover sources with different political orientation, whereas for the latter, we used online services such as MediaBias-FactCheck[5] and NewsGuard.[6] We extracted the article texts either using Trafilatura (Barbaresi, 2021) or, in few cases, ad hoc procedures.

We annotated each text for genre, framing, and persuasion techniques using the taxonomy described in Section 2. While genre and framing were annotated at the document level, we annotated the persuasion techniques at the span level. We had about 40 annotators, who were either media analysts, disinformation specialists or NLP experts, most of which had prior experience in performing linguistic annotations. All annotators were either native or near-native speakers of the language they annotated for. We used the INCEpTION (Klie et al., 2018) platform for carrying out the annotations. The annotation interface for an example document using INCEpTION is shown in Figure 2.

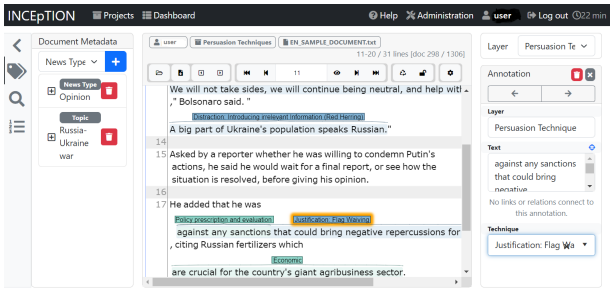

Figure 2: Example of a multi-label annotation using Inception: news genre is annotated as document metadata (left), while the persuasion techniques and the framings are highlighted in blue and green, respectively.

As regards English, we exploited the texts from Da San Martino et al. (2019), but the annotations for persuasion techniques have been slightly modified in order to match the extended taxonomy, most notably *Whataboutism* included two meanings: distracting from the main argument and calling the hypocrisy of the speaker; the latter meaning is now covered by the technique *Appeal to Hypocrisy*. Moreover, we added annotations for framing and news genre.

			train			
lang	#docs	#chars	#spans	#ne-par	#avg-fr	#avg-pt
English	446	2,431K	7201	9498	3.7	16.1
French	158	737K	5,595	2196	3.0	35.4
German	132	581K	4,501	1484	4.3	34.1
Italian	227	927K	6,027	2552	3.8	26.6
Polish	145	765K	2,839	2294	5.0	19.6
Russian	143	590K	3,399	1876	2.5	23.8
			development			
lang	#docs	#chars	#spans	#ne-par	#avg-fr	#avg-pt
English	90	403K	1,801	3,127	5.1	20.0
French	53	222K	1,586	610	3.0	29.9
German	45	171K	1,236	522	4.6	27.5
Italian	76	287K	1,934	882	3.9	25.4
Polish	49	264K	985	800	4.9	20.1
Russian	48	163K	739	515	2.3	15.4
			test			
lang	#docs	#chars	#spans	#ne-par	#avg-fr	#avg-pt
English	54	228K	1,775	910	4.7	32.9
French	50	181K	1,681	510	3.1	33.6
German	50	259K	1,904	790	5.7	38.1
Italian	61	245K	2,351	593	3.8	38.5
Polish	47	349K	1,491	1006	5.9	31.7
Russian	72	161K	944	601	1.2	13.1
Georgian	29	46K	218	161	1.7	7.5
Greek	64	248K	691	947	2.9	10.8
Spanish	30	109K	546	330	2.3	18.2

Table 1: Statistics about the *training*, the *development*, and the *test* datasets: total number of documents (#docs), total number of characters (#chars), total number of text spans annotated (#spans), total number of non-empty paragraphs (#ne-par), average number of frames per document (#avg-fr), and average number of persuasion techniques per document (#avg-pt).

	train			**dev**			**test**		
lang	op	rep	sat	op	rep	satire	op	rep	sat
English	382	41	10	20	54	9	32	17	5
French	103	43	11	35	15	4	37	7	6
German	86	27	19	29	9	7	33	12	5
Italian	174	44	8	59	15	3	41	13	7
Polish	104	25	15	35	9	6	35	10	2
Russian	93	41	8	32	14	3	45	18	9
Georgian	-	-	-	-	-	-	19	10	0
Greek	-	-	-	-	-	-	39	22	3
Spanish	-	-	-	-	-	-	14	9	7
all	942	221	71	210	116	32	295	118	44

Table 2: Number of documents from each genre across the languages: opinion (op), reporting (rep), satire (sat).

Each document was annotated by at least two annotators. Once the individual annotations for a document have been accomplished, a curator (an experienced annotator) with the help of annotators

[3] https://news.google.com
[4] https://emm.newsbrief.eu
[5] https://mediabiasfactcheck.com
[6] https://www.newsguardtech.com

consolidated the final annotation. The consolidation consisted of: (a) merging complementary annotations (tagged only by one annotator), (b) deciding whether overlapping annotations are to be kept as they are (multi-labels) or joined into a single-labeled annotation, and (c) carrying out global consistency analysis. The detailed description of the annotation and the consolidation process are described in a detailed technical report (Piskorski et al., 2023). In order to assess the annotation quality, we computed the Inter-Annotator Agreement (IAA) using Krippendorf's α: the value was 0.342, which is lower than the recommended threshold of 0.667, but we should note that this value represents the agreement before the consolidation, and as such, it is more representative of the consolidation difficulty rather than of the quality of the final consolidated annotations. Actually, we used IAA to allocate consolidation roles and to eliminate low-performing annotators.

We further studied the IAA by ranking the annotators by their performance with respect to the ground truth on the subset of documents they annotated. We then split them into two groups, *top* and *low*, based on the median micro-F_1 scores. Their respective α scores were 0.415 and 0.250. Finally, we considered the value of α of the group of annotators, based on Italian, the only language with two curators, achieving 0.588, which is lower but close to the recommended value.

The annotated data, consisting of 2,049 documents in total, were divided into *train*, *dev*, and *test* datasets, whose high-level statistics are provided in Table 1. Georgian, Greek, and Spanish-annotated data was used only for testing (surprise languages). Table 2 shows the distribution of articles per language in terms of genre. Detailed statistics about the fine-grained persuasion techniques are shown in Table 17 in Annex A.

5 Evaluation Framework

5.1 Evaluation Measures

Subtask 1 is a multi-class classification problem. We used $macro\ F_1$ as the official evaluation measure, but we also computed $micro\ F_1$.
Subtask 2 is a multi-label multi-class classification problem. We used $micro\ F_1$ as the official evaluation measure, but we also computed $macro\ F_1$.
Subtask 3 is a multi-label multi-class classification problem. We used $micro\ F_1$ as the official evaluation measure. The official score was computed using the 23 fine-grained persuasion technique labels. We also computed $macro\ F_1$.

5.2 Task Organization

The shared task was run in two phases:
Development Phase: initially, only *training* and *development* data were made available to the participants, and no gold labels were provided for the latter. The participants competed against each other to achieve the best performance on the development set. They could make an unlimited number of submissions, and the best score for each team, regardless of the submission time, was shown in real time on a public leaderboard.
Test Phase: in the second phase, the gold labels for the *development* and the *test* sets were released, and the participants were given a week to submit their final predictions on the *test* set. It is important to note that the test data contained news in three additional languages, i.e., Georgian, Greek, and Spanish, which were not known upfront to the participants (surprise languages). The participants could again submit multiple runs, but they would not get any feedback on their performance. Only the latest submission of each team was considered as official and was used for the final team ranking. Overall, 41 teams made official submissions to all the tasks, where 27, 22, and 22 teams submitted results for ST1, ST2, ST3, respectively. Moreover, 13, 14, and 14 teams submitted results for all languages for ST1, ST2, ST3, respectively.

The results for the development and the test phases are available on the leaderboard[7] page. After the competition was over, we left the submission system open for the test dataset for post-shared task evaluations and to monitor the state of the art for the different subtasks across the languages.

