A Corpus of Persuasion Techniques in Slavic Languages

Jakub Piskorski[⋆], Dimitar Iliyanov Dimitrov[†], Marina Ernst[‡], Jacek Haneczok[♢], Michał Marcińczuk[♠], Arkadiusz Modzelewski[§], Roman Yangarber[♡]

[⋆]Institute of Computer Science, Polish Academy of Science, Poland,
[†]Sofia University "St. Kliment Ohridski", Bulgaria, [‡]University of Koblenz, Germany,
[♢]Visa Technology Europe, [♠]CodeNLP,
[§]University of Padua, Italy, [♡]University of Helsinki, Finland
jpiskorski@gmail.com, ilijanovd@fmi.uni-sofia.bg, marinaernst@uni-koblenz.de,
jacek.haneczok@gmail.com, marcinczuk@gmail.com, contact@amodzelewski.com,
roman.yangarber@helsinki.fi

Abstract

Persuasion techniques are powerful rhetorical devices used to sway public opinion in a wide range of media. We present a new corpus of persuasion techniques, focusing on Slavic languages. The corpus contains documents in Bulgarian, Polish, and Russian, annotated with persuasion techniques at the coarse-grained text-span level and fine-grained sentence level. The techniques are drawn from a taxonomy of 25 fine-grained persuasion techniques, grouped under six broad categories of rhetorical persuasion strategies. The corpus contains approximately 7500 text spans from 222 documents that cover topics hotly debated at the national and international levels. We describe the corpus creation process, provide detailed statistics, and examine correlations between topics and persuasion techniques. We use classic ML-based and generative AI-based models to provide baselines and benchmark results for the detection and classification of persuasion techniques at the text-span level and sentence level.

Keywords: persuasion techniques, text classification, linguistic resources, Slavic languages, machine learning

1. Introduction

Persuasion is central to political debates, affecting policy outcomes, and is heavily used in a wide range of contexts, including by social media influencers, to sway public opinion. Persuasion techniques function as psychological mechanisms designed to steer the reader's beliefs and actions, such as voting. Many are based on flawed or unsound reasoning in constructing arguments, while others deliberately trigger emotional responses—such as *appeal to patriotism*—to secure agreement when factual support is missing or weak.

In this paper, we present a new corpus of texts annotated with persuasion techniques in Slavic languages. It contains documents from parliamentary debates in Bulgarian and Polish, and social media in Russian, annotated with persuasion techniques, using a taxonomy of 25 fine-grained persuasion techniques. The corpus consists of approximately 7,500 text fragments annotated with persuasion techniques, in 222 documents. The corpus covers highly debated topics at the international—e.g., the Ukraine-Russia war—and national level—e.g., abortion legislation.

The main impetus behind the creation of the presented corpus is to foster and stimulate research on the detection of persuasion techniques and the classification of texts in Slavic languages. We aim to cover the domains of parliamentary debates and social media, for which there are few or no resources on persuasion techniques in Slavic languages. Our contributions can be summarized as follows:

- we release a new corpus of persuasion techniques, annotated at the text-span and sentence level in three Slavic languages—Bulgarian, Polish, and Russian—for two text genres: parliamentary debates and social media posts,
- we provide the characteristics of this corpus, including statistics, topics, and correlations among the labels,
- we release and evaluate classical ML- and generative-AI-based baseline models for detection and classification of persuasion techniques at the text-span and sentence levels.

It is important to emphasize that text-span annotations of persuasion techniques may cover single words, sub-parts of a sentence, entire sentences or even bigger chunks of text that span more than one sentence. The sentence-level annotations are obtained by mapping the text-span annotations to sentences.

The paper is organized as follows. Section 2 covers related work. Section 3 describes the creation of the corpus. Section 4 presents insights into correlations between topics and persuasion techniques. Section 5 introduces baseline models for

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3599–3616
11–16 May 2026.

persuasion technique detection and classification. Section 6 concludes and outlines future work.

2. Related Work

Research on automated detection of five types of logical fallacies was reported by Habernal et al. (2017, 2018). Da San Martino et al. (2019b) presented a corpus of English news articles labeled using a more fine-grained taxonomy of 18 persuasion techniques at span and sentence level, and reported on experiments of automated solutions to detect them. This corpus was used in *NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection* (Da San Martino et al., 2019a) and *SemEval-2020 Task 11 on Detection of Persuasion Techniques in News Articles* (Da San Martino et al., 2020), focused on the detection of persuasion techniques in text fragments and document-level classification, with an initial taxonomy of 18 techniques, in English news articles.

Piskorski et al. (2023c) introduced an extended taxonomy of 23 persuasion techniques, grouped in 6 different categories, and a corpus of 1.7K news articles in 6 languages (including two Slavic languages, Polish and Russian) annotated at the text-span level using this taxonomy. The usefulness and applicability of this taxonomy was demonstrated by Modzelewski et al. (2025), who employed it to enhance disinformation detection through the Persuasion-Augmented Chain-of-Thought method. The corpus was used in SemEval-2023 Task 3 on Detecting Category, Framing, and Persuasion Techniques in Online News in a Multi-lingual Setup (Piskorski et al., 2023b). The CLEF 2024 Task 3 on Persuasion Techniques (Piskorski et al., 2024) followed up on Semeval-2023 Task 3, by including new articles in five languages, namely, Arabic, *Bulgarian*, English, Portuguese, and *Slovene*, (with two additional Slavic languages), and evaluation of persuasion detection and classification at text-span level.

Recently, work has been reported on detection and classification of persuasion techniques in other text genres, e.g., the shared task *DIPROMATS 2023*: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers (Moral et al., 2023) and DIPROMATS 2024: Detection, characterization and tracking of propaganda in messages from diplomats and authorities of world powers (Moral et al., 2024), which resulted in a release of a dataset of approximately 21K tweets in English and Spanish, posted by authorities of China, Russia, the United States, and the European Union.

Detection and classification of persuasion techniques in parliamentary debates and social media was the subject of the *SlavicNLP 2025 Shared Task* (Piskorski et al., 2025). In particular, the task focused on the classification of persuasion techniques at the paragraph level, covered 5 Slavic languages (Bulgarian, Croatian, Polish, Russian, and Slovene), and adapted and extended the persuasion technique taxonomy presented in Piskorski et al. (2023c) to 25 techniques. Our new dataset, presented in this paper, has been annotated using the same taxonomy.

Kyslyi et al. (2025) reports on a shared task on manipulation detection and classification in Ukrainian social media, for which a new dataset containing approximately 9.5K posts from Ukrainian Telegram in both Russian and Ukrainian were introduced, and labeled with a subset variant of the persuasion taxonomies used in the SemEval tasks mentioned earlier. Moreover, Modzelewski et al. (2024) proposed a taxonomy comprising 11 manipulation techniques and released a Polish disinformation dataset annotated based on this taxonomy. Finally, Alzahrani et al. (2024) presented a study on the detection and classification of persuasion techniques in Arabic social media.

Another line of research focused on detecting persuasion techniques in multimodal content. For instance, detection of persuasion techniques in memes was the subject of the shared tasks: *SemEval-2021 Task 6 on Detection of Persuasion Techniques in Texts and Images* (Dimitrov et al., 2021) and *SemEval-2024 Task 4 on Multilingual Detection of Persuasion Techniques in Memes* (Dimitrov et al., 2024). These tasks extended the Semeval-2020 taxonomy to detect persuasion in visual content as well, totaling 22 techniques (20 multimodal and 2 vision-only). Apart from English, the second task covered two Slavic languages: Bulgarian and Macedonian.

Accurate analysis of persuasion techniques is critical for text-understanding tasks; for example, in information extraction (Piskorski and Yangarber, 2013; Huttunen et al., 2002), persuasion techniques introduce the complex interplay between the factual and the rhetorical dimensions of the content: does the content convey objective reporting of events or attempt to influence the reader's perception of events. Relying on event detection methods alone is insufficient to assure reliability and trust (Atkinson et al., 2011; Yangarber, 2006). As persuasive content becomes ever more pervasive online, our ability to accurately detect it and analyze all of its aspects becomes more important.

To the best of our knowledge, the corpus we present is the first for Bulgarian and Polish that covers text-span and sentence-level annotations in the domain of parliamentary debates, and the first for Russian social media, annotated with fine-grained persuasion techniques. This constitutes part of the corpus used for the shared task de-

scribed in (Piskorski et al., 2025), which focused on paragraph-level detection and classification of persuasion techniques in Slavic languages.

3. Corpus Creation

3.1. Taxonomy

To label the texts in our corpus, we use the 2-tier taxonomy from the SlavicNLP 2025 shared task on persuasion techniques (Piskorski et al., 2023c). At the top level are 6 coarse-grained *persuasion strategies*: *Attack on Reputation*, *Justification*, *Simplification*, *Distraction*, *Call*, and *Manipulative Wording*—described in detail below.

Attack on reputation: The argument shifts the focus—from addressing the topic, toward targeting the participants—their personality, experience, deeds, etc.—in order to question or undermine their credibility. The target of the argument can be a group of individuals, organization, or activity.

