On Experiments of Detecting Persuasion Techniques in Polish and Russian Online News: Preliminary Study

Nikolaos Nikolaidis[1] **Nicolas Stefanovitch**[2] **Jakub Piskorski**[3]
[1]Athens University of Economics and Business, Athens, Greece nnikon@aueb.gr
[2]European Commission Joint Research Centre, Ispra, Italy nicolas.stefanovitch@ec.europa.eu
[3]Polish Academy of Sciences, Warsaw, Poland jpiskorski@gmail.com

Abstract

This paper reports on the results of preliminary experiments on the detection of persuasion techniques in online news in Polish and Russian, using a taxonomy of 23 persuasion techniques. The evaluation addresses different aspects, namely, the granularity of the persuasion technique category, i.e., coarse- (6 labels) versus fine-grained (23 labels), and the focus of the classification, i.e., at which level the labels are detected (subword, sentence, or paragraph). We compare the performance of mono- verus multi-lingual-trained state-of-the-art transformed-based models in this context.

1 Introduction

Nowadays, readers of online content are exposed more than ever to manipulation, disinformation and propaganda, which can potentially influence their opinion on relevant topics, such as, e.g., elections, health crises, migration crises, military conflicts, etc. Thus, the analysis of online media landscape is essestial in order to get a deeper insight on the presented narratives around certain topics across countries, to detect and identify manipulation attempts and to enchance users' media literacy. As a result, in the recent years, one could observe an ever-growing trend of research on automated methods supporting the detection of potentially deceptive and manipulative content, on narrative extraction, and on tools for comparative analysis of online media of different political orientations.

In this paper, we present the results of some preliminary experiments on the detection of persuasion techniques in online news in Polish and Russian. In order to perform our experiments, we exploit the datasets used in the *SemEval 2023 Shared Task 3: Detecting the Genre, the Framing, and the Persuasion Techniques in Online News in a Multilingual Setup* (Piskorski et al., 2023), whose one specific subtask focuses on the detection of persuasion techniques at paragraph level in nine languages, including, i.a., Polish and Russian, which, to the best of our knowledge, constitutes the first ever annotated resource for persuasion technique detection for these languages at intra-document level. While the aforementioned shared task revolves solely around paragraph-level detection and classification of persuasion techniques using a taxonomy of 23 techniques, in our work, we focus on the evaluation with different settings: (a) the granularity of the data after aggregating the results of the classifier: fine-grained (23 labels) versus coarse-grained (6 labels); and (b) the focus of the classification, i.e., at which level the labels are aggregated: subword, sentence, and paragraph level. The main drive behind the inclusion of these different dimensions in the evaluation is to gain a better understanding about the usability of automated persuasion technique detection for practical applications. The primary focus is to compare the performance of mono- versus multi-lingual-trained state-of-the-art transformed-based models in this context.

The rest of this paper is organized as follows. First, we report on related work in Section 2. Next, the persuasion technique detection task and the underlying taxonomy is introduced in Section 3. Subsequently, in Section 4 we report on the carried out experiments, including the description of the dataset, evaluation methodology, models explored, the results, and some rudimentary error analysis. We end up with the conclusions and future outlook in Section 5.

2 Related Work

The work on automated detection of persuasion techniques in text is related to work on propaganda detection. The work in the latter area initially focused on document-level analysis and predictions. For example, Rashkin et al. (2017) reports on prediction of four classes (*trusted*, *satire*, *hoax*, and *propaganda*) of documents, whereas

Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), pages 155–164

Barrón-Cedeno et al. (2019) developed a corpus of documents tagged either as *propaganda* or *non-propaganda*) and further investigated writing style and readability level.

In parallel to the above, other research work focused on the detection of specific persuasion techniques in text. Habernal et al. (2017, 2018) presented a corpus with 1.3k arguments annotated with 5 fallacies that directly relate to propaganda techniques. A more fine-grained analysis was done by Da San Martino et al. (2019a), who developed a corpus of English news articles labelled with 18 propaganda techniques at span and sentence level, and proposed a deep learning-based solutions for this task. Improved models were proposed addressing the limitations of transformers by Chernyavskiy et al. (2021), whereas the topic of interpretable propaganda detection was addressed by Yu et al. (2021). Somewhat related is also the work on 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). Bonial et al. (2022) reported on the creation of annotated text snippet dataset with logical fallacies for Covid-19 domain and evaluation or ML-based approaches using this corpus. Sourati et al. (2022) presents three-stage evaluation framework of detection, coarse-grained, and fine-grained classification of logical fallacies through adapting existing evaluation datasets, and evaluate various state-of-the-art models using this framework. Jin et al. (2022) proposed the task of logical fallacy detection and a new dataset of logical fallacies found in climate change claims. Noteworthy, all the persuasion techniques and logical fallacy taxonomies introduced in the aforementioned research works do, in principle, overlap to a very high degree, but are structured differently, and different naming conventions are used.