6 Participants and Results

6.1 Subtask 1: News Genre Categorization

The full results for Subtask 1 are shown in Tables 3 and 4 (surprise languages). We used a linear SVM[8] with class balancing, trained on 5-char n-grams as a baseline (highlighted in blue in the tables). Table 5 shows an overview of the approaches. Almost all participants used transformers. The scarcity of the annotated data was dealt with either by combining the datasets for all languages, e.g., via multilingual

[7] https://propaganda.math.unipd.it/semeval2023task3/leaderboard.php
[8] https://scikit-learn.org

English			Italian			Russian			French			German			Polish		
TEAM	mac	mic	TEAM	mac	mic	TEAM	mac	mic	TEAM	mac	mic	TEAM	mac	mic	TEAM	mac	mic
MELODI	.784	.815	Hitachi	.768	.852	Hitachi	.755	.750	UMUTeam	.835	.880	UMUTeam	.820	.820	FTD	.786	.936
MLModeler5	.616	.630	QUST	.767	.836	SheffieldVeraAI	.729	.722	QCRI	.767	.800	SheffieldVeraAI	.820	.820	Hitachi	.779	.872
SheffieldVeraAI	.613	.704	DSHacker	.720	.836	FTD	.668	.694	Hitachi	.744	.780	DSHacker	.813	.820	SheffieldVeraAI	.765	.851
HHU	.594	.611	SheffieldVeraAI	.720	.836	UMUTeam	.645	.681	DSHacker	.710	.720	SinaAI	.782	.760	MELODI	.709	.851
DSHacker	.591	.630	MELODI	.587	.754	MELODI	.586	.625	SheffieldVeraAI	.682	.740	MELODI	.779	.780	UMUTeam	.664	.809
Unisa	.586	.611	UnedMediaBias	.584	.623	QCRI	.567	.653	FTD	.671	.780	Hitachi	.777	.760	SinaAI	.663	.809
Hitachi	.553	.593	UMUTeam	.553	.754	DSHacker	.559	.597	MELODI	.656	.740	FTD	.713	.720	DSHacker	.661	.809
UnedMediaBias	.524	.574	QCRI	.541	.787	Spoke	.490	.653	SinaAI	.638	.680	QCRI	.667	.660	kb	.653	.809
SinaAI	.506	.667	FTD	.517	.754	QUST	.472	.514	QUST	.621	.700	SATLab	.644	.700	SATLab	.571	.830
QUST	.506	.630	SinaAI	.502	.738	SinaAI	.443	.472	Baseline	.568	.740	Baseline	.630	.760	QCRI	.571	.830
UMUTeam	.413	.593	HHU	.455	.639	HHU	.426	.472	UnedMediaBias	.465	.480	QUST	.626	.660	QUST	.528	.596
UM6P	.394	.519	Riga	.436	.574	Baseline	.398	.653	SATLab	.447	.640	HHU	.611	.740	UnedMediaBias	.507	.553
Riga	.349	.537	Baseline	.389	.672	UnedMediaBias	.365	.444	Riga	.356	.580	FramingFreaks	.569	.700	Baseline	.490	.830
FTD	.329	.463	FramingFreaks	.360	.656	Riga	.271	.389	JUSTR00	.347	.660	Riga	.412	.480	Riga	.433	.468
kb	.299	.574	SATLab	.319	.623	MaChAmp	.256	.625	FramingFreaks	.341	.660	UnedMediaBias	.362	.420	MaChAmp	.285	.745
Baseline	.288	.611	JUSTR00	.317	.574	FramingFreaks	.236	.319	MaChAmp	.284	.740	JUSTR00	.265	.660	FramingFreaks	.282	.702
QCRI	.281	.593	MaChAmp	.268	.672	E8IJS	.175	.306	E8IJS	.080	.120	MaChAmp	.265	.660	E8IJS	.063	.085
Spoke	.265	.444	E8IJS	.121	.164							E8IJS	.118	.180			
JUSTR00	.257	.370															
FramingFreaks	.248	.593															
MaChAmp	.248	.593															
ssnNlp	.248	.593															
SATLab	.243	.574															
UTB-NLP	.243	.574															
E8IJS	.075	.093															

Table 3: Results for Subtask 1 for the six main languages: $macro\ F_1$ (mac), $micro\ F_1$ (mic), ordered by the former, which is the official score.

language models or by automatic translation, or by looking for similar datasets in the literature; ensemble methods have also been very popular.

Spanish			Greek			Georgian		
TEAM	mac	mic	TEAM	mac	mic	TEAM	mac	mic
DSHacker	.563	.567	SinaAI	.806	.813	Riga	1,000	1,000
QUST	.552	.633	UMUTeam	.767	.797	SheffieldVeraAI	.963	.966
QCRI	.489	.567	HHU	.750	.750	FTD	.888	.897
SheffieldVeraAI	.443	.500	QCRI	.708	.813	QCRI	.622	.897
MELODI	.443	.600	FTD	.698	.766	DSHacker	.597	.862
UMUTeam	.438	.500	SheffieldVeraAI	.687	.734	UMUTeam	.582	.862
FTD	.400	.433	MELODI	.637	.703	QUST	.537	.690
Riga	.385	.500	DSHacker	.593	.641	SATLab	.519	.724
UnedMediaBias	.336	.367	UnedMediaBias	.521	.563	MELODI	.490	.724
HHU	.327	.433	QUST	.492	.609	UnedMediaBias	.486	.690
SinaAI	.323	.433	Riga	.412	.578	SinaAI	.468	.690
FramingFreaks	.317	.467	SATLab	.254	.406	MaChAmp	.396	.655
SATLab	.282	.433	MaChAmp	.252	.609	Baseline	.256	.345
E8IJS	.235	.300	FramingFreaks	.234	.344	FramingFreaks	.255	.621
MaChAmp	.212	.467	Baseline	.171	.344	E8IJS	.000	.000
Baseline	.154	.300	E8IJS	.057	.063			

Table 4: Results for Subtask 1 for the three surprise languages: $macro\ F_1$ (mac), $micro\ F_1$ (mic), ordered by the former, which is the official score.

Below, we give a short description of the system papers that were top-ranked for at least one language.

SinaAI (EL) used multilingual languages models, XLM, mBERT and LABSE, and ensembles thereof. They further used data augmentation by selecting 30% of the sentences of each document to create new synthetic examples.

DSHacker (ES) created synthetic texts for each class using the OpenAI GPT-3 Davinci language model. Each language was augmented by approximately 500 articles per genre, producing roughly 13,500 artificially generated articles. Then, they fine-tuned a single XLM-RoBERTalarge on the original and the augmented data.

Team Name	transformers	other representations	additional data	data augmentation	ensembles	preprocessing	trained on all languages	knowledge base	chunk processing	data translation
DSHacker	✓		✓	✓			✓			
FTD	✓		✓		✓	✓	✓			
HHU	✓			✓	✓	✓			✓	
Hitachi	✓		✓		✓		✓			
MELODI	✓					✓	✓			
MLModeler5	✓		✓	✓		✓				
MaChAmp	✓		✓				✓			
NLUBot101										
QCRI	✓			✓			✓			
QUST	✓				✓	✓	✓		✓	
Riga	✓				✓		✓			✓
SheffieldVeraAI	✓		✓	✓	✓	✓	✓			
UM6P	✓	✓			✓					
UMUTeam	✓					✓	✓		✓	
UTB-NLP		✓	✓		✓	✓				
UnedMediaBiasTeam	✓		✓				✓			
Unisa	✓			✓		✓				✓
kb	✓									

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.

FTD (PL): They experimented with monolingual and multilingual models, ensembles, additional data, and uncertainty estimation. For Russian and English, they fine-tuned models pretrained on the FTD dataset for genre classification. For English, they added 1,000 reporting texts from Gigaword. For Polish and German, their best results were achieved by fine-tuning a monolingual Polish BERT and a monolingual German Electra, respec-