Justification: The argument has two parts: a statement and an explanation or appeal, where the latter is used to justify or 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 focus away from the main topic 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., to impact the reader emotionally.

These six strategies, subdivided into 25 fine-grained persuasion techniques (see Figure 1), are defined in detail with examples in Annex A.

ATTACK ON REPUTATION
- Name Calling or Labelling
- Guilt by Association
- Casting Doubt
- Appeal to Hypocrisy
- Questioning the Reputation

JUSTIFICATION
- Flag-Waiving, appeal to patriotism
- Appeal to Authority
- Appeal to Popularity
- Appeal to Fear, Prejudice
- Appeal to Values

DISTRACTION
- Strawman
- Whataboutism
- Red Herring
- Appeal to Pity

SIMPLIFICATION
- Causal Oversimplification
- False Dilemma or No Choice
- Consequential Oversimplification
- False Equivalence

CALL
- Slogans
- Conversation Killer
- Appeal to Time

MANIPULATIVE WORDING
- Loaded Language
- Obfuscation, Intentional Vagueness, Confusion
- Exaggeration or Minimisation
- Repetition

Figure 1: Two-tier Persuasion Technique taxonomy.

3.2. Document Acquisition

To create the corpus, we collected documents in three languages—Bulgarian, Polish, and Russian—covering various controversial topics. For Bulgarian and Polish, we use transcripts of the respective parliamentary sessions. For Russian, we use social media posts, particularly from the Telegram platform, focusing on community-based channels.

Topics covered in Bulgarian parliamentary debates include: foreign policy (with a focus on military aid to Ukraine), acceptance into the Eurozone and Schengen, and national sovereignty concerns. Domestic political discourse (priorities, integrity, identity) is intertwined with international matters (e.g., ongoing conflicts).

Topics covered in the Polish debates cover the highly-disputed abortion legislation, national security and defense policy, Poland's role within the European Union, legislation against hate speech and discrimination, socio-economic matters—vaccination, forest management, mass layoffs, mental health awareness, etc.

The Russian documents focus predominantly on the Ukraine-Russia war, e.g., Putin-Trump negotiations, Russia's opposition with the West, disinformation, demographic challenges (such as migration, integration), criticism of state policies in this context, and civilian resistance in conflict zones.

Overall, the documents in the three languages are multiply labeled with the following broader topics / categories: *Ukraine-Russia war* (URW), *Defense*, *Israel-Hamas conflict*, *Migration*, *Demographics*, *Abortion*, *EU*, *Schengen*, *External affairs*, *Climate change*, *Hate speech*, *History*, *Other*.

3.3. Annotation Process

The annotation process included five steps:

1. For each language, a team of at least two annotators, one curator, and one language coordinator was set up, where all team members were native speakers and had prior experience in linguistic annotations, including annotation in disinformation and manipulative content,
2. We provided all teams with detailed annotation guidelines (Piskorski et al., 2023a), and organized live sessions with annotation trainings,
3. Each document was annotated by two annotators; ad-hoc meetings were organized to discuss difficult and unclear cases,
4. The curator for each language verified the adherence of the annotations to the guidelines, and corrected them as needed,
5. Regular meetings were held and an information exchange platform was set up to safeguard and maintain consistency in annotations across languages, given the complexity of the task at hand (Stefanovitch and Piskorski, 2023).

We adapted INCEpTION (Klie et al., 2018)—a web-based framework for collaborative annotation.

In the remainder of this section, we outline the types of persuasion techniques that proved most challenging for the annotators to detect reliably. Their subtlety lies not in overt manipulation but in how they blend into everyday reasoning, requiring the reader to track context, nuance, and argument structure to recognize them.

Whataboutism and *red herrings* are particularly tricky in complex debates, where topic-switching feels relevant. They rely on the reader losing track of the original issue as the discussion expands. Spotting these techniques requires considering the broad context and noticing when the focus has shifted without justification. *False equivalence* can be equally deceptive, as it presents opposing ideas in a "balanced" format that appears fair, requiring a deep understanding of the underlying issues to see that the comparison is flawed. *Strawman* arguments distort the opposing position slightly, often in long or nuanced exchanges; detecting them requires remembering what was said earlier and recognizing subtle misrepresentations. Causal or consequential *oversimplification* appeals to our desire for clear, linear explanations, though real-world problems often involve multiple causes, feedback loops, and uncertainties. Identifying this type of fallacy requires a cognitive effort to weigh the missing variables and alternative outcomes. Finally, *obfuscation*, intentional vagueness, and confusion hide flawed reasoning behind abstract or technical language; the reader feels something meaningful has been said but can't specify what. Detecting this tactic requires noticing the absence of clarity rather than the presence of persuasion, i.e., it is a task that demands sustained attention and critical reflection across the broader context.

	BG	PL	RU
Documents	79	53	90
Sentences	4984	4515	2074
Sentences with PTs	2491	2092	1218
Text spans annotated	3275	3048	1089
AVG words/document	1146	1215	359

Table 1: Dataset statistics across languages.

3.4. Annotation Format

For each language in the dataset, there is one annotation file, where each line contains annotations at **sentence level** and has the following format:

`<FILE> <START> <END> <PT-LABELS>`

where `<FILE>`, is the name of the file, `<START>` and `<END>` are the start/end character position of the sentence, and `<PT-LABELS>` is a list of persuasion technique labels. If the given sentence does not contain any persuasion techniques, then the list is empty. All elements in each line are separated by tabs.

For each language, the dataset contains a corresponding file with **single text-span annotations**, where each line has the format:

`<FILE> <START> <END> <PT> <NORM>`

where `<PT>` contains a unique persuasion technique label and `<NORM>` contains whitespace-normalized text which corresponds to the text fragment starting/ending at `<START>`/`<END>`.

All documents in the corpus are provided in plain text format encoded in UTF-8.

3.5. Statistics

The corpus consists of 222 documents in three languages—Bulgarian, Polish, and Russian—with a total of 7412 instances of persuasion techniques. The overall statistics of the corpus are provided in Table 1. The proportion of sentences labeled with at least one persuasion technique for Bulgarian, Polish, and Russian is 50.0%, 46.3%, and 58.7%, respectively. The dataset covers 25 persuasion techniques, with a highly unbalanced distribution. Figure 2 provides the distribution of the persuasion techniques and comparison across languages. *False Equivalence* is the least frequent class (74 instances), while *Loaded language* is the most frequent (1110 instances).

The heatmap in Figure 3 shows the co-occurrence of persuasion-technique annotations within sentences across all languages. Each cell shows the number of sentences that contain both persuasion techniques corresponding to the row and the column. The diagonal cells indicate the total number of sentences annotated with a specific persuasion technique. Darker colors indicate higher counts, highlighting techniques that