A comprehensive survey on computational propaganda detection is presented in (Da San Martino et al., 2020b).

Various shared tasks related to persuasion technique detection were organized in the recent years. For instance, *SemEval-2020 task 11 on Detection of Persuasion Techniques in News Articles* (Da San Martino et al., 2020a) focused on the detection of persuasion techniques (at text span level) in news articles. The *NLP4IF-2019 task on Fine-Grained Propaganda Detection* task proposed a similar-in-nature task with a taxonomy of 18 persuasion techniques. The *SemEval-2021 task 6 on Detection of Persuasion Techniques in Texts and Images* focused on the detection of propaganda techniques deployed in memes, and used a taxonomy of 22 techniques (Dimitrov et al., 2021b). Finally, WANLP'2022 (Alam et al., 2022) shared task centred around the detection of 20 propaganda techniques in Arabic tweets (Alam et al., 2022), while the *SemEval 2023 Shared Task 3: Detecting the Genre, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup* (Piskorski et al., 2023) has a subtask revolving around the detection of persuasion techniques at paragraph level in nine languages, including: English, French, Georgian, German, Greek, Italian, Polish, Russian, Spanish.

The work on persuasion detection for Polish and Russian is scarce, and focused mainly on the analysis of the use of persuasion techniques, not their detection. For instance (Stepaniuk K., 2021) studies the use of linguistic cues defined as Persuasive Linguistic Tricks (PLT) in social media (SM) marketing communication. (Andrusyak, 2019) studied the use of propaganda techniques in the Russian news in the context of the Russian military intervention in Ukraine in 2014, and also explored NLP-based models for their automated detection, however, this is done at document level, i.e., classification of articles into persuasive and non-persuasive ones, which is different from our work which is at the intra-document level. To our best knowledge, the resources used in the context of the SemEval 2023 Shared Task 3 (Piskorski et al., 2023) constitute the only resource for persuasion technique detection for Polish and Russian at text span level, on top of which we carry out our research reported in this paper.

3 Persuasion Technique Detection

Persuasion techniques are tools and strategies used by individuals to influence others' opinions or to motivate them to undertake or support some action or adopt new behaviour(s). In order to perform our set of experiments, we exploit the persuasion techniques taxonomy from the *SemEval 2023 Shared Task 3: Detecting the Genre, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup* (Piskorski et al., 2023), which is an extended version of the the taxonomy introduced

in Da San Martino et al. (2019a,b). At the top level, there are 6 coarse-grained types of persuasion techniques:

- **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 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.

These six types are further subdivided into 23 fine-grained techniques. Figure 1 gives an overview of the two-tier taxonomy and a short definition of all fine-grained techniques.

The persuasion technique detection is a multi-class multi-label classification task. Some examples of persuasion techniques for Polish and Russian are provided in Figure 2.

ATTACK ON REPUTATION

Name Calling or Labelling: 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: attacking the opponent or an activity by associating it with a another group, activity or concept that has sharp negative connotations for the target audience.
Casting Doubt: questioning the character or personal attributes of someone or something in order to question their general credibility or quality.
Appeal to Hypocrisy: the target of the technique is attacked on its reputation by charging them with hypocrisy/inconsistency.
Questioning the Reputation: 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: justifying an idea by exhaling the pride of a group or highlighting the benefits for that specific group.
Appeal to Authority: 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: 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: a weight is given to an idea by linking it to values seen by the target audience as positive.
Appeal to Fear, Prejudice: promotes or rejects an idea through the repulsion or fear of the audience towards this idea.

DISTRACTION

Strawman: 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 replaced with a false one.
Red Herring: consists in diverting the attention of the audience from the main topic being discussed, by introducing another topic, which is irrelevant.
Whataboutism: a technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

SIMPLIFICATION

Causal Oversimplification: assuming a single cause or reason when there are actually multiple causes for an issue.
False Dilemma or No Choice: 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: 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: a brief and striking phrase, often acting like emotional appeals, that may include labeling and stereotyping.
Conversation Killer: words or phrases that discourage critical thought and meaningful discussion about a given topic.
Appeal to Time: the argument is centred around the idea that time has come for a particular action.