English			Italian			Russian			French			German			Polish		
TEAM	mic	mac	TEAM	mic	mac	TEAM	mic	mac	TEAM	mic	mac	TEAM	mic	mac	TEAM	mic	mac
SheffieldVeraAI	.579	.539	MarsEclipse	.617	.545	MarsEclipse	.450	.303	MarsEclipse	.553	.537	MarsEclipse	.711	.660	MarsEclipse	.673	.638
TeamAmpa	.567	.510	QCRI	.599	.479	SheffieldVeraAI	.441	.356	BERTastic	.537	.520	QCRI	.660	.606	SheffieldVeraAI	.645	.603
MarsEclipse	.562	.490	Hitachi	.598	.515	QCRI	.434	.364	SheffieldVeraAI	.534	.520	SheffieldVeraAI	.653	.601	QCRI	.642	.599
Hitachi	.543	.472	TeamAmpa	.597	.483	TeamAmpa	.409	.294	Hitachi	.514	.488	TeamAmpa	.632	.573	UMUTeam	.642	.593
mCPT	.535	.482	mCPT	.584	.469	mCPT	.409	.367	TeamAmpa	.506	.479	Hitachi	.629	.567	Hitachi	.634	.584
QUST	.513	.462	UMUTeam	.576	.447	BERTastic	.393	.265	TheSyllogist	.486	.462	mCPT	.622	.564	SATLab	.620	.570
QCRI	.513	.419	SheffieldVeraAI	.571	.491	TheSyllogist	.385	.290	QCRI	.480	.430	QUST	.616	.545	TeamAmpa	.614	.555
BERTastic	.512	.446	TheSyllogist	.554	.444	UMUTeam	.385	.288	UMUTeam	.477	.438	UMUTeam	.614	.565	MaChAmp	.597	.582
UMUTeam	.508	.415	BERTastic	.545	.469	Hitachi	.370	.326	mCPT	.469	.429	BERTastic	.603	.562	mCPT	.597	.555
ACCEPT	.507	.502	QUST	.502	.465	Riga	.315	.222	ACCEPT	.456	.443	MaChAmp	.582	.564	Baseline	.594	.532
MaChAmp	.506	.493	Riga	.499	.321	ACCEPT	.254	.249	QUST	.447	.438	SATLab	.572	.519	QUST	.591	.533
TheSyllogist	.487	.409	ACCEPT	.495	.439	QUST	.250	.213	Riga	.376	.287	FTD	.555	.299	FTD	.588	.516
MLModeler5	.477	.427	Baseline	.486	.372	Baseline	.230	.218	SATLab	.375	.352	FramingFreaks	.545	.496	BERTastic	.587	.535
FTD	.453	.362	SATLab	.474	.416	FramingFreaks	.219	.159	MaChAmp	.359	.355	TheSyllogist	.537	.465	FramingFreaks	.560	.460
JUSTR00	.443	.363	FTD	.459	.227	FTD	.198	.117	Baseline	.329	.276	Riga	.509	.375	TheSyllogist	.553	.501
Riga	.420	.313	FramingFreaks	.452	.355	MaChAmp	.161	.151	FramingFreaks	.327	.300	ACCEPT	.496	.460	Riga	.542	.412
SATLab	.378	.317	MaChAmp	.424	.403	SinaAI	.113	.128	FTD	.255	.105	Baseline	.487	.418	ACCEPT	.510	.490
Baseline	.350	.274	SinaAI	.251	.200	DigDemLab	.070	.055	DigDemLab	.220	.192	DigDemLab	.335	.279	SinaAI	.475	.446
UTB-NLP	.341	.309	DigDemLab	.237	.173				SinaAI	.187	.157	SinaAI	.302	.265	DigDemLab	.392	.348
IA2022Grupa1	.326	.265															
SinaAI	.266	.226															
DigDemLab	.207	.172															
FramingFreaks	.196	.142															

Table 6: Results for Subtask 2 for the six main languages: $micro\ F_1$ (mic), $macro\ F_1$ (mac), ordered by the former, which is the official score.

Spanish			Greek			Georgian		
TEAM	mic	mac	TEAM	mic	mac	TEAM	mic	mac
mCPT	.571	.455	SheffieldVeraAI	.546	.454	SheffieldVeraAI	.654	.679
UMUTeam	.558	.465	TeamAmpa	.544	.444	MarsEclipse	.645	.639
SheffieldVeraAI	.508	.432	UMUTeam	.534	.404	TheSyllogist	.561	.493
TeamAmpa	.506	.387	TheSyllogist	.530	.440	BERTastic	.552	.408
Riga	.489	.426	BERTastic	.526	.444	UMUTeam	.529	.411
QCRI	.488	.390	QCRI	.519	.400	QCRI	.517	.457
MarsEclipse	.477	.404	mCPT	.516	.410	TeamAmpa	.517	.379
BERTastic	.477	.428	MarsEclipse	.498	.402	Riga	.424	.381
TheSyllogist	.473	.387	QUST	.414	.392	mCPT	.400	.291
ACCEPT	.388	.387	FramingFreaks	.380	.154	FramingFreaks	.352	.344
MaChAmp	.385	.269	Riga	.377	.195	MaChAmp	.313	.225
SATLab	.383	.293	ACCEPT	.355	.370	QUST	.311	.260
QUST	.374	.353	Baseline	.345	.057	Baseline	.260	.251
FTD	.265	.201	MaChAmp	.293	.206	ACCEPT	.220	.290
FramingFreaks	.215	.211	SinaAI	.140	.123	SinaAI	.133	.205
SinaAI	.181	.163	SATLab	.068	.037	SATLab	.053	.184
Baseline	.120	.095						

Table 7: Results for Subtask 2 for the three surprise languages: $micro\ F_1$ (mic), $macro\ F_1$ (mac), ordered by the former, which is the official score.

tively. For the other languages, their best systems used multilingual BERT, XLM-RoBERTa, or ensembles thereof. In all cases, they truncated the input to the first 510 tokens. They further upsampled the data to balance the distribution between the classes (the results without upsampling were low).

Hitachi (IT, RU) augmented the dataset for subtask 1 by collecting labelled examples from similar datasets. They pretrained (XLM-)RoBERTa in multi-task (one language, subtasks 1 and 2), multilingual (one subtask, all languages), and multilingual multi-task (subtasks 1 and 2 in all languages) settings. Besides using the single models, they report experiments with ensembles of base models and different hyper-parameter values.

MELODI (EN) fine-tuned the domain-specific language model trained on English data, POLITICS, on the English input articles and on the articles in all other languages, which were automatically translated. In addition, in order to use whole articles as input, they used a sliding window and aggregated each window representation using mean-pooling. They also tested other multilingual approaches, such as XLM-RoBERTa, and were able to process long documents (Longformer), which were generally less effective.

UMUTeam (FR, DE) used a multilingual model based on XML-RoBERTa, which was fine-tuned on all languages at once and a sentence transformer model to extract the most important chunk of text. The input data was truncated to 200 tokens with 50 overlaps using the sentence-transformer model to obtain the subset of text most related to the article's title.

SheffieldVeraAI (DE) deployed an ensemble of three fine-tuned mBERT models and one mBERT model with a bottleneck adapter. All used bert-base-multilingual-cased. The pool of training data was also extended by integration additional "satire" resources for English. The final predictions were drawn as a majority-voting predicted class.

6.2 Subtask 2: Framing

The full results for subtask 2 on framing classification are provided in Table 6 and 7 (surprise languages). We used linear SVM trained using word unigrams and bigrams as a baseline (highlighted in blue in the tables). Table 8 shows an overview of the approaches. Since the models were all

transformer-based, what differentiated the participating systems were once again the pre-processing and the data augmentation techniques. The vast majority of teams trained their systems on all languages and used ensembles.

Team Name	transformers	other representations	additional data	data augmentation	ensembles	preprocessing	trained on all languages	knowledge base	chunk processing	data translation
ACCEPT	✓				✓	✓	✓	✓	✓	✓
BERTastic	✓			✓		✓	✓			✓
FTD	✓									
Hitachi	✓				✓		✓			
MLModeler5	✓		✓	✓		✓				
MaChAmp	✓									
MarsEclipse	✓					✓	✓			
QCRI	✓			✓			✓			
QUST	✓				✓	✓	✓		✓	
Riga	✓				✓		✓			✓
SheffieldVeraAI	✓		✓		✓	✓	✓			✓
TheSyllogist	✓						✓			✓
UMUTeam	✓					✓	✓			
UTB-NLP		✓	✓		✓	✓				
mCPT	✓						✓			

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.

MarsEclipse (IT, RU, FR, DE, PL): This team used a multi-label contrastive loss for fine-tuning XLM-RoBERTa using SimCLR and SimCSE and adapting the loss function to a multilabel setup.

mCPT (ES): This team used a two-phase training procedure of a transformer model, first by pretraining jointly on all the languages and then by fine-tuning for each language. In both phases, a multi-label contrastive loss was used.

SheffieldVeraAI (EN, EL, KA): The team achieved the best average rank score over all the languages. They used two different ensembles of MUPPET large and of XLM-RoBERTa large with adapters and task-adaptive MLM pretraining on the train+dev+test data. Their data was preprocessed and truncated. The models were trained both with and without class weighting.

6.3 Subtask 3: Persuasion Techniques Detection

The full results for subtask 3 on persuasion techniques detection are given in Tables 9 and 10 (surprise languages). We used linear SVM trained using word uni-grams and bigrams as a baseline (highlighted in blue in the tables). Table 11 shows an overview of the approaches used by the participating systems. The big picture is very similar to the previous subtasks: multilingual transformer models were used by all participants, and what differentiated the approaches was again the preprocessing and data augmentation strategies, for example, a few teams made use of the span-level annotations.

APatt (EN): The team combined different fine-tuned transformer models (XLNet, RoBERTa, BERT, ALBERT, and DeBERTa) through a weighted average. For English, they weighted the predictions of the models to give higher importance to certain models.

KInITVeraAI (IT, RU, DE, PL, EL, KA): This team performed overall the best, using a fine-tuned XLM-RoBERTa-large transformer model trained on all the input data. They carefully adjusted the prediction threshold for each language using a principled approach. They truncated the input, and also found that preprocessing did not impact the quality much.

NAP (FR): The team presented an approach combining predictions of several models in an ensemble, which differ in three main aspects: a) training data, b) model architecture, and c) input format to the model. They leveraged translation as data augmentation strategies using available MarianMT models. Model architectures included XLM-RoBERTa models, Adapters, SetFit, and linguistically-informed heuristics for underrepresented techniques which were fine-tuned on different combinations of original and augmented data. They fine-tuned models on both paragraph- and span-level information.