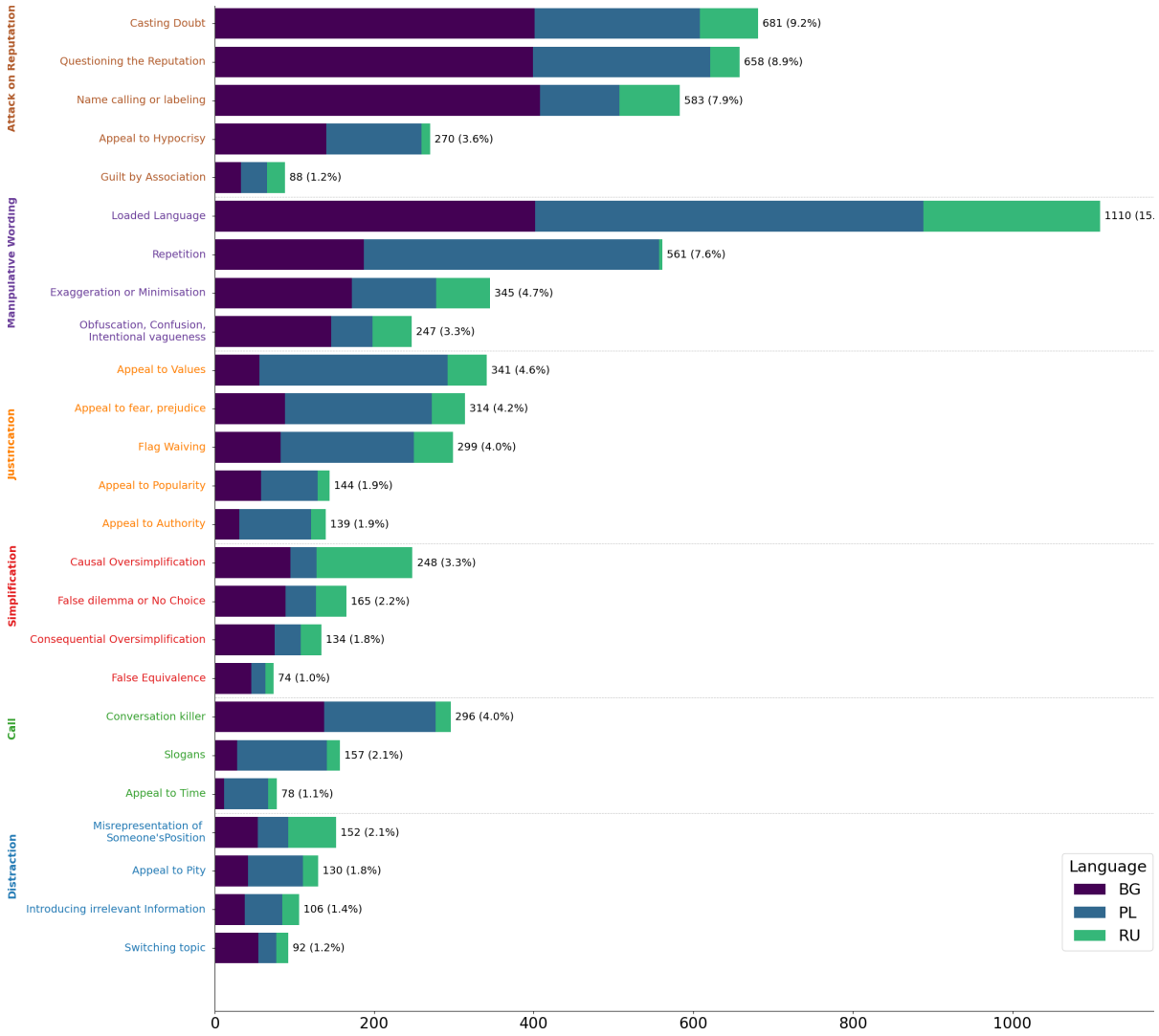

Figure 2: Distribution of persuasion technique annotations by language.

Topic	BG	%%	PL	%%	RU	%%
Abortion	-		14	26	1	1
Climate change	2	2.5	8	15	-	
Defense	14	17	22	41	-	
Demographics	1	1.2	5	9	9	10
Hate speech	-		8	15	-	
EU	28	35	15	28	-	
External affairs	2	2.5	13	25	21	23
History	-		2	4	-	
Israel-Hamas conflict	2	2.5	-		-	
Migration	7	8.5	8	15	13	14.5
Schengen	7	8.5	-		-	
Ukraine-Russia war	42	52	5	9	52	58
Other	4	5	1	2	4	4.5

Table 2: Topic distribution across languages.

frequently recur within the same sentence. The axes list all persuasion techniques, grouped and color-coded by their broader categories (e.g., *Attack on Reputation*, *Manipulative Wording*, *Justification*, *Simplification*, *Call*, *Distraction*).

Table 2 summarises the distribution of documents by topic across the Bulgarian (BG), Polish (PL), and Russian (RU) datasets. The figures reveal significant cross-linguistic variation. All three languages engage with the `URW` topic, which accounts for 44% of documents across languages, though it is particularly prevalent in Russian (58%) and Bulgarian (52%) texts. `External affairs` emerges as another cross-linguistic theme, particularly relevant in Russian (23%) and Polish (25%) materials. In contrast, while `EU` dominates Bulgarian texts (35%), `Abortion` debates appear almost exclusively in Polish discourse (26%)

4. Topic vs. Persuasion Technique Analysis

To better understand how different persuasion techniques and strategies are employed in public discourse, we analyze their co-occurrence with the topics under discussion. Figure 4 presents a heatmap of persuasion-topic co-occurrence: the

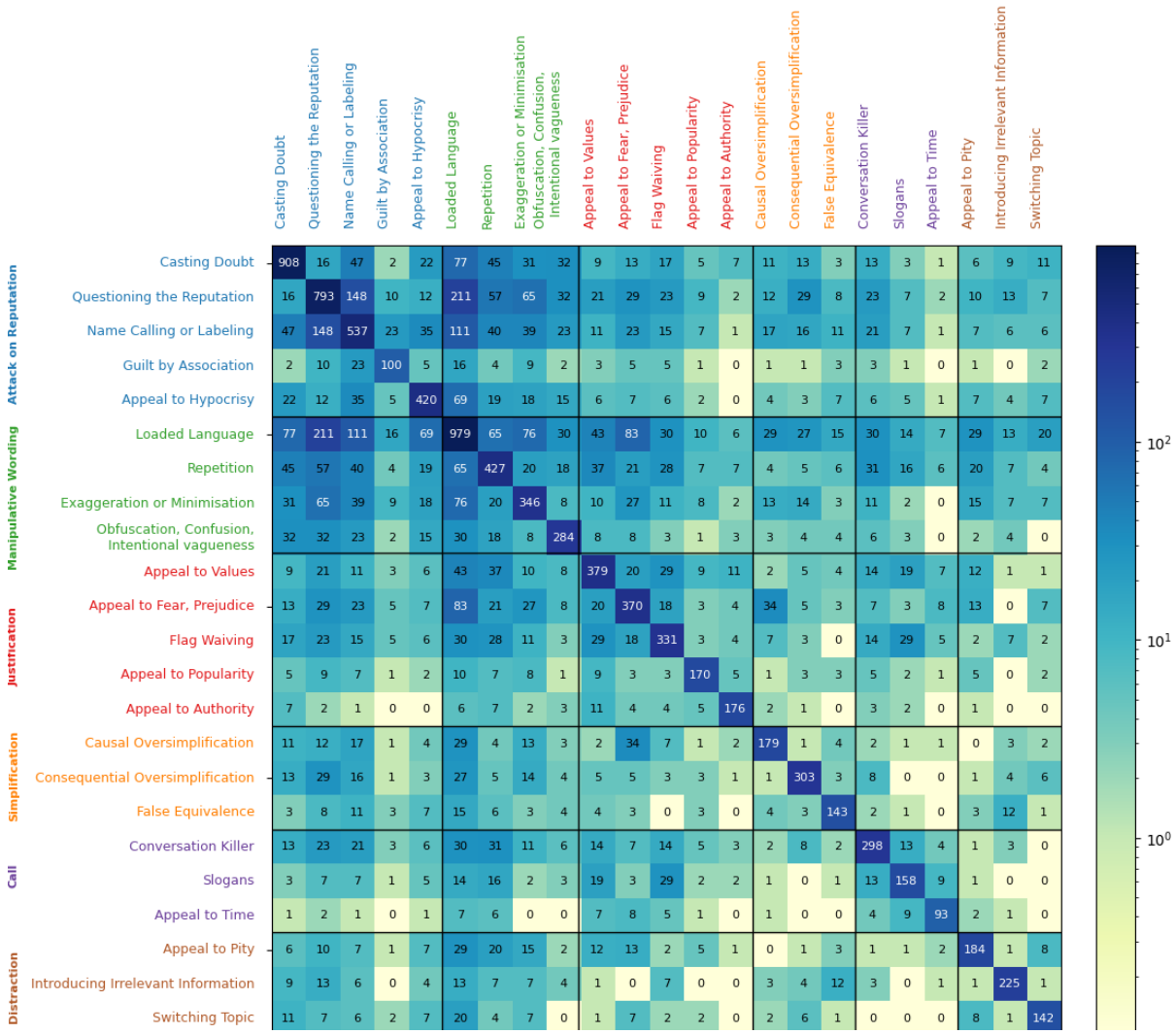

Figure 3: Co-occurrence of persuasion technique annotations within sentences.

rows represent persuasion techniques, grouped into strategies, and the columns represent topics. Color intensity reflects the prevalence of each strategy within a topic, while the numerical values provide precise percentages.

Across all topics, *Manipulative Wording* and *Attack on Reputation* are the most prominent persuasion strategies, with *Loaded Language* being the most prevalent technique. *Doubt* and *Questioning the Reputation* frequently occur in politically sensitive discussions, particularly those related to `EU`, `Defense`, `Schengen`, `Migration`.

Several techniques exhibit topic-specific spikes. For example, *Appeals to values* occurs more frequently in documents related to `Abortion` and `Demographic`, whereas *Appeal to Fear* becomes particularly important in `Hate speech`. In contrast, *Doubt* and *Questioning the Reputation* appear frequently in documents about `URW`, the most common topic in the corpus. It is important to note that *Loaded language* is a dominant technique in the Russian data, while Bulgarian contributes most with *Doubt*. Notably, the topics `Schengen` and `Migration` evoke similar persuasion techniques.

5. Baseline Models

In this Section, we present baseline models for the task of detection and classification of persuasion techniques. Since this paper focuses on the corpus, we provide these models and results as benchmarks (rather than as state-of-the-art)—to highlight the *complexity* of the classification task.