MANIPULATIVE WORDING

Loaded Language: 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: use of words that are deliberately not clear, vague or ambiguous so that the audience may have its own interpretations.
Exaggeration or Minimisation: consists of either representing something in an excessive manner or making something seem less important or smaller than it really is.
Repetition: the speaker uses the same phrase repeatedly with the hopes 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.

4 Experiments

We explore the performance of state-of-the-art transformer-based models for the task at hand, on the two languages of interest, namely, Polish and Russian, and the effect of cross-lingual transfer learning using multi-lingual models. Specifically, we compared the performance of mono-lingual models with the current multi-lingual model XLM-RoBERTa (Conneau et al., 2020), we measured

Ci zaś, którzy nie pamiętają PRL, mogą sobie skojarzyć styl telewizji Jacka Kurskiego z Chinami albo innymi krajami Wschodu.
Guilt by Association [POLISH]

Już nigdy nie pozwolimy, by na polskiej ziemi stanęła noga rosyjskiego żołnierza – dmie w sztandar narodowej dumy premier.
Flag Waiving [POLISH]

Według najnowszych danych agencji badawczej Inquiry, aż 47 proc. respondentów w tej grupie deklaruje, że nie będzie się szczepić. Czy naprawdę w Polsce jesteśmy gotowi ryzykować życiem i zdrowiem naszych dzieci? Appeal to fear, prejudice [POLISH]

Jak słyszeliśmy dzisiaj prezydenta Niemiec, który mówi, że Nord Stream 2 to jest formuła reparacji czy spłaty długu za okropności, jakie zostały wyrządzone przez Niemcy Rosjanom w czasie drugiej wojny światowej, muszę powiedzieć, że nabiera to nowego znaczenia. Jeśli ten projekt tak miałby być traktowany, to Niemcy są gotowe do dyskusji o reparacjach dla Polski. Strawman [POLISH]

Była zastępczyni rzecznika praw obywatelskich w rozmowie z Interią stwierdziła, że „potrzebna jest partia, która w sposób pryncypialny podejdzie do kwestii walki z katastrofą klimatyczną i bezkompromisowo do praw zwierząt". - Bez weganizmu taka perspektywa nie będzie możliwa - oceniła. False Dilemma or No Choice [POLISH]

Taka jest prawda i koniec. Conversation killer [POLISH]

Aborcja to tylko zabieg medyczny Minimisation [POLISH]

Решение суда будет иметь пугающие последствия для всей Америки Casting doubt [RUSSIAN]

Решение суда будет иметь пугающие последствия для всей Америки Loaded language [RUSSIAN]

собрало беспрецедентно широкую поддержку Exaggeration or Minimisation [RUSSIAN]

Лавров сорвал маски и выдвинул требование Appeal to Hypocrisy [RUSSIAN]

Или вы говорите, что президент Зеленский герой, или вы пропутинская марионетка False dilemma [RUSSIAN]

Отмечается, что в первые дни спецоперации люди стремились поддержать Украину, однако сейчас фокус их внимания заострен на более актуальных проблемах Obfuscation [RUSSIAN]

Figure 2: Examples of text snippets in Polish and Russian with persuasion techniques. The text fragments highlighted in yellow are the actual text spans annotated.

	TRAIN			DEVELOPMENT			TEST
lang	#docs	#spans	A_{pt}	#docs	#spans	A_{pt}	#docs
PL	145	2839	19.6	49	985	20.1	47
RU	143	3399	23.8	48	739	15.4	72
FR	158	5595	35.4	53	1586	29.9	-
EN	446	7201	16.1	90	1801	20.0	-
DE	132	4501	34.1	45	1236	27.5	-
IT	227	6027	26.6	76	1934	25.4	-

Table 1: Dataset statistics: total number of documents (#docs), total number of text spans annotated (#spans), average number of persuasion techniques per document (A_{pt}).

the effect of training with extra annotations from different languages (English, French, German, and Italian).

4.1 Experiments Settings

For monolingual models we used HerBERT (Mroczkowski et al., 2021) and RuRoBERTa[1], for Polish and Russian respectively, and for multi-lingual data we used XLM-RoBERTa (Conneau et al., 2020). We used the *large* variants for all the models from *Huggingface*. Regarding hyper-parameters, from our previous experimentation with this task on a multi-lingual setting, we found the optimal settings to be $batch\ size = 12$, $lr = 3e-3$, $weight\ decay = 0.01$, and early stopping of with a patience of 750 steps. We used the aforementioned BERT variants in a multi-label token classification configuration where we added a sigmoid layer on the output of the last layer with binary-cross entropy as loss function. This way, for each token we get 23 predictions, one per label (then aggregated to 6 in the coarse-gained setting). Each token in this setting corresponds to a subword, emitted by the model's tokenizer. Using subword-level predictions, we further aggregated them in sentences and paragraphs in post-processing for additional evaluation.