TeamAmpa (ES): The team used different oversampling strategies, data truncation, and monolingual and multilingually trained models, combined in an ensemble for the English Task 3 data. The surprise languages were handled using the multilingual model only, which where trained using XLM-R on all languages with oversampling, for one of these languages the team ranked first.

6.4 Aggregated results

Tables 12-14 report the average micro F_1 scores of the teams who, for each subtask, submitted solutions for multiple languages: the 6 for which we provided training data (6L), the 3 surprise ones (3L), all of them (9L). Results are ranked by decreasing value on all.

English			Italian			Russian			French			German			Polish		
TEAM	mic	mac	TEAM	mic	mac	TEAM	mic	mac	TEAM	mic	mac	TEAM	mic	mac	TEAM	mic	mac
APatt	.376	.129	KInITVeraAI	.550	.214	KInITVeraAI	.387	.189	NAP	.469	.322	KInITVeraAI	.513	.233	KInITVeraAI	.430	.179
SheffieldVeraAI	.368	.172	NAP	.539	.266	TeamAmpa	.378	.227	TeamAmpa	.434	.305	NAP	.510	.272	NAP	.422	.246
AppealForAtt	.363	.166	SheffieldVeraAI	.525	.282	QCRI	.361	.182	KInITVeraAI	.432	.214	QCRI	.498	.231	DSHacker	.390	.170
KInITVeraAI	.362	.133	TeamAmpa	.521	.264	NLUBot101	.323	.201	SheffieldVeraAI	.414	.324	APatt	.484	.177	TeamAmpa	.389	.236
NLUBot101	.361	.197	FTD	.516	.176	SheffieldVeraAI	.318	.205	QCRI	.401	.226	TeamAmpa	.476	.266	QCRI	.378	.156
FTD	.346	.088	QCRI	.513	.209	AppealForAtt	.312	.173	NLUBot101	.396	.254	SheffieldVeraAI	.447	.237	APatt	.366	.150
TeamAmpa	.325	.158	DSHacker	.496	.153	APatt	.306	.117	DSHacker	.388	.201	NLUBot101	.420	.179	SheffieldVeraAI	.347	.191
QCRI	.320	.133	ReDASPersuasion	.448	.106	NAP	.305	.193	APatt	.384	.191	AppealForAtt	.418	.218	AppealForAtt	.344	.201
DSHacker	.320	.140	APatt	.441	.166	MaChAmp	.271	.148	AppealForAtt	.374	.203	DSHacker	.408	.154	FTD	.327	.122
CLaC	.309	.071	Riga	.436	.092	DSHacker	.257	.083	kb	.362	.266	MaChAmp	.405	.178	NLUBot101	.320	.169
NL4IA	.308	.142	NLUBot101	.435	.164	kb	.253	.117	MaChAmp	.345	.207	ReDASPersuasion	.384	.078	kb	.314	.179
Unisa	.298	.109	SATLab	.433	.183	Riga	.252	.064	SATLab	.338	.241	kb	.373	.201	MaChAmp	.307	.170
MaChAmp	.295	.149	AppealForAtt	.431	.211	FTD	.235	.058	Riga	.306	.078	Riga	.373	.060	SATLab	.300	.143
Riga	.280	.062	MaChAmp	.422	.166	ReDASPersuasion	.219	.050	ReDASPersuasion	.301	.115	FTD	.363	.110	ReDASPersuasion	.238	.112
NAP	.263	.082	kb	.399	.201	Baseline	.207	.086	FTD	.298	.126	SATLab	.355	.163	UnedMediaBias	.237	.103
SATLab	.259	.103	Baseline	.397	.122	CLaC	.193	.057	Baseline	.240	.099	UnedMediaBias	.318	.106	Riga	.228	.038
ReDASPersuasion	.251	.045	UnedMediaBias	.317	.111	UnedMediaBias	.183	.100	CLaC	.239	.066	Baseline	.317	.083	CLaC	.190	.050
UnedMediaBias	.241	.078	CLaC	.313	.063	SinaAI	.139	.057	UnedMediaBias	.236	.121	CLaC	.248	.055	Baseline	.179	.059
Baseline	.195	.069	QUST	.213	.155	QUST	.100	.080	QUST	.209	.164	QUST	.153	.112	QUST	.097	.074
IA2022Grupa1	.193	.072	SinaAI	.203	.064				SinaAI	.195	.063	SinaAI	.042	.034	SinaAI	.064	.025
SinaAI	.141	.022															
QUST	.135	.103															
kb	.060	.031															

Table 9: Results for subtask 3 for the six main languages: $micro\ F_1$ (mic), $macro\ F_1$ (mac), ordered by the former, which is the official score.

Spanish			Greek			Georgian		
TEAM	mic	mac	TEAM	mic	mac	TEAM	mic	mac
TeamAmpa	.381	.244	KInITVeraAI	.267	.126	KInITVeraAI	.457	.328
KInITVeraAI	.380	.155	QCRI	.265	.129	QCRI	.414	.339
NAP	.370	.181	NAP	.258	.164	NAP	.413	.306
QCRI	.350	.157	TeamAmpa	.238	.171	TeamAmpa	.408	.259
AppealForAtt	.317	.139	MaChAmp	.215	.129	Riga	.362	.209
NLUBot101	.305	.151	AppealForAtt	.206	.119	MaChAmp	.301	.221
FTD	.281	.074	SheffieldVeraAI	.174	.110	AppealForAtt	.280	.261
MaChAmp	.276	.139	Riga	.164	.036	CLaC	.271	.199
SheffieldVeraAI	.275	.130	CLaC	.156	.055	NLUBot101	.254	.172
CLaC	.267	.048	NLUBot101	.150	.097	SheffieldVeraAI	.249	.296
Baseline	.248	.020	kb	.150	.121	UnedMediaBias	.180	.221
kb	.245	.143	SinaAI	.114	.029	kb	.150	.100
UnedMediaBias	.227	.078	UnedMediaBias	.106	.026	SinaAI	.139	.040
Riga	.199	.045	Baseline	.088	.006	Baseline	.138	.141
SATLab	.193	.057	QUST	.057	.047	QUST	.091	.115
SinaAI	.178	.028	SATLab	.000	.000	SATLab	.076	.158
QUST	.126	.099						

Table 10: Results for Subtask 3 for the three surprise languages: $micro\ F_1$ (mic), $macro\ F_1$ (mac), ordered by the former, which is the official score.

Team Name	transformers	other representations	additional data	data augmentation	ensembles	preprocessing	trained on all languages	knowledge base	chunk processing	data translation
APatt	✓				✓					
AppealForAtt	✓			✓		✓	✓		✓	✓
DSHacker	✓	✓	✓	✓	✓			✓		
FTD	✓									
KInITVeraAI	✓				✓	✓	✓		✓	✓
NAP	✓			✓	✓	✓	✓		✓	
NL4IA	✓								✓	
NLUBot101	✓			✓		✓	✓		✓	✓
QCRI	✓			✓			✓			
QUST	✓				✓	✓	✓		✓	
ReDASPersuasion	✓									✓
Riga	✓				✓		✓			
SheffieldVeraAI	✓			✓			✓			
UnedMediaBiasTeam	✓		✓				✓			
Unisa	✓					✓			✓	
kb	✓					✓				

Table 11: 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.

7 Conclusions and Future Work

We presented SemEval-2023 Task 3 on *Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup*. The task attracted a lot of attention: 181 teams registered for the task, 41 teams eventually madevan official submission on the test set, and 32 teams also submitted a task description paper.

In future work, we plan to further increase the data size, cover additional languages, and explore different ways of evaluation of the persuasion technique detection, e.g., by changing the focus (sentence- and text span-level evaluation).

8 Limitations

Dataset Representativeness Our dataset covers a range of topics of public interest (COVID-19, climate change, abortion, migration, the Russo-Ukrainian war, and local elections) as well as media from all sides of the political spectrum. However, it should not be seen as representative of the media in any country, nor should it be seen as perfectly balanced in any specific way.