5.1. Text-span level

To demonstrate the complexity of the Persuasion Technique Classification problem, we build a baseline model that, given a text fragment known to contain one or more persuasion techniques and no additional context, determines the persuasion

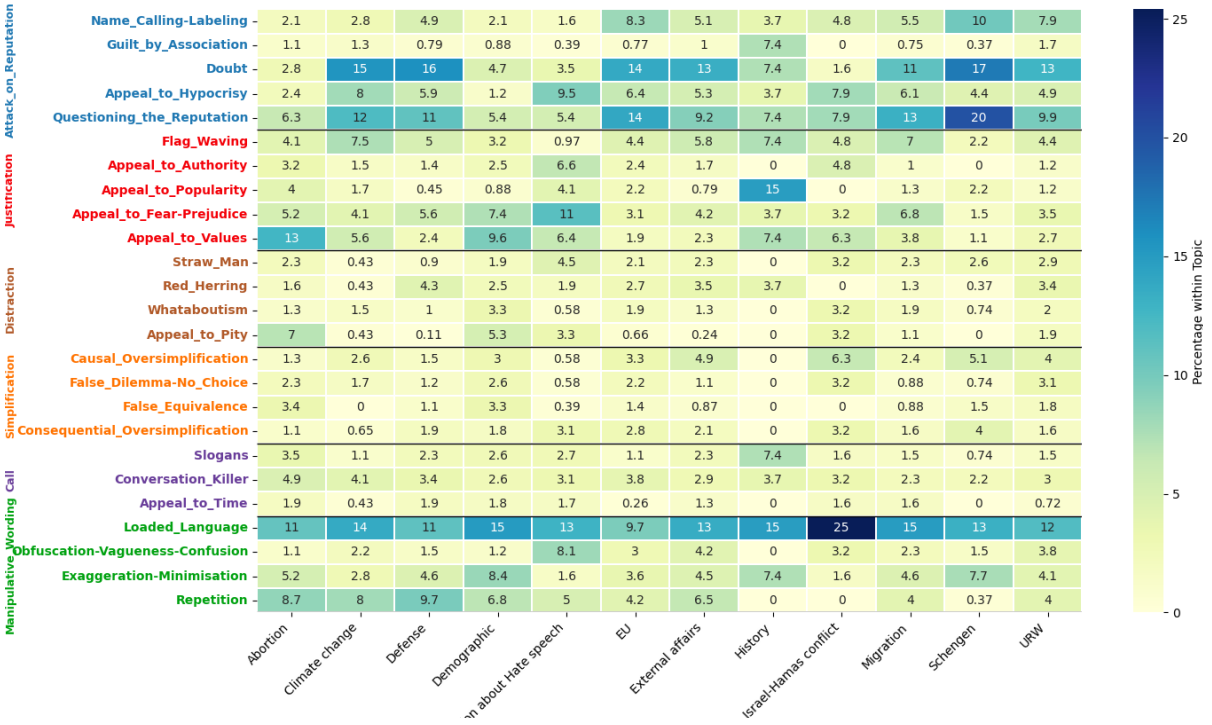

Figure 4: Correlation between persuasion technique and topics across languages.

technique. For this purpose, we trained a SVM[1] on text spans annotated with persuasion techniques, which uses 3-5 character n-grams as features, with vector normalization, and ran 5-fold cross-validation on the entire data for each language. The classification results (i.e., F_1 scores) for individual techniques are provided in Table 3. The macro (micro) F_1 for the persuasion technique classification task is .19 (.34), .18 (.35), and .04 (.11) for Bulgarian, Polish, and Russian, respectively. For the individual techniques, we can observe that lexical features have some discriminative power for the classification of *Loaded Language*, *Name Calling and Labeling*, *Repetition*, *Conversational Killer*, *Flag Waiving*, *Appeal to Values*, where these classes appear to have a higher number of instances in the corpus. On the other hand, *Distractions* and *Simplifications* are less frequent, and intuitively harder to detect. In general, results for Russian are worse than those for Bulgarian and Polish, likely due to the smaller amount of available data for Russian.

	BG		PL		RU	
Technique	F_1	sup	F_1	sup	F_1	sup
Questioning the Reputation	.26	393	.19	222	.00	37
Name-Calling, Labeling	.56	343	.37	90	.09	76
Guilt by Association	.00	31	.00	32	.00	22
Doubt	.40	396	.19	207	.05	73
Appeal to Hypocrisy	.22	139	.09	118	.00	11
Appeal to Popularity	.22	57	.09	71	.00	15
Appeal to Values	.03	56	.41	236	.15	49
Appeal to Authority	.11	31	.28	89	.00	18
Appeal to Fear/Prejudice	.10	86	.15	184	.00	42
Flag Waving	.24	83	.39	167	.04	49
Causal Oversimplification	.05	95	.00	33	.07	120
False Dilemma-No Choice	.10	89	.04	38	.00	38
Consequential Oversimplification	.02	75	.00	33	.00	26
False Equivalence	.07	46	.00	18	.00	10
Straw Man	.06	54	.00	38	.00	60
Whataboutism	.03	55	.21	22	.00	15
Red Herring	.05	35	.07	47	.00	21
Appeal to Pity	.14	42	.11	69	.00	19
Appeal to Time	.00	12	.14	55	.15	11
Slogans	.12	26	.12	112	.00	16
Conversation Killer	.46	129	.31	136	.00	19
Loaded Language	.60	363	.68	461	.32	222
Obfuscation-Vagueness-Confusion	.14	145	.12	52	.00	49
Exaggeration-Minimization	.14	164	.06	106	.00	67
Repetition	.59	123	.59	298	.00	4
ALL (micro)	.34	3068	.35	2934	.11	1089
ALL (macro)	.19	3068	.18	2934	.04	1089

Table 3: Text-span classification—F_1 and support—of character n-gram based SVM. Scores above .5, in range .4–.5, and in range .3–.4 highlighted in green, lime green, and yellow, respectively.

5.2. Sentence level

5.2.1. Coarse-Grained Persuasion Strategy Classification with LLMs

Experimental Setup: To evaluate the capability of large language models (LLMs) to detect persuasion at the sentence level, we assess LLM performance on two classification tasks:

- **Binary Classification** – Each persuasion strategy is treated as an independent binary classi-

[1]We used the Liblinear library at: www.csie.ntu.edu.tw/~cjlin/liblinear

	BG		PL		RU	
Model	F_1 Macro	F_1 Micro	F_1 Macro	F_1 Micro	F_1 Macro	F_1 Micro
gemini-2.0-flash	.28	.35	.27	.34	.20	.26
google/gemma-3-27b-it	.27	.40	.29	.36	.20	.28
gpt-4.1-mini	.26	.33	.27	.34	.21	.25
gpt-4o-mini	.29	.35	.28	.32	.23	.27
meta-llama/Llama-3.3-70B-Instruct	.25	.33	.24	.32	.22	.28

Table 4: Results for multi-label classification of high-level persuasion strategies. Results presented for six different LLMs. Experiments done at the sentence level with zero-shot prompting. Scores above .5, in range .4–.5, and in range .3–.4 highlighted in green, lime green, and yellow, respectively.

	Model	Attack on Reputation	Call	Distraction	Justification	Manipulative Wording	Simplification
BG	gemini-2.0-flash	.45	.18	.18	.19	.47	.32
	google/gemma-3-27b-it	.56	.16	.22	.22	.46	.16
	gpt-4.1-mini	.44	.20	.26	.15	.47	.36
	gpt-4o-mini	.45	.23	.28	.10	.47	.39
	meta-llama/Llama-3.3-70B-Instruct	.42	.12	.08	.14	.48	.18
PL	gemini-2.0-flash	.40	.22	.14	.38	.50	.14
	google/gemma-3-27b-it	.44	.10	.10	.32	.47	.28
	gpt-4.1-mini	.43	.27	.15	.28	.46	.18
	gpt-4o-mini	.41	.20	.18	.22	.48	.17
	meta-llama/Llama-3.3-70B-Instruct	.44	.15	.22	.15	.46	.22
RU	gemini-2.0-flash	.26	.12	.08	.24	.40	.28
	google/gemma-3-27b-it	.37	.12	.12	.17	.33	.18
	gpt-4.1-mini	.46	.18	.10	.16	.41	.29
	gpt-4o-mini	.42	.11	.14	.00	.38	.27
	meta-llama/Llama-3.3-70B-Instruct	.41	.12	.10	.08	.40	.28

Table 5: Binary classification results for each of six high-level persuasion strategies. Experiments done at the sentence level with zero-shot prompting. Scores above .5, in range .4–.5, and in range .3–.4 highlighted in green, lime green, and yellow, respectively.

fication problem. Models are evaluated on their ability to detect the presence or absence of a specific persuasion strategy.

- **Multilabel Classification** – LLMs are assessed on their ability to identify multiple persuasion strategies that may co-occur within a single sentence, formulated as a multilabel classification.

We evaluate five state-of-the-art LLMs accessed via their respective APIs[2]: *GPT-4o Mini*, *GPT-4.1 Mini*, *Gemini 2.0 Flash*, *Gemma 3 27B IT*, and *Llama 3.3 70B*. To ensure that the results are as deterministic as possible, all models were prompted with the temperature parameter set to zero. Evaluations were conducted in a zero-shot setting for all languages. In total, the experiments involved approximately 55,000 API calls. The complete set of prompts and the accompanying codebase are available in our public repository to support transparency and reproducibility.[3]

Results: Overall performance is moderate and comparable across languages, reflecting the difficulty of zero-shot persuasion detection. In the multilabel setup (Table 4), micro F_1 ranges from .26 to .40, with Gemma 3 27B achieving the best results, particularly for Bulgarian and Polish.