4.2 Dataset

We exploit the dataset consisting of new articles with annotated persuasion techniques for Polish and Russian from the SemEval 2023 Shared Task 3 (Piskorski et al., 2023)[2]. This dataset contains span- and paragraph-level annotations of persuasion techniques, where the latter were simply derived from the span-level annotations. We also used the data for English, German, French and Italian from the same shared task to explore how exploitation of multi-lingual data boosts the performance for the target languages. The entire dataset is subdivided into *train*, *development* and *test* dataset. The overview of the high-level statistics of all three datasets[3] is provided in Table 1.

Detailed statistics on the coarse- and fine-grained persuasion techniques for Polish and Russian for

[1] https://huggingface.co/sberbank-ai/ruRoberta-large/blob/main/README.md

[2] https://propaganda.math.unipd.it/semeval2023task3/index.html

[3] The golden labels for the test dataset are currently not publicly available, however the shared task provides a web interface to carry out evaluations on this dataset

the *training* and *development* datasets are provided in Table 2. One can observe that these datasets are highly imbalanced. *Attack on Reputation* instances account for approx. 50% of the entire dataset for both languages, where *Name Calling-Labelling* (approx. 18-27% for Polish, 7-10% for Russian) and *Doubt* (approx. 11-12% for Polish, 18-22% for Russian) are the most prevalent fine-grained techniques. The second most populated coarse-grained class is *Manipulative Wording* (ca. 19-21% and 35-37% for Polish and Russian respectively), where *Loaded Language* is the most prominent fine-grained class (approx. 11-15% and 28-29% for Polish and Russian respectively). Finally, the third most populated coarse-grained class for Polish is *Justification* (approx. 15-22%), whereas it it significantly less populated for Russian (approx. 4-5% only).

4.3 Evaluation Methodology

For the purpose of evaluating different models we use $micro$ and $macro$, $recall$ and $precision$ and F_1 measures.

Additionally, we evaluate different settings: (a) the granularity of the data after aggregating the results of the classifier: fine-grained (23 labels), coarse-grained (6 labels); and (b) the focus of the classification, i.e., at which level the labels are aggregated: paragraph level (split at new lines), sentence level (using an ad-hoc language-aware sentence splitter) and natively at subword level.

4.4 Results

Tables 3 and 4 provide overall evaluation results for all models on fine- and coarse-grained classification task at different focus levels of evaluation, i.e., subword, sentence, and paragraph level, for Polish and Russian resp. All models were trained using *train* dataset and evaluated on the *development* dataset. The XLM-RoBERTa version trained on all multilingual data (6 languages) is referred to with XLM-RoBERTa$_{multi}$.

First, we can observe that including the other languages (XLM-RoBERTa$_{multi}$), yields the highest performance boost in almost all settings, especially in terms of macro scores, and that overall results for Russian are better than for Polish. Second, the performance in both micro and macro F_1 for Polish grows with the broader focus level of the evaluation, ranging for macro F_1 from .187 (.224) to .324 (.487) for fine-grained (coarse-grained) classification, and for Russian from .190 (.267) to .306 (.464). The mono-lingual HerBERT used for Polish performs worst in almost all settings, whereas the mono-lingual Russian ruRoBERTa-based model exhibits slightly better performance vis-a-vis XLM-RoBERTa and outperforms XLM-RoBERTa$_{multi}$ only in micro F_1 at the subword level. Since this is noticeable only at this level, we speculate that it is an effect of the difference in script (latin to cyrillic).