Biases Human data annotation involves some degree of subjectivity. To mitigate this, we created a comprehensive 60-page guidelines document, which we updated from time to time to clarify newly arising important cases during the annotation process. We further had quality control steps

team	6L	3L	9L
SheffieldVeraAI	0.779	0.733	0.764
Hitachi	0.768	-	-
MELODI	0.761	0.676	0.732
UMUTeam	0.756	0.720	0.744
DSHacker	0.735	0.690	0.720
FTD	0.725	0.699	0.716
QCRI	0.720	0.759	0.733
Baseline	0.711	0.330	0.584
SinaAI	0.688	0.645	0.673
MaChAmp	0.672	0.577	0.641
QUST	0.656	0.644	0.652
FramingFreaks	0.605	0.477	0.562
UnedMediaBiasTeam	0.516	0.540	0.524
Riga	0.505	0.693	0.567
E8IJS	0.158	0.121	0.146

Table 12: Average macro score across language for the teams participating in all 'provided' six languages (6L), the three surprise languages (3L), all nine languages (9L) for subtask 1.

team	6L	3L	9L
MarsEclipse	0.594	0.540	0.576
SheffieldVeraAI	0.571	0.570	0.570
QCRI	0.555	0.508	0.539
TeamAmpa	0.554	0.522	0.543
Hitachi	0.548	-	-
mCPT	0.536	0.496	0.523
UMUTeam	0.534	0.541	0.536
BERTastic	0.530	0.518	0.526
TheSyllogist	0.500	0.521	0.507
QUST	0.486	0.366	0.446
ACCEPT	0.453	0.321	0.409
Riga	0.444	0.430	0.439
MaChAmp	0.438	0.330	0.402
FTD	0.418	-	-
Baseline	0.412	0.242	0.356
FramingFreaks	0.383	0.316	0.361
SinaAI	0.266	0.151	0.227
DigDemLab	0.244	-	-

Table 13: Average micro score across language for the teams participating in all 'provided' six languages (6L), the three surprise languages (3L), all nine languages (9L) for subtask 2.

in the data annotation process, and we have been excluding low-performing annotators. Despite all this, we are aware that some degree of intrinsic subjectivity will inevitably be present in the dataset and will eventually be learned by models trained on it.

Acknowledgments

We are greatly indebted to all the annotators from different organisations, including, i.a., the European Commission, the European Parliament, the University of Padua, and the Qatar Computing Research Institute, HBKU, who took part in the annotations, and notably to the language curators whose patience and diligence have been fundamental for the quality of the dataset. We are also thankful to Nikolaidis Nikolaos for the preparation of the baseline models and for sharing various ideas.

Part of this work was supported by IDKT Fund TDF 03-1209-210013: *Tanbih: Get to Know What You Are Reading*.

team	6L	3L	9L
KInITVeraAI	0.446	0.368	0.420
TeamAmpa	0.420	0.343	0.395
NAP	0.418	0.347	0.394
QCRI	0.412	0.343	0.389
SheffieldVeraAI	0.403	0.233	0.346
APatt	0.393	-	-
DSHacker	0.376	-	-
NLUBot101	0.376	0.236	0.329
APatt	0.374	0.268	0.338
FTD	0.347	-	-
MaChAmp	0.341	0.264	0.315
Riga	0.312	0.242	0.289
ReDASPersuasion	0.307	-	-
kb	0.294	0.182	0.256
Baseline	0.256	0.158	0.223
UnedMediaBias	0.255	0.171	0.227
CLaC	0.249	0.231	0.243
QUST	0.151	0.091	0.131
SinaAI	0.131	0.144	0.135

Table 14: Average micro score across language for the teams participating in all 'provided' six languages (6L), the three surprise languages (3L), all nine languages (9L) for subtask 3.

References

Ahmed Al-Qarqaz and Malak Abdullah. 2023. Team justr00 at semeval-2023 task 3: Transformers for news articles classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1213–1216, Toronto, Canada. Association for Computational Linguistics.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, Abu Dhabi, UAE.

Hamza Alami, Abdessamad Benlahbib, Abdelkader El Mahdaouy, and Ismail Berrada. 2023. Um6p at semeval-2023 task 3: News genre classification based on transformers, graph convolution networks and number of sentences. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 856–861, Toronto, Canada. Association for Computational Linguistics.

Sergiu Amihaesei, Laura Cornei, and George Stoica. 2023. Appeal for attention at semeval-2023 task 3:

Data augmentation extension strategies for detection of online news persuasion techniques. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 616–623, Toronto, Canada. Association for Computational Linguistics.

Micaela Bangerter, Giuseppe Fenza, Mariacristina Gallo, Vincenzo Loia, Alberto Volpe, Carmen De Maio, and Claudio Stanzione. 2023. Unisa at semeval-2023 task 3: A shap-based method for propaganda detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 885–891, Toronto, Canada. Association for Computational Linguistics.

Katarzyna Baraniak and M Sydow. 2023. Kb at semeval-2023 task 3: On multitask hierarchical bert base neural network for multi-label persuasion techniques detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1395–1400, Toronto, Canada. Association for Computational Linguistics.

Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Rosina Baumann and Sabrina Deisenhofer. 2023. Framingfreaks at semeval-2023 task 3: Detecting the category and the framing of texts as subword units with traditional machine learning. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 922–926, Toronto, Canada. Association for Computational Linguistics.

Fabian Billert and Stefan Conrad. 2023. Hhu at semeval-2023 task 3: An adapter-based approach for news genre classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1166–1171, Toronto, Canada. Association for Computational Linguistics.

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*, ACL-IJCNLP '15, pages 438–444, Beijing, China.

Loretta H Cheeks, Tracy L Stepien, Dara M Wald, and Ashraf Gaffar. 2020. Discovering news frames: An approach for exploring text, content, and concepts in online news sources. In *Cognitive Analytics: Concepts, Methodologies, Tools, and Applications*, pages 702–721. IGI Global.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2021. Transformers: "The end of history" for NLP? In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, ECML-PKDD'21.

Nelson Filipe Costa, Bryce Hamilton, and Leila Kosseim. 2023. Clac at semeval-2023 task 3: Language potluck roberta detects online persuasion techniques in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1613–1618, Toronto, Canada. Association for Computational Linguistics.

Juan Cuadrado, Elizabeth Martinez, Anderson Morillo, Daniel Peña, Kevin Sossa, Juan Carlos Martinez-Santos, and Edwin Puertas. 2023. Utb-nlp at semeval-2023 task 3: Weirdness, lexical features for detecting categorical framings, and persuasion in online news. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1551–1557, Toronto, Canada. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF '19, pages 162–170, Hong Kong, China.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, IJCAI-PRICAI '20, pages 4826–4832.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeno, and Preslav Nakov. 2020c. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 287–293. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP '19, pages 5636–5646, Hong Kong, China.

Nicolas Devatine, Philippe Muller, and Chloé Braud. 2023. Melodi at semeval-2023 task 3: In-domain pre-training for low-resource classification of news articles. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 108–113, Toronto, Canada. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021a. Detecting propaganda techniques in memes. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '21, pages 6603–6617.

Dimiter Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. Task 6 at SemEval-2021: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, pages 390–397.

Neele Falk, Annerose Eichel, and Prisca Piccirilli. 2023. Nap at semeval-2023 task 3: Is less really more? (back-)translation as data augmentation strategies for detecting persuasion techniques. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1433–1446, Toronto, Canada. Association for Computational Linguistics.

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 17–21, Amsterdam, Netherlands. Association for Computing Machinery.

Maurício Gruppi, Benjamin D. Horne, and Sibel Adali. 2020. NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv*, 2003.08444.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP '17, pages 7–12, Copenhagen, Denmark.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018. Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *LREC*. European Language Resources Association (ELRA).

Maram Hasanain, Ahmed Oumar El-Shangiti, Rabindra Nath Nandi, Preslav Nakov, and Firoj Alam. 2023. Qcri at semeval-2023 task 3: News genre, framing and persuasion techniques detection using multilingual models. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1237–1244, Toronto, Canada. Association for Computational Linguistics.

Philipp Heinisch, Moritz Plenz, Anette Frank, and Philipp Cimiano. 2023. Accept at semeval-2023 task 3: An ensemble-based approach to multilingual framing detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1347–1358, Toronto, Canada. Association for Computational Linguistics.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv*, 1703.09398.

Kristina Hristakieva, Stefano Cresci, Giovanni Da San Martino, Mauro Conti, and Preslav Nakov. 2022. The spread of propaganda by coordinated communities on social media. In *Proceedings of the 14th ACM Web Science Conference*, WebSci '22, pages 191–201, Barcelona, Spain.

Timo Hromadka, Timotej Smolen, Tomas Remis, Branislav Pecher, and Ivan Srba. 2023. Kinitveraai at semeval-2023 task 3: Simple yet powerful multilingual fine-tuning for persuasion techniques detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 629–637, Toronto, Canada. Association for Computational Linguistics.

Ye Jiang. 2023. Team qust at semeval-2023 task 3: A comprehensive study of monolingual and multilingual approaches for detecting online news genre, framing and persuasion techniques. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 300–306, Toronto, Canada. Association for Computational Linguistics.

Arjun Khanchandani, Nitansh Jain, and Jatin Bedi. 2023. Mlmodeler5 at semeval-2023 task 3: Detecting the category and the framing techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1096–1101, Toronto, Canada. 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).