For the binary classification task (Table 5), variation across persuasion strategies is substantial. The models perform best on *Attack on Reputation* and *Manipulative Wording*, often exceeding the F_1 score of .45, while strategies such as *Call* and *Distraction* remain challenging. Performance is generally consistent across languages, though Russian shows slightly lower scores.

5.2.2. Fine-Grained Persuasion Technique Classification with Custom SLMs

Experimental Setup: We fine-tuned a transformer-based model for each language—*PKOBP/polish-roberta-8k*[4] for Polish, *ai-forever/ruRoberta-large*[5] (Zmitrovich et al., 2024) for Russian, and *FacebookAI/xlm-roberta-large*[6] (Conneau et al., 2019) for Bulgarian. Training was conducted for 60 epochs with a learning rate of 5e-6, batch size of 16, maximum sequence length of 512 tokens, and weight decay of 0.01. For each language-specific model, the checkpoint yielding the highest micro F_1

[2]For open-source LLMs we used APIs provided by DeepInfra: deepinfra.com/models

[3]Link to repository: github.com/SlavicNLP/SlavicPersuasionTechniques/

[4]huggingface.co/PKOBP/polish-roberta-8k

[5]huggingface.co/ai-forever/ruRoberta-large

[6]huggingface.co/FacebookAI/xlm-roberta-large

	BG			PL			RU		
Technique	train	valid	test	train	valid	test	train	valid	test
Questioning the Reputation	315	139	41	181	45	33	22	13	4
Name-Calling, Labeling	233	97	38	66	12	11	44	13	22
Guilt by Association	24	8	2	28	9	4	13	8	4
Doubt	316	152	62	221	43	26	47	19	20
Appeal to Hypocrisy	129	73	26	135	25	15	8	1	7
Appeal to Popularity	39	22	3	57	21	8	13	1	5
Appeal to Values	41	14	8	176	60	19	32	11	18
Appeal to Authority	35	13	6	71	16	12	12	5	4
Appeal to Fear/Prejudice	70	33	13	144	29	32	27	10	12
Flag Waving	67	21	10	124	34	17	32	9	16
Causal Oversimplification	80	19	10	31	8	2	90	27	36
False Dilemma-No Choice	64	25	12	35	8	3	23	11	17
Consequential Oversimplification	61	22	15	34	6	6	21	10	3
False Equivalence	49	22	8	34	9	4	10	4	3
Straw Man	42	19	10	50	12	6	39	22	10
Whataboutism	57	28	8	16	7	3	11	3	5
Red Herring	85	30	15	43	12	7	16	7	10
Appeal to Pity	31	5	8	86	11	8	13	8	13
Appeal to Time	10	2	1	42	15	8	9	3	3
Slogans	13	7	7	86	21	5	10	3	6
Conversation Killer	92	35	15	96	25	16	11	2	6
Loaded Language	223	93	37	294	58	44	122	44	63
Obfuscation-Vagueness-Confusion	102	27	15	49	12	11	30	19	19
Exaggeration-Minimisation	87	48	21	75	22	21	37	15	20
Repetition	82	37	16	176	53	45	8	2	3
Sentences	3138	1346	500	2809	703	460	1101	473	500
With any PT	1548	676	267	1496	362	222	654	261	303
Without any PT	1590	670	233	1313	341	238	447	212	197

Table 6: Number of sentences with each type of PT for the train-valid-split.

score on the validation subset was selected, and final results were reported on the corresponding held-out test set. Table 6 presents the sentence-level statistics, showing the number of sentences with each type of PT per language and subset (train, validation, and test).

Results: Table 7 presents results achieved by transformer-based classifiers[7]. The low performance for Russian may be attributable to the limited data relative to other languages—approximately 1,000 sentences, compared to 3,000 for Polish and Bulgarian in the train subset (see Table 6). Consequently, the setup for Russian was adjusted to train on the combined training and validation subsets. The final model, after 60 epochs, achieves modest recognition for some techniques: *Name-Calling, Labeling* (F_1 = .27), *Appeal to Values* (F_1 = .27), *Causal Oversimplification* (F_1 = .26).

For Bulgarian and Polish, each with approximately 3,000 instances, the most frequent techniques yielded relatively high F_1 values: *Questioning the Reputation* (.42 for Bulgarian; .40 for Polish), *Loaded Language* (.42 for Bulgarian; .59 for Polish). Certain techniques performed better in one language than in others—which is a pattern that correlates with the frequency of instances: for Bulgarian, *Doubt* (.53) and *Consequential Oversimplification* (.42); for Polish, *Appeal to Values* (.61) and *Appeal to Authority* (.61).

6. Conclusion and Future Work

This paper describes a new corpus of excerpts from parliamentary debates in Bulgarian and Polish, and social media texts in Russian, annotated with fine-grained persuasion techniques at the text-span level and at the sentence level. The corpus covers highly controversial topics discussed in both national and global discourse. The paper gives an overview of the corpus creation process, detailed statistics about the datasets, and an analysis of the correlation of topic and persuasion techniques. It presents baseline models for the task of persuasion technique detection and classification at the text-span and sentence level, and reports on the performance of these models in order to demonstrate the complexity of these tasks. We make the corpus freely available for research purposes. eWe believe that together with the accompanying baseline models, it will be useful for the community working on the detection of manipulative content in general, and in Slavic languages in particular.

In the future, we plan to: (a) expand the size of the corpus, (b) cover a wider range of Slavic languages, (c) improve the models by generating synthetic data, and (d) explore in depth the detection of the more challenging persuasion techniques, in particular, those that fall under the strategies of *Simplifications* and *Distractions*.

	BG		PL		RU	
Technique	F_1	sup	F_1	sup	F_1	sup
Questioning the Reputation	.42	41	.40	33	.0	4
Name-Calling, Labeling	.68	38	.27	11	.27	22
Guilt by Association	.0	2	.0	4	.0	4
Doubt	.53	62	.22	26	.15	20
Appeal to Hypocrisy	.33	26	.35	15	.0	0
Appeal to Popularity	.0	3	.17	8	.0	5
Appeal to Values	.0	8	.61	19	.27	18
Appeal to Authority	.0	6	.61	12	.0	4
Appeal to Fear/Prejudice	.40	13	.41	32	.0	12
Flag Waving	.40	10	.17	17	.11	16
Causal Oversimplification	.11	10	.0	2	.26	36
False Dilemma-No Choice	.20	12	.0	3	.0	17
Consequential Oversimplification	.42	15	.0	6	.0	3
False Equivalence	.0	8	.0	4	.0	3
Straw Man	.29	10	.0	6	.0	10
Whataboutism	.17	8	.0	3	.0	5
Red Herring	.40	15	.40	7	.0	10
Appeal to Pity	.0	8	.29	8	.0	13
Appeal to Time	.0	1	.33	8	.0	3
Slogans	.0	7	.0	5	.0	6
Conversation Killer	.36	15	.36	16	.0	6
Loaded Language	.42	37	.59	44	.17	63
Obfuscation-Vagueness-Confusion	.08	15	.0	11	.0	19
Exaggeration-Minimization	.21	21	.09	21	.0	20
Repetition	.17	16	.21	45	.0	3
Without any PT	.64	259	.63	246	.57	198
ALL (micro)	.46	666	.44	612	.28	527
ALL (macro)	.24	666	.23	612	.07	527

Table 7: Sentence level classification—F_1 and support for each class—for BERT-architecture models. Scores above .5, in range .4–.5, and in range .3–.4 are highlighted in green, lime green, and yellow.

[7] huggingface.co/collections/SlavicNLP/PersuasionTechniques-LREC2026

Acknowledgments

Work carried out in Bulgaria was partially supported by the European Union—NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria [Grant Project No. BG-RRP-2.004-0008].

Work carried out in Finland was supported in part by the Research Council of Finland, Project *"Know-AI"* (Grant 1359285), and by European Regional Development Fund (ERDF) Project *"Generative AI and Knowledge Management"* (Grant 4740347). We are grateful for their support.

Ethics

Intended Use and Misuse Potential: The corpus presented in this paper was created to foster research on detection and classification of persuasion techniques for the domain of parliamentary debates and social media. Given the potential risks of exploiting this corpus to produce manipulative content, we strongly advise responsible use of the data.

Fairness: The annotators were either: (a) researchers from the institutions of the authors of this manuscript, (b) students from the respective academic organizations, or (c) experts from a contracted professional annotation company. The annotators in the first two groups were fairly remunerated as part of their job, whereas the experts in the third group were compensated at rates based on their country of residence.

Limitations

Dataset Representativeness: The corpus covers parliamentary debates and propaganda narratives in various countries, and we strove to include utterances of speakers covering a wide political spectrum in each of these countries. However, we must emphasize that these datasets should not be considered representative of the political landscape in any specific country or region, nor should they be considered as balanced in any way.

Biases: We have invested significant effort in training the annotators and acquainting them with the specifics of the persuasion technique taxonomy. Furthermore, cross-language quality control mechanisms have been implemented to ensure the highest quality of annotations. Nevertheless, some degree of intrinsic subjectivity might be present in the datasets. Therefore, models trained using these datasets might exhibit certain biases.

References

Abdurahmman Alzahrani, Eyad Babkier, Faisal Yanbaawi, Firas Yanbaawi, and Hassan Alhuzali. 2024. Investigating persuasion techniques in Arabic: An empirical study leveraging large language models.

Martin Atkinson, Jakub Piskorski, Erik van der Goot, and Roman Yangarber. 2011. Multilingual real-time event extraction for border security intelligence gathering. In U. Kock Wiil, editor, *Counterterrorism and Open Source Intelligence*, pages 355–390. Springer Lecture Notes in Social Networks, Vol. 2.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019a. 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*, pages 162–170, Hong Kong, China. Association for Computational Linguistics.

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

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *EMNLP*.

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

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav

Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *EMNLP*.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018. Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *LREC: Conference on Language Resources and Evaluation*.

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

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.

Roman Kyslyi, Nataliia Romanyshyn, and Volodymyr Sydorskyi. 2025. The UNLP 2025 shared task on detecting social media manipulation. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 105–111, Vienna, Austria (online). Association for Computational Linguistics.

Arkadiusz Modzelewski, Giovanni Da San Martino, Pavel Savov, Magdalena Anna Wilczyńska, and Adam Wierzbicki. 2024. MIPD: Exploring manipulation and intention in a novel corpus of Polish disinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19769–19785, Miami, Florida, USA. Association for Computational Linguistics.

Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. 2025. PCoT: Persuasion-augmented chain of thought for detecting fake news and social media disinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24959–24983, Vienna, Austria. Association for Computational Linguistics.

Pablo Moral, Jesús M Fraile, Guillermo Marco, Anselmo Peñas, and Julio Gonzalo. 2024. Overview of dipromats 2024: Detection, characterization and tracking of propaganda in messages from diplomats and authorities of world powers. *Procesamiento del lenguaje natural*, 73:347–358.

Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de Albornoz, and Iván Gonzalo-Verdugo. 2023. Overview of dipromats 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers. *Procesamiento del lenguaje natural*, 71:397–407.

Jakub Piskorski, Dimitar Iliyanov Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michal Marcinczuk, Arkadiusz Modzelewski, Ivo Moravski, and Roman Yangarber. 2025. SlavicNLP 2025 shared task: Detection and classification of persuasion techniques in parliamentary debates and social media. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 254–275, Vienna, Austria. Association for Computational Linguistics.

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

Jakub Piskorski, Nicolas Stefanovitch, 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. 2023a. News categorization, framing and persuasion techniques: Annotation guidelines. Technical report, European Commission Joint Research Centre, Ispra (Italy).

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023b. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in

online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

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

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

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

Roman Yangarber. 2006. Verification of facts across document boundaries. In *Proceedings of the International Workshop on Intelligent Information Access (IIIA-2006)*, Helsinki, Finland.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 507–524, Torino, Italy.

A. Persuasion Technique Definitions

Below, we provide definitions of the persuasion techniques accompanied by examples in English (in blue) and in the Slavic languages (in brown) of the Shared Task. The text fragments highlighted in bold are the text spans to be annotated according to the guidelines presented in (Piskorski et al., 2023a).

The definitions of the persuasion techniques are taken directly from the Annex of (Piskorski et al., 2023c), with two new persuasion techniques, *Appeal to Pity* and *False Equivalence*, added for this task.

A.1. Attack on Reputation

Name Calling or Labeling: a form of argument in which loaded labels are directed at an individual or a group, typically in an insulting or demeaning way. An object is labeled as something the target audience fears, hates, or, on the contrary, finds desirable or loves. This technique calls for a qualitative judgement that disregards facts and focuses solely on the essence of the subject being characterized. This technique is also in a way manipulative wording, as it appears as a nominal group rather than being a full-fledged argument with a premise and a conclusion. For example, in political discourse, typically one uses adjectives and nouns as labels that refer to political orientation, opinions, personal characteristics, and association to some organisations, as well as insults. What distinguishes it from *Loaded Language* (see A.6), is that it is concerned only with the characterization of the subject.
Example: ***'Fascist' Anti-Vax*** *Riot Sparks COVID Outbreak in Australia.*
Example: *Trzeba zrozumieć, że bronią także i polskich granic przeciwko* ***rosyjskiemu imperializmowi****, którego ducha wskrzesił Władimir Putin—* ***prezydent zbrodniarz****.* (*It is necessary to understand that they are also defending the Polish borders against* ***Russian imperialism****, whose spirit has been revived by Vladimir Putin—****the criminal president.****)

Guilt by Association: Attacking an opponent or an activity by associating it with another group, activity, or concept that has sharply negative connotations for the target audience. The most common example, which has given its name in the literature to this technique (i.e., *Reduction ad Hitlerum*) is making comparisons with Hitler and the Nazi regime. However, it is important to emphasize, that this technique is not restricted to comparisons to that group only. More precisely, this can be done by claiming a link or an equivalence between the target of the technique and any individual, group, or event in the present or in the past, which is or was negatively perceived (e.g., was considered a failure), or is depicted in such a way.
Example: ***Manohar is a big supporter for equal pay for equal work. This is the same policy that all those extreme feminist groups support.*** *Extremists like Manohar should not be taken seriously.*
Example: *Мы часто забываем, что после Второй мировой наши типа союзники, французы* ***(на самом деле настоящие союзники Гитлера)****, стали срочно восстанавливать*

свою империю. (*We often forget that after WWII our so-called allies, the French* ***(in fact, Hitler's allies)****, immediately started rebuilding their empire.*)

Casting Doubt: Casting doubt on the character or the personal attributes of someone or something in order to question their general credibility or quality, rather than using a proper argument relevant to the topic. This can be done for instance, by speaking about the target's professional background, as a way to discredit their argument. Casting doubt can also be done by referring to some actions or events carried out or planned by some entity that are/were not successful, or appear as resulting in not achieving the planned goals.

Example: *This task is quite complex.* ***Is his professional background, experience and the time left sufficient to accomplish the task at hand?***

Example: *Има един-единствен кореспондент от българска страна и по нашите медии се твърди, че те били обективни, представяли реално гледната точка, казвали истината и прочее.* (***There was only one reporter from Bulgaria, and our media claimed that they were objective, presented a realistic point of view, told the truth, and so on.***)

Appeal to Hypocrisy: The reputation of the target is attacked by charging them with hypocrisy or inconsistency. This can be done explicitly by calling out hypocrisy directly, or implicitly by underlining the contradictions between different positions that were held or actions that were done in the past. A common way of calling out hypocrisy is by saying that someone who criticizes you for something you have done, has done it himself in the past.

Example: ***How can you demand that I eat less meat to reduce my carbon footprint if you yourself drive a big SUV and fly for holidays to Bali?***

Example: *Иначе СЕМ твърди, че е безпристрастен, но когато става въпрос за безпочвени обвинения към Русия или манипулиране на общественото мнение по този начин, някак си СЕМ пропуска това.* (***Otherwise, the CEM claims to be impartial, but when it comes to groundless accusations against Russia or manipulating public opinion in this way, the CEM somehow misses the mark.***)

Questioning the Reputation: This technique is used to attack the reputation of the target by making strong negative claims about it, focusing on undermining its character and moral stature rather than relying on an argument about the topic. Whether the claims are true is irrelevant for the effective use of this technique. Smears can be used at any point in a discussion. One way of using this technique is to preemptively call into question the reputation/credibility of an opponent, before he has a chance to express himself, therefore biasing the audience's perception. Hence, one of the names for this technique is "poisoning the well."

The main difference between *Casting Doubt* (above) and *Questioning the reputation* is that the former focuses on questioning the capacity, capabilities, and credibility of the target, while the latter aims to undermine the overall reputation, moral qualities, behaviour, etc.

Example: *I hope I presented my argument clearly. Now,* ***my opponent will attempt to refute my argument by his own fallacious, incoherent, illogical version of history.***

Example: ***Jedni i drudzy rządziliście Polską od 20 lat i nie przygotowaliście nas do obrony na czas wojny.*** (***Together, you have governed Poland for 20 years and failed to prepare us for defense in times of war.***)

A.2. Justification

Flag Waving: Justifying or promoting an idea by appealing to the pride of a group or highlighting the benefits for that specific group. The stereotypical example would be national pride, and hence the name of the technique; however, the target may be any group, e.g., related to race, gender, political preference, etc. The connection to nationalism, patriotism, or benefit for an idea, group, or country might be inappropriate and is usually based on the presumption that the recipients already hold certain beliefs, biases, and prejudices about the given issue. It can be seen as an appeal to emotions instead to logic of the audience aiming to manipulate them to win an argument. As such, this technique can also appear outside well-constructed arguments, by making statements that resonate with the particular group and as such setting up a context for further arguments.