Yuta Koreeda, Ken-ichi Yokote, Hiroaki Ozaki, Atsuki Yamaguchi, Masaya Tsunokake, and Yasuhiro Sogawa. 2023. Hitachi at semeval-2023 task 3: Exploring cross-lingual multi-task strategies for genre and framing detection in online news. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1702–1711, Toronto, Canada. Association for Computational Linguistics.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A systematic media frame analysis of 1.5 million New York Times articles from 2000 to 2017. In *Proceedings of the 12th ACM Conference on Web Science*, WebSci '20, pages 305–314, Southampton, United Kingdom.

Mikhail Lepekhin and Serge Sharoff. 2023. Ftd at semeval-2023 task 3: News genre and propaganda detection by comparing mono- and multilingual models with fine-tuning on additional data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 549–555, Toronto, Canada. Association for Computational Linguistics.

Qisheng Liao, Meiting Lai, and Preslav Nakov. 2023. Marseclipse at semeval-2023 task 3: Multi-lingual and multi-label framing detection with contrastive learning. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 83–87, Toronto, Canada. Association for Computational Linguistics.

Genglin Liu, Yi Fung, and Heng Ji. 2023. Nlubot101 at semeval-2023 task 3: An augmented multilingual nli approach towards online news persuasion techniques detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1636–1643, Toronto, Canada. 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 US gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, CoNLL '19, pages 504–514, Hong Kong, China.

Tarek Mahmoud and Preslav Nakov. 2023. Bertastic at semeval-2023 task 3: Fine-tuning pretrained multilingual transformers – does order matter? In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 58–63, Toronto, Canada. Association for Computational Linguistics.

Arkadiusz Modzelewski, Witold Sosnowski, Magdalena Wilczynska, and Adam Wierzbicki. 2023. Dshacker at semeval-2023 task 3: Genres and persuasion techniques detection with multilingual data augmentation through machine translation and text generation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1582–1591, Toronto, Canada. Association for Computational Linguistics.

Osama Mohammed Afzal and Preslav Nakov. 2023. Team thesyllogist at semeval-2023 task 3: Language-agnostic framing detection in multi-lingual online news: A zero-shot transfer approach. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 2058–2061, Toronto, Canada. Association for Computational Linguistics.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021a. COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '21.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021b. A second pandemic? Analysis of fake news about COVID-19 vaccines in Qatar. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '21.

Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adali. 2019. NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the Thirteenth International Conference on Web and Social Media*, ICWSM '19, pages 630–638, Munich, Germany. AAAI Press.

Ronghao Pan, José Antonio García-Díaz, Miguel Ángel Rodríguez-García, and Rafael Valencia-García. 2023. Umuteam at semeval-2023 task 3: Multilingual transformer-based model for detecting the genre, the framing, and the persuasion techniques in online news. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 609–615, Toronto, Canada. Association for Computational Linguistics.

Amalie Pauli, Rafael Pablos Sarabia, Leon Derczynski, and Ira Assent. 2023. Teamampa at semeval-2023 task 3: Exploring multilabel and multilingual roberta models for persuasion and framing detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 847–855, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, Jason Gonin, Camelia Ignat, Bonka Kotseva, Eleonora Mantica, Lorena Marcaletti, Enrico Rossi, Alessio Spadaro, Marco Verile, Giovanni Da San Martino, Firoj Alam, and Preslav Nakov. 2023. News categorization, framing and persuasion techniques: Annotation guidelines. Technical Report JRC-132862, European Commission Joint Research Centre, Ispra (Italy).

Albert Pritzkau. 2023. Nl4ia at semeval-2023 task 3: A comparison of sequence classification and token

classification to detect persuasive techniques. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 794–799, Toronto, Canada. Association for Computational Linguistics.

Antonio Purificato and Roberto Navigli. 2023. Apatt at semeval-2023 task 3: The sapienza nlp system for ensemble-based multilingual propaganda detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 382–388, Toronto, Canada. Association for Computational Linguistics.

Fatima Zahra Qachfar and Rakesh Verma. 2023. Redaspersuasion at semeval-2023 task 3: Persuasion detection using multilingual transformers and language agnostic features. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 2124–2132, Toronto, Canada. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 2931–2937, Copenhagen, Denmark.

Markus Reiter-Haas, Alexander Ertl, Kevin Innerhofer, and Elisabeth Lex. 2023. mcpt at semeval-2023 task 3: Multilingual label-aware contrastive pre-training of transformers for few- and zero-shot framing detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 941–949, Toronto, Canada. Association for Computational Linguistics.

Francisco-Javier Rodrigo-Ginés, Laura Plaza, and Jorge Carrillo-de Albornoz. 2023. Unedmediabiasteam @ semeval-2023 task 3: Can we detect persuasive techniques transferring knowledge from media bias detection? In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 787–793, Toronto, Canada. Association for Computational Linguistics.

Aryan Sadeghi, Reza Alipour, Kamyar Taeb, Parimehr Morassafar, Nima Salemahim, and Ehsaneddin Asgari. 2023. Sinaai at semeval-2023 task 3: A multilingual transformer language model-based approach for the detection of news genre, framing and persuasion techniques. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 2168–2173, Toronto, Canada. Association for Computational Linguistics.

Ben Wu, Olesya Razuvayevskaya, Freddy Heppell, João Leite, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Sheffieldveraai at semeval-2023 task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1995–2008, Toronto, Canada. Association for Computational Linguistics.

Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. 2021. Interpretable propaganda detection in news articles. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '21, pages 1597–1605.

Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James Glass, and Preslav Nakov. 2019. Tanbih: Get to know what you are reading. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: System Demonstrations*, EMNLP-IJCNLP '19, pages 223–228, Hong Kong, China.

A Supplementary Corpus Information

A.1 Statistics

This section contains additional statistical information related to the corpus.

Table 15 provides the statistics for genre for all languages. One can observe that *opinion* and *satire* are the most and least populated classes across the languages respectively.

Analogously, Table 16 shows the number of different framing dimensions per language. *Political* and *Security and Defense* framings constitute the two most frequent ones across all languages. The distribution of the different framings varies across the languages though.

Finally, Table 17 reports the exact count of fine-grained persuasion techniques per language for the entire dataset. The two most frequent techniques irrespective of the language are *Loaded Language* and *Name Calling-Labelling*, which account for 18.5% and 23.7% of the dataset, trumping by several order of magnitude the lower populated classes. They are followed by *Casting Doubt* (12.5%), *Questioning the Reputation* (7.6%), *Appeal to Fear-Prejudice* (4.8%), and *Exageration-Minimisation* (4.7%). These six classes together represent 71.8% of the entire dataset.

B Participant Systems

In the following we list the systems of all participants who submitted a system description paper. The team name used for the submission is in bold; in case the team used a different name on the leaderboard, it is appended in parentheses; the list of subtasks the team participated in is also given in brackets; in case the team was ranked first for at least

	opinion	reporting	satire
English	434	112	24
French	175	65	21
German	148	48	31
Italian	274	72	18
Polish	174	44	23
Russian	170	73	20
Georgian	19	10	0
Greek	39	22	3
Spanish	14	9	7
ALL	1447	455	147

Table 15: Statistics for the genre labels for all languages and the entire dataset.

a subtask-language pair, the list of all such pairs where it is ranked first is given; a list of keywords; and finally, a short description of the system.

ACCEPT [ST2] (Heinisch et al., 2023) (Keywords: *XLM-RoBERTa, ConceptNet*) They used an ensemble combining XLM-RoBERTa with static multilingual and monolingual word embeddings; for the latter, they translated the non-English-texts to English using Google Translate. They further experimented with external common sense knowledge graphs, specifically ConceptNet.

APatt [ST3] (Purificato and Navigli, 2023) (first for: ST3:EN) (Keywords: *XLNet, RoBERTa, BERT, ALBERT, DeBERTa*) They used an ensemble of pre-trained language models fine-tuned on the propaganda dataset: BERT, RoBERTa, ALBERT, XLNet, DistilBERT, and HerBERT. Whenever more LLMs for the same language were available, their output is combined through a weighted average.

Appeal for attention (AppealForAtt) [ST3] (Amihaesei et al., 2023) (Keywords: *XLM-RoBERTa-Large, WordNet, data augmentation*) They focused on data augmentation techniques. They translated the datasets from each language into all other languages using the DeepTranslator API, and they extracted synonyms from WordNet to generate new sentences. Finally, they trained XLM-RoBERTa-large on that augmented data.

BERTastic [ST2] (Mahmoud and Nakov, 2023) (Keywords: *mBERT, XLM-RoBERTa*) They used cross-lingual transformers, mBERT and XLM-RoBERTa, using different orderings of the languages when doing fine-tuning. They further used test data augmentation via translation of both the training and the test sets.