Example: ***We should make America great again, and restrict the immigration laws.***

Example: ***Wolna Ukraina i silna Unia Europejska, silna Polska stanowią podstawę polskiej racji stanu, to podstawa naszego bezpieczeństwa.*** (***A free Ukraine and a strong European Union, a strong Poland, are the foundation of the Polish national interest, they are the basis of our security.***)

Appeal to Authority: attempting to add weight to an argument, an idea or information by simply stating that a particular entity considered to be an authority is the source of the information. The entity mentioned as an authority may, but does not need to be, an actual authority in the specific domain to discuss a particular topic or to serve as an expert. What is important, and makes it different from simply sourcing information, is that the tone of the text capitalizes on the weight of the

alleged authority in order to justify some claim or conclusion. Referencing a valid authority is not a logical fallacy, while referencing an invalid authority is a logical fallacy, and both are captured within this label. In particular, a self-reference as an authority falls under this technique as well.
Example: ***Since the Pope said that this aspect of the doctrine is true we should add it to the creed.***
Example: *Глава ЦБ РФ Эльвира Набиуллина назвала новые реалии тектоническими изменениями в мировой торговле, и с учётом всех нюансов происходящего это ещё очень деликатная формулировка.* (***The head of the Central Bank of Russia Elvira Nabiullina called the new situation a "tectonic shift in global trade," and considering all the nuances of what is happening, this is still a very delicate formulation.***)

Appeal to Popularity: This technique gives weight to an argument or idea by justifying it on the basis that allegedly "*everyone*" (or the vast majority) agrees with it, or "*nobody*" disagrees with it. The target audience is encouraged to gregariously adopt the same idea by considering "*everyone*" as an authority, and to join in and take the same course of action. Here, "*everyone*" might refer to the general public, key entities and actors in a certain domain, countries, etc. Analogously, an attempt to persuade the audience not to do something because "*nobody else is taking the same action*" falls under our definition of *Appeal to Popularity*.
Example: ***Because everyone else goes away to college, it must be the right thing to do.***
Example: *По предната точка Ви казах за последното социологическо проучване, в което 78% от българските граждани не искат да се предоставя оръжие на Украйна, а Вие правите точно това, което не искат българските граждани.* (***In the previous point, I told you about the latest sociological survey, in which 78% of Bulgarian citizens do not want weapons to be provided to Ukraine, and you are doing exactly what Bulgarian citizens do not want.***)

Appeal to Values: This technique gives weight to an idea by linking it to values seen by the target audience as positive. These values are presented as an authoritative reference in order to support or to reject an argument. Examples of such values are, for instance: tradition, religion, ethics, age, fairness, liberty, democracy, peace, transparency, etc. When such values are mentioned outside the context of a proper argument by simply using certain adjectives or nouns as a way of characterizing something or someone, such references fall under another label, namely, *Loaded Language*, which is a form of *Manipulative Wording* (see A.6).
Example: ***It's standard practice to pay men more than women so we'll continue adhering to the same standards this company has always followed.***
Example: *В очередной раз удар нанесён по одной из самых чувствительных сфер—религиозным правам и свободам.* (***Another attack has been made on one of the most sensitive areas—religious rights and freedoms.***)

Appeal to Fear, Prejudice: This technique aims at promoting or rejecting an idea through the repulsion or fear the audience feels toward this idea (e.g., via exploiting some preconceived judgements) or toward its alternative. The alternative could be the status quo, in which case the current situation is described in a scary way with *Loaded Language*. If the fear is linked to the consequences of a decision, it is often the case that this technique is used simultaneously with *Appeal to Consequences* (see Simplification techniques in A.4), and if there are only two alternatives that are stated explicitly, then it is used simultaneously with the *False Dilemma* technique (see A.4).
Example: *It is a great disservice to the Church to maintain the pretense that there is nothing problematical about Amoris laetitia.* ***A moral catastrophe is self-evidently underway*** *and it is not possible honestly to deny its cause.*
Example: *Много, много други такива неща са се случвали и за съжаление,* ***ние отиваме по едни стъпки, които са изключително опасни, изключително наистина тревожни за бъдещето на нашата държава.*** (*Many, many other such things have happened, and unfortunately,****we are taking extremely dangerous steps, extremely worrying for the future of our country.***)

A.3. Distraction

Strawman: This technique consists in creating an illusion of refuting the argument of the opponent's proposition, while the real subject of the argument was not addressed or refuted, but instead replaced with a false one. Often, this technique is referred to as a misrepresentation of the argument. First, a new argument is created via the covert replacement of the original argument with something that appears related, but is actually a different, distorted, exaggerated, or misrepresented version of the original proposition, which is referred to as "*setting up a strawman*." Subsequently, the newly created '*false* argument (strawman) is refuted, which is referred to as "*knocking down the strawman*." Often, the strawman argument is created in such a way that it is easier to refute, and

thus, creating the illusion of having defeated an opponent's real proposition. Fighting a strawman is easier than fighting a real person, which explains the name of this technique. In practice, it appears often as an abusive reformulation or explanation of what the opponent *actually* means or intends.
Example: *Referring to your claim that providing medicare for all citizens would be costly and a danger to the free market, I infer* ***that you don't care if people die from not having healthcare, so we are not going to support your endeavour****.*
Example: *Има огромно значение, господин Иванов, дали българското знаме е отляво, или отдясно. Това нещо го знаете по протокол.* ***Ако казвате, че няма значение, това означава, че за Вас няма значение какъв точно ще бъде статутът на българското знаме в България, статутът на българския държавен герб и къде точно ще се полага*** (*It makes a huge difference, Mr Ivanov, whether the Bulgarian flag is on the left or the right. You know this from protocol.****If you say that it does not matter, it means that it does not matter to you exactly what the status of the Bulgarian flag will be in Bulgaria, the status of the Bulgarian state coat of arms and exactly where it will be placed.***)

Red Herring: This technique consists in diverting the attention of the audience from the main topic being discussed, by introducing another topic. The aim of attempting to redirect the argument to another issue is to focus on something the person doing the redirecting can better respond to or to leave the original topic unaddressed. The name of that technique comes from the idea that a fish with a strong smell (such as a herring) can be used to divert dogs from the scent of someone they are following. A strawman (defined earlier) is a specific type of a red herring in that it distracts from the main issue by presenting the opponent's argument in an inaccurate light.
Example: *Lately, there has been a lot of criticism regarding the quality of our product.* ***We've decided to have a new sale in response, so you can buy more at a lower cost!****.*
Example: *Недавно она прочитала лекцию о необходимости войны с ухоженным газоном, потому что «это символ сексизма, расизма и экологического разрушения».* ***Среди друзей Аджубей много проукраинских активистов и адептов движений Black lives matter и ЛГБТ.*** (*She recently gave a lecture on the need for a war on manicured lawns because "they are a symbol of sexism, racism and ecological destruction"* ***Adzhubey's friends include many pro-Ukrainian activists and adherents of the Black lives matter and LGBT movements.***)

Whataboutism: Attempt to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument. Rather than answering a critical question or argument, an attempt is made to retort with a critical counter-question that expresses a counter-accusation, e.g., mentioning double standards, etc. The intent is to distract from the content of a topic and to actually switch the topic. There is a fine distinction between this technique and *Appeal to Hypocrisy*, introduced earlier: the former is an attack on the argument and introduces irrelevant information to the main topic, while the latter is an attack on reputation and highlights the hypocrisy of double standards on the same or a closely related topic.
Example: ***A nation deflects criticism of its recent human rights violations by pointing to the history of slavery in the United States.***
Example: *Добре, на Хърватия е пораснал—окей.* ***А Естония и Финландия, които са на минус, и Ирландия, които са в еврозоната, какво правим?*** (*Okay, Croatia's has grown—okay.* ***And what about Estonia and Finland, which are in the red, and Ireland, which are in the eurozone, what do we do?***)

Appeal to Pity: Evokes feelings of pity, sympathy, compassion or guilt in audience to distract it from focusing on evidence, rational analysis and logical reasoning, so that it accepts the speaker's conclusion as truthful solely based on these emotions. It is an attempt to sway opinions and fully substitute logical evidence in an argument with a claim intended to elicit pity or guilt.
Example: ***If this person is found guilty of this crime, his ten children will be left without a parent at home, therefore the jury must submit a verdict of innocence.***
Example: ***Напуганные, изнурённые отсутствием спокойствия и элементарных условий для жизни****, женщины всё равно не были сломлены и не потеряли надежду на освобождение российскими подразделениями их родного хутора.* (***Frightened, exhausted by the insecurity and lack of basic living conditions****, the women were still not broken and did not lose hope for the liberation of their village by Russian troops*)

A.4. Simplification

Causal Oversimplification: Assuming a single cause or reason when there are actually multiple causes for an issue. This technique has the following logical form(s): (a) *Y occurred after X; therefore, X was the only cause of Y*, or (b) *X caused Y; therefore, X was the only cause of Y (although A, B, C...etc. also contributed to Y.)*
Example: *School violence has gone up and aca-*

demic performance has gone down since video games featuring violence were introduced. ***Therefore, video games with violence should be banned, resulting in school improvement.***
Example: *И если собственно украинские возможности к сопротивлению закончились к концу марта 22го, что и привело Киев к Стамбулу,* ***то выигрыш 3х недель, вселил уверенность в запад, и поэтому было решение Джонсона, продолжать войну.*** (*If Ukraine's own capacity for resistance had run out by the end of March 22, bringing Kyiv to Istanbul, then* ***the three-week gain gave the West confidence, and that is why Johnson decided to continue the war.***)