CLaC (CLAC) [ST3] (Costa et al., 2023) (Keywords) *RoBERTa* augmented the dataset translating examples from other languages, focusing on articles having least represented techniques in order to balance the dataset. They used a RoBERTa-base model trained on the English language and made predictions on the other languages by first translating the text into English. They report better F_1 scores with such approach on French and German than using Large Language Models trained directly on the target language.

DSHacker [ST1, ST3] (Modzelewski et al., 2023) (first for: ST1:ES) (Keywords: *XLMRoBERTa-large, GPT3-Davinci, sequence classification, text generation, Ensemble Learning, XGBoost, Logistic Regression, LightGBM, BERT, Data Augmentation, Summarization*) created, for ST1, synthetic texts for each class using the OpenAI GPT-3 Davinci language model. Each language was augmented by approximately 500 articles per genre, producing roughly 13,500 artificially generated articles. Single XLM-RoBERTa-large model was trained using the original and augmented data. For ST3 they developed for Polish an ensemble consisting of three one-vs-rest classifiers: eXtreme Gradient Boosting, Logistic Regression and Light Gradient-Boosting Machine with HerBERT embeddings and various stylometric features using StyloMetrix library. For all other languages BERT-based pre-trained models were deployed and they used summarization applied to longer paragraphs. Furthermore, training data was augmented through machine translation utilizing the DeepL API.

FramingFreaks [ST1, ST2] (Baumann and Deisenhofer, 2023) (Keywords: *SVM, Logistic Regression*) classified texts by splitting them into subwords and then using these tokens as input to a Support Vector Machines for ST1 and to Logistic regression for ST2.

FTD [ST1, ST2, ST3] (Lepekhin and Sharoff, 2023) (first for: ST1:PL) (Keywords: *SLM-RoBERTa, multilingual BERT, Electra, monolingual BERT-based models, fine-tuning, uncertainty estimation, ensembles*) focused on ST1, where they experimented with monolingual and multilingual models, ensembles, additional data, and uncertainty estimation. For Russian and English, they fine-tuned models pre-trained on the FTD dataset for genre classification. For English, they added 1,000 reporting texts from Gigaword. For Polish and German, their best results were achieved by fine-tuning a monolingual Polish BERT and a monolingual German Electra, respectively. For the other languages, their best systems used multilingual BERT,

Framing	English	French	German	Italian	Polish	Russian	Georgian	Greek	Spanish	ALL
Capacity and resources	56	62	104	120	88	34	1	10	11	486
Crime and punishment	274	22	44	57	57	51	3	11	4	523
Cultural identity	42	34	46	43	48	13	1	8	0	235
Economic	74	79	108	142	144	68	2	14	4	635
External regulation and reputation	214	85	91	132	86	44	9	9	3	673
Fairness and equality	131	30	35	52	39	21	0	8	2	318
Health and safety	86	60	107	97	144	37	4	8	3	546
Legality, Constitutionality, jurisprudence	281	41	65	73	56	44	0	23	7	590
Morality	231	62	39	62	63	31	2	5	7	502
Policy prescription and evaluation	154	38	70	129	110	15	2	12	7	537
Political	343	108	130	178	144	55	10	43	6	1017
Public opinion	68	34	50	58	74	22	4	10	3	323
Quality of life	115	40	53	89	85	32	0	5	3	422
Security and defense	222	89	121	155	105	90	10	19	10	821

Table 16: Statistics for the framing labels for all languages and the entire dataset.

Persuasion technique	English	French	German	Italian	Polish	Russian	Georgian	Greek	Spanish	ALL
Name Calling-Labeling	1,945	903	2,818	1,470	1,391	483	42	37	42	9,131
Guilt by Association	84	210	216	98	234	59	4	8	12	925
Doubt	887	679	606	2,295	574	957	40	129	37	6,204
Appeal to Hypocrisy	82	220	307	149	329	167	1	77	14	1,346
Questioning the Reputation	162	662	837	819	555	598	23	35	66	3,757
Flag Waving	434	87	100	72	176	152	3	19	6	1,049
Appeal to Values	47	219	163	264	246	93	6	26	14	1,078
Appeal to Popularity	76	149	119	79	86	38	0	7	11	565
Appeal to Fear-Prejudice	554	443	339	589	245	135	1	12	41	2,359
Appeal to Authority	207	175	377	118	133	22	2	5	6	1,045
Causal Oversimplification	265	228	62	88	22	65	2	29	9	770
False Dilemma, No Choice	241	169	55	149	28	59	2	10	22	735
Consequential Oversimplification	21	215	50	53	47	110	8	33	7	544
Strawman	64	242	27	111	25	46	2	13	6	536
Red Herring	97	72	52	48	23	6	10	19	4	331
Whataboutism	25	93	33	11	22	17	0	13	7	221
Conversation Killer	176	352	235	468	126	172	12	24	26	1,591
Slogans	234	230	176	122	64	113	3	15	16	973
Appeal to Time	4	71	33	53	24	41	0	3	8	237
Loaded Language	3,467	2,533	604	2,878	654	1,347	26	88	134	11,731
Obfuscation-Vagueness-Confusion	37	185	108	42	66	62	6	34	9	549
Exaggeration-Minimisation	730	527	307	261	197	225	5	37	35	2,324
Repetition	938	198	17	75	48	115	20	18	14	1,443

Table 17: Statistics for the fine-grained persuasion techniques for all languages and the entire dataset.

XLM-RoBERTa, or ensembles thereof. In all cases, they truncated the input to the first 510 tokens. They further upsampled the data to balance the distribution between the classes (the results without upsampling were low). For English they further, experimented with uncertainty estimation, and showed that replacing the model predictions that have high uncertainty with the majority class on the training data was helpful on the dev set. For ST2, for each language, they used the model and the setup that worked best for ST1, and just retrained it on the ST2 data.

HHU [ST1] (Billert and Conrad, 2023): (Keywords: *XLM-RoBERTa, Adapters, AdapterFusion*) used an Adapter-based configuration: Using XLM-RoBERTa as a base, they stacked first a language-specific adapter and then a task-specific adapter on top of it. Moreover, they augmented each dataset by translating the articles from the datasets in the other languages.

Hitachi [ST1, ST2] (Koreeda et al., 2023) (first for: ST1:IT, ST1:RU) (Keywords: *XLM-RoBERTa, RoBERTa*) augmented the dataset for ST1 by collecting labelled examples from similar datasets. They pretrained (XLM-)RoBERTa in multi-task (one language, ST1 and ST2), multilingual (one subtask, all languages) and multilingual multi-task (ST1 and ST2 in all languages) settings. Besides using the single models, they report experiments with ensemble of base models with different hyper-

parameters.

JUSTR00 [ST1, ST2] (Al-Qarqaz and Abdullah, 2023): (Keywords: *LongFormer, BERT, RoBERTa, mBERT, XLM-RoBERTa, BigBird* experimented with many state-of-the-art transformer-based language models, both monolingual and multilingual. Their top performing model is based on a transformer called "Longformer".

kb [ST1, ST3] (Baraniak and Sydow, 2023) (Keywords: *BERT, Bert, hierarchical learning, multitask learning*) tackled ST3 by training a BERT model to identify the start of a text fragment with a technique, then the first index of predicted span was used to get the BERT embedding for classification.

KInITVeraAI (KInIT) [ST3] (Hromadka et al., 2023): (first for: ST3:IT, ST3:RU, ST3:DE, ST3:PL, ST3:EL, ST3:KA) (Keywords: *XLM-RoBERTa large and base, mBERT base, monolingual RoBERTa base and large, monolingual BERT base, distilBERT, language model fine-tuning with different layer freezing strategies*) used a fine-tuned XLM-RoBERTa-large transformer model trained on all the input data. They carefully adjusted the prediction threshold for each language using a principled approach. They truncated the input, and also found that pre-processing did not impact the quality much.

MarsEclipse [ST2] (Liao et al., 2023): (first for: ST2:IT, ST2:RU, ST2:FR, ST2:DE, ST2:PL) (Keywords: *XLM-RoBERTa, mBERT, SimCSE, SimCLR*) used a multi-label contrastive loss for fine-tuning pre-trained language models in a multilingual setting. They followed the general architecture of SimCLR and SimCSE to do contrastive learning, but modified the contrastive loss to make it fit for a multi-label setup. This yielded very competitive results for ST2, and this was the winning system for five of the languages.

mCPT (PolarIce) [ST2] (Reiter-Haas et al., 2023) (first for: ST2:ES) (Keywords: *paraphrase-multilingual-MiniLM-L12-v2, contrastive pre-training*) used a two-phase training procedure of a transformer model, first by pre-training jointly on all the languages and then by fine-tuning for each language. In both phases, a multi-label contrastive loss was used.

MELODI [ST1] (Devatine et al., 2023): (first for: ST1:EN) (Keywords: *Translation + POLITICS (RoBERTa)*) fine-tuned the domain-specific language model trained on English data, POLITICS, on the English input articles and on the articles in all other languages automatically translated. In addition, in order to use whole articles as input, they used a sliding window approach and aggregated each window representation with mean pooling. They also tested other multilingual approaches, such as XLM-RoBERTa, and approaches able to process long documents (Longformer), which were in general less effective.

MLModeler5 [ST1, ST2] (Khanchandani et al., 2023) (Keywords: *RoBERTa, ALBERT*) provided a solution for English only. For ST1 they pre-trained the RoBERTa, ALBERT and other deep learning models using the original training data in English and translated versions of the data in other languages and performed NLP augmentation using NLPAUG library on it. For ST2 a similar-in-nature approach was used.

NAP [ST3] (Falk et al., 2023) (first for: ST3:FR) (Keywords: *XLM-RoBERTa (base and large), setfit, adapters, translation and backtranslation of paragraphs*) presented an approach combining predictions of several models in an ensemble, which differ in three main aspects: a) training data, b) model architecture, and c) input format to the model. They leveraged (back-)translation as data augmentation strategies using available MarianMT models. Model architectures included XLM-RoBERTa models, Adapters, SetFit, and linguistically-informed heuristics for under-represented techniques which were fine-tuned on different combinations of original and augmented data. They fine-tuned models on both paragraph- and span-level information.

NL4IA [ST3] (Pritzkau, 2023) (Keywords: *RoBERTa*) used RoBERTa and exploited the span level annotations framing ST3 as a token-level classification one, but report better results when treating the subtask as a sequence classification one.

NLUBot101 [ST1,ST3] (Liu et al., 2023) (Keywords: *mDeBERTa*) built, for ST3, a solution on top of mDeBERTa NLI model and exploit cross-lingual data augmentation. The performance could be improved through the exploitation of the expanded definitions of the persuasion technique guidelines from the official annotation guidelines vis-a-vis the usage of single words or phrases. Their system achieved the highest macro F_1 score for the English language.

QCRI (QCRITeam) [ST1, ST2, ST3] (Hasanain et al., 2023) (Keywords: *XLM-RoBERTa, French Europeana BERT, Gottbert-base, Italian BERT,*

HerBERT) used, for all subtasks, data augmentation and then fine-tuned a BERT model specifically for each language, in addition to fine-tuning XLM-RoBERTa on all languages at once.

QUST [ST1,ST2,ST3] (Jiang, 2023): (Keywords) *XLM-RoBERTa* Their model is build on top of XLM-RoBERTa, which is fine-tuned with the pre-calculated class weights and sample weights to combat the imbalanced data. The class weights are multiplied by the loss to make the model focus more on the minority class. The sample weights are combined with a weighted sampler to resample the distribution of the training batch. In addition, two types of fine-tuning strategies, the task-agnostic and the task-dependent, where the latter proved to help the multilingual model to learn the shared information between subtasks. The submitted system achieves the second best result for Italian and Spanish (zero-shot) in ST1.

ReDASPersuasion [ST3] (Qachfar and Verma, 2023) (Keywords: *XLM-RoBERTa*) uses XLM-RoBERTa as a backbone model, incorporating language agnostic features, computed over the articles translated using Google translation. Such features target specific techniques, including sentiment- and polarity- based features targeting appeal to fear and slogans, indefinite pronouns indicative of exaggeration and minimisation, a profanity language detection to capture loaded language. XLM-RoBERTa has been proved to be a powerful multilingual pre-trained language model compared against other models like Multilingual BERT (M-BERT).

SheffieldVeraAI (vera) [ST1, ST2, ST3] (Wu et al., 2023) (first for: ST1: DE, ST2:EN, ST2:EL, ST2:KA) (Keywords: *mBERT, adapters, text preprocessing, upsampling, XLM-Roberta, Pfeiffer Adapters, MUPPET, Task-adaptive Pre-training, RoBERTa, class weighting*) deployed an ensemble of three fine-tuned mBERT models and one mBERT model with a bottleneck adapter for ST1. All used bert-base-multilingual-cased. For the fine-tuned mBERT models, they pre-processed the data by filtering out non-informative sentences. The pool of training data was also extended by integration additional "satire" resources for English. In the cases where the length of the tokenised article was more than 512 tokens, an equal number of sentences from the beginning and the end of the article was selected until the size of 512 tokens in a concatenated text is reached. The final predictions were drawn as a majority-voting predicted class

For ST2 they used two different ensembles of MUPPET large, and of XLM-R with adapters and task-adaptive MLM pre-training on the train+dev+test data. Their data was pre-processed and truncated. The models were trained both with and without class weighting.

For ST3 they trained a monolingual RoBERTa-Base model for English and a multilingual mBERT-cased model for the remaining languages. They used class weighting to account for class imbalance. They also experimented with augmenting data through translation, which improved the performance for the surprise languages.

SinaAI (SinaaAI) [ST1, ST2, ST3] (Sadeghi et al., 2023) (first for: ST1:EL) (Keywords: *XLM, mBERT, LaBSE*) used multilingual languages models such as XLM, mBERT and LaBSE, which they combined in an ensemble. For ST1 and ST2, they further used data augmentation by selecting 30% of the sentences of each document to create new synthetic examples.

TeamAmpa [ST2, ST3] (Pauli et al., 2023) (first for: ST3:ES) (Keywords: RoBERTa, XML-R, ensemble models) used different oversampling strategies, data truncation, and multilingual and monolingually trained models, combined in an ensemble for the English data. The surprise languages were handled using the multilingual model with oversampling on English data and data from low-represented classes.

TheSyllogist [ST2] (Mohammed Afzal and Nakov, 2023): (Keywords: *BERT*) participated in ST2, and experimented with zero-shot transfer: translating the data for all languages into English (using Google Translate), and then training and applying an English system. They used fine-tuned BERT (bert-base-uncased) with mean-pooling.

UM6P [ST1, ST3] (Alami et al., 2023): (Keywords:*Longformer, RoBERTa, GCN*) fine-tuned Longformer and RoBERTa transformers for both ST1 and ST3. They further added a graph convolution network, and a classifier based on the number of sentences in each document. Finally, they used an ensemble to combine the predictions of these models.

UMUTeam [ST1, ST2] (Pan et al., 2023) (first for: ST1:FR, ST1:DE) (Keywords: *Sentence transformers, XML-RoBERTa*) used a multilingual model based on XML-RoBERTa, which they fine-tuned on all languages at once and a sentence transformer to extract the most important chunk of text

for ST1 and ST2. They further truncated the input data to 200 tokens with 50 tokens of overlap using the sentence-transformer model to obtain the subset of text most related to the article title.

UnedMediaBiasTeam [ST1, ST3] (Rodrigo-Ginés et al., 2023) (Keywords: *XLM-RoBERTa, bert-base-multilingual-cased*) solutions are based on two-stage fine-tuned multilingual models. For ST1 they exploit the media bias detection datasets called BABE and MBIC and XLM-RoBERTa model fine-tuned in two stages: first with the BABE and MBIC datasets, and later with the data provided for the task. For ST3 a similar approach is deployed, where instead of training a single model in two phases, two models are trained and the cascading inference is carried out.

Unisa [ST1, ST3] (Bangerter et al., 2023) (Keywords: *DistilBert, SHAP*) built solutions on top of DistilBert and leverage the application of the eXplainable Artificial Intelligence (XAI) method, Shapley Additive Explanations (SHAP). In ST1, data augmentation was exploited through translation data to the target language (English) on top of which the model was trained with only the first 512 tokens of the articles being considered as input to the model. SHAP was used to understand what was driving the model to fail so that it could be improved.

In ST3, a re-calibration of the Attention Mechanism is realized by extracting critical tokens for each technique. XAI is exploited for countering the overfitting of the resulting model and attempting to improve the performance when there are few training samples. First, a binary model first processes a new incoming paragraph to predict whether it contains any persuasion attempt. If the text is predicted to be propaganda, it is compared with SHAP Vocabularies previously created, which represent the most important words associated with each persuasion technique. Such comparison defines the additional input to pass to the final multi-class model trained to focus on the span that identifies the text that characterizes the persuasion technique.

UTB-NLP (UTBNLP) [ST1, ST2] (Cuadrado et al., 2023) (Keywords:) used a feature-based representation: they extracted noun phrases and represented them as tf-idf vectors; they considered several features specific for each of the three classes of ST1, such as psycholinguistic, writing style, readability, structural characteristics, conceptual embeddings and argumentation-based features. In addition, they used SMOTE to oversample the minority classes.

ST2 was tackled by collecting extra texts from Wikipedia related to the frames, pre-processing them to create a frame-related lexicon and then to use it to create a bag-of-words representation for each input article.