False Dilemma or No Choice: Sometimes called the *either-or* fallacy, a false dilemma is a logical fallacy that presents only two options or sides when there are actually many. One of the alternatives is depicted as a *no-go* option, hence the only choice is the other option. In extreme cases, the author tells the audience exactly what actions to take, eliminating any other possible choices (also referred to as *Dictatorship*).
Example: ***There is no alternative to Pfizer Covid-19 vaccine. Either one takes it or one dies.***
Example: ***Debatujemy dzisiaj o bardzo ważnej i – ośmielę się powiedzieć – kluczowej dla nas sprawie, sprawie życia i śmierci.*** (***Today we are debating a very important and, I dare say, crucial issue for us, a matter of life and death.***)

Consequential Oversimplification: An argument or an idea is rejected and instead of discussing whether it makes sense and/or is valid, the argument affirms, without proof, that accepting the proposition would imply accepting other propositions that are considered negative. This technique has the following logical form: *if A will happen then B, C, D, ... will happen*. The core essence behind this fallacy is an assertion one is making of some '*first*' event/action leading to a domino-like chain of events that have some significant negative effects and consequences that appear to be ludicrous. This technique is characterized by *ignoring and/or understating the likelihood of the sequence of events from the first event leading to the end point* (last event). In order to take into account symmetric cases, i.e., using *Consequential Oversimplification* to promote or to support certain action in a similar way, we also consider cases when the sequence of events leads to positive outcomes (i.e., encouraging people to undertake a certain course of action(s), with the promise of a major positive event in the end).
Example: ***If we begin to restrict freedom of speech, this will encourage the government to infringe upon other fundamental rights, and eventually this will result in a totalitarian state where citizens have little to no control of their lives and decisions they make.***
Example: *Сокрытие правды и подмена понятий приведет к тому, что управлять умами и историей будет противник на нашей территории, выдавая правду с нужным ему уклоном.* (***Concealing the truth and substituting concepts will result in the enemy controlling minds and history on our territory, spreading the truth with an intended bias.***)

False Equivalence: A technique that attempts to treat scenarios that are significantly different as if they had equal merit or significance. In particular, an emphasis is placed on one specific shared characteristic between the items of comparison in the argument that is off by an order of magnitude, oversimplified, or important additional factors have been ignored. The introduction of certain shared characteristics of the scenarios is then used to consider them equivalent. This technique has the following logical form: *A and B share some characteristic X. Therefore, A and B are equivalent*.
Example: ***The introduction or restrictive hours of alcohol sales boosted the black market industry, and analogously, one can expect that the introduction of too restrictive anti-abortion regulations will lead to growth of the illegal abortion business.***
Example: ***To właśnie Führer jako pierwszy wprowadził wolną aborcję dla Polek oraz dla innych kobiet z narodów podbitych. Chodziło o fizyczne zniszczenie ludności niearyjskiej i zdobycie lebensraumu dla Niemców. Hitler rozumiał, że jeśli zalegalizuje aborcję, stanie się ona zjawiskiem masowym i spowoduje spadek urodzeń. Na ziemiach podbitych przez Niemcy dzieci niearyjskie uważano za zagrożenie, więc wdrażano politykę sprzyjającą aborcji. Równocześnie za to samo, za zabicie dziecka niemieckiego w Niemczech groziła kara śmierci. A dyktator groził: osobiście zastrzelę tego idiotę, który chciałby wprowadzić w życie przepisy zabraniające aborcji na wschodnich terenach okupowanych. Jaka jest analogia? Kto powiedział: każda odmowa aborcji będzie zgłaszana do prokuratury? Premier rządu rewolucji*** (***It was the Führer who first introduced free abortion for Polish women and other women from conquered nations. The idea was to physically destroy the non-Aryan population and gain Lebensraum for the Germans. Hitler understood that if he legalized abortion, it would become a mass phenomenon and cause a decrease in births. In the lands conquered by Germany, non-Aryan children were considered a threat, so a policy favoring abortion was implemented. At the same time, the same thing,***

killing a German child in Germany, was punishable by death. And the dictator threatened: I will personally shoot this idiot who would want to implement regulations prohibiting abortion in the occupied eastern territories. What is the analogy? Who said: every refusal to have an abortion will be reported to the prosecutor's office? The prime minister of the government of the revolution)

Example: *В 1990-х годах были скинхеды—группы асоциальной молодежи, которые толпой нападали на лиц неевропейской наружности, на так сказать «черных». **Теперь скинхеды—это группы асоциальной молодежи среднеазиатской наружности, которые так и не смогли гармонично жить рядом с русскими, и толпой избивают русских парней и насилуют русских девочек*** (*In the 1990s, there were the skinheads—groups of antisocial youth who mobbed people of non-European appearance, the so-called "blacks". **Now skinheads are groups of antisocial youth of Central Asian appearance, who failed to live peacefully next to Russians, and crowds of them beat up Russian guys and rape Russian girls***)

A.5. Call

Slogans: A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.

Example: ***Immigrants welcome, racist not!***

Example: *Да живее България!* (***Long live Bulgaria!***)

Conversation Killer: This includes words or phrases that discourage critical thought and meaningful discussion about a given topic. They are a form of *Loaded Language*, often passing as folk wisdom, intended to end an argument and quell cognitive dissonance.

Example: *I'm not so naïve or simplistic to believe we can eliminate wars. **You can't change human nature.***

Example: ***Takie są fakty i taka jest polska racja stanu.*** (***These are the facts, and this is the Polish national interest.***)

Appeal to Time: The argument is centered around the idea that the time has come for a particular action. The very timeliness of the idea is part of the argument.

Example: *This is no time to engage in the luxury of cooling off or to take the tranquilizing drug of gradualism. **Now is the time to make real the promises of democracy. Now is the time to rise from the dark and desolate valley of segregation to the sunlit path of racial justice.***

Example: ***Można powiedzieć, że w wielu wymiarach nastał czas prawdy.*** (***It can be said that in many ways, the moment of truth has arrived.***)

A.6. Manipulative Wording

Loaded Language: use of specific words and phrases with strong emotional implications (either positive or negative) to influence and to convince the audience that an argument is valid. It is also known as *Appeal to Argument from Emotive Language*.

Example: *They keep feeding these people with **trash**. They should stop.*

Example: ***Nękanie*** *zasłużonej dla szerzenia polskości instytucji **bezzasadnymi** pozwami odbierane jest m.in. przez moich wyborców jako działania mające na celu **sparaliżowanie** funkcjonowania tej fundacji.* (*The **harassment** of an institution that has earned merit in promoting Polish identity through **groundless** lawsuits is perceived, among others by my constituents, as actions aimed at **paralyzing** the functioning of this foundation.*)

Obfuscation, Intentional Vagueness, Confusion: This fallacy uses words that are deliberately unclear, so that the audience may have its own interpretations. For example, an unclear phrase with multiple or unclear definitions is used within the argument and, therefore, does not support the conclusion. Statements that are imprecise and intentionally do not fully or vaguely answer the posed question fall under this category.

Example: ***Feathers cannot be dark, because all feathers are light!*** Example: *Затем следует «вишенка»: дорогой доллар превращается в «пылесос» для капиталов со всего мира.* (***Then comes the "cherry on top": the expensive dollar turns into a "vacuum cleaner" for capital from all over the world.***)

Exaggeration or Minimisation: This technique consists of either representing something in an excessive manner—by making things larger, better, worse (e.g., *the best of the best*, *quality guaranteed*)—or by making something seem less important or smaller than it really is (e.g., saying that an insult was just a joke), downplaying the statements and ignoring the arguments and the accusations made by an opponent.

Example: *From the seminaries, to the clergy, to the bishops, to the cardinals, **homosexuals are present at all levels, by the thousand**.*

Example: ***Europa prowadzi również najbardziej dramatyczną wojnę, wojnę demograficzną, którą przegrywa.*** (***Europe is also fighting its most dramatic war, the demographic war, which it is losing.***)

Repetition: The speaker uses the same word, phrase, story, or imagery repeatedly in the hope

that the repetition will persuade the audience.
Example: ***Hurtlocker deserves an Oscar***. *Other films have potential, but they do not* ***deserve an Oscar like Hurtlocker does***. *The other movies may deserve an honorable mention but* ***Hurtlocker deserves the Oscar***.
Example: *И няма да бъдем готови, защото имаме структурни, чисто фундаментални проблеми и неща, които трябва да решим,* ***и това няма да се случи за 6 месеца, няма да се случи за година, няма да се случи и за две години.*** (*And we will not be ready, because we have structural, purely fundamental problems and issues that we need to resolve,* ***and this will not happen in six months, it will not happen in a year, it will not happen in two years.***)