In order to get a deeper insight into the performance of the best performing classifier, namely, XLM-RoBERTa$_{multi}$ we provide in Table 7 precision, recall, and F_1 results per each persuasion technique evaluated at sentence level for both Polish and Russian. The classes obtaining best results (i.e., F_1 measure above .3) are highlighted in bold. One can observe some that the two models perform best in the same techniques. In both models, the best performing classes are *Name Calling-Labelling* .56 (.63), *Appeal to Fear-Prejudice* .47 (.46), and *Loaded Language* .46 (.46) for Polish (Russian). We also observe that the worst performing classes are also common. i.e., for both languages *Red Herring*, *Whataboutism*, *Obfuscation-Vagueness-Confusion* obtain zero scores. We hypotheisze the poor performance is most likely due to data scarcity, something observed for most languages of the dataset. We also compared the results of XLM-RoBERTa$_{multi}$ with the models trained without transfer learning from other languages on a per-class basis. We observed that transfer learning provides a noticeable boost on low-performing classes: the count of classes not predicted at all goes down from 9 to 3 for both Polish and Russian.

For the sake of completeness, in Table 6, we present the results of the models when trained on *train* with *development* as validation and evaluated on the *test* dataset only on sentence level using the fine-grained taxonomy. Due to the imbalanced nature of the data, and the high number of under-performing classes, we focus on the macro F_1 score. Here, we can clearly see that XLM-RoBERTa$_{multi}$ also provides a noticeable boost in both cases, while the micro scores remain at the same level as in the other cases. As before, we hypothesize that this effect is due to a boost in under-represented labels where the number of annotations in the target language is very low, but the contribution of annotations from other languages is sufficient to enable the detection of those labels.

We have carried an additional experiment to sim-

	Polish				Russian			
	TRAIN		DEV		TRAIN		DEV	
technique	#num	%	#num	%	#num	%	#num	%
Attack on Reputation	1620	57.06	484	49.14	1601	47.10	341	46.14
Name Calling-Labeling	764	26.91	177	17.97	331	9.74	56	7.58
Guilt by Association	111	3.91	37	3.76	32	0.94	12	1.62
Doubt	349	12.29	111	11.27	732	21.54	133	18.00
Appeal to Hypocrisy	192	6.76	91	9.24	125	3.68	19	2.57
Questioning the Reputation	204	7.19	68	6.90	381	11.21	121	16.37
Justification	413	14.55	218	22.13	185	5.44	36	4.87
Flag Waving	97	3.42	33	3.35	50	1.47	10	1.35
Appeal to Authority	43	1.51	50	5.08	10	0.29	2	0.27
Appeal to Values	111	3.91	60	6.09	54	1.59	9	1.22
Appeal to Popularity	31	1.09	28	2.84	8	0.24	2	0.27
Appeal to Fear-Prejudice	131	4.61	47	4.77	63	1.85	13	1.76
Simplification	49	1.73	22	2.23	147	4.32	31	4.19
Causal Oversimplification	12	0.42	5	0.51	40	1.18	6	0.81
Consequential Oversimplification	25	0.88	9	0.91	76	2.24	14	1.89
False Dilemma-No Choice	12	0.42	8	0.81	31	0.91	11	1.49
Distraction	40	1.41	14	1.42	30	0.88	16	2.17
Strawman	19	0.67	3	0.30	21	0.62	11	1.49
Red Herring	12	0.42	7	0.71	2	0.06	1	0.14
Whataboutism	9	0.32	4	0.41	7	0.21	4	0.54
Calls	115	4.05	58	5.89	211	6.21	39	5.28
Slogans	42	1.48	7	0.71	84	2.47	12	1.62
Conversation Killer	58	2.04	45	4.57	91	2.68	26	3.52
Appeal to Time	15	0.53	6	0.61	36	1.06	1	0.14
Manipulative Wording	602	21.20	189	19.19	1225	36.04	276	37.35
Loaded Language	422	14.86	112	11.37	971	28.57	216	29.23
Obfuscation-Vagueness-Confusion	37	1.30	11	1.12	20	0.59	10	1.35
Exaggeration-Minimisation	128	4.51	48	4.87	149	4.38	30	4.06
Repetition	15	0.53	18	1.83	85	2.50	20	2.71
all	2839		985		3399		739	

Table 2: Dataset statistics for the fine-grained persuasion techniques for *train* and *development* datasets.

	Fine-grained classification																	
	Subword						Sentence						Paragraph					
	micro			macro			micro			macro			micro			macro		
model	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
HerBERT	.236	.089	.129	.162	.056	.083	.331	.197	.247	.212	.110	.145	.423	.306	.355	.296	.170	.216
XLM-RoBERTa	.245	.096	.138	.176	.061	.091	.341	.204	.255	.227	.108	.146	.445	.336	.383	.289	.170	.214
XLM-RoBERTa$_{multi}$	.390	.154	**.221**	.331	.130	**.187**	.502	.254	**.337**	.382	.189	**.253**	.612	.338	**.435**	.473	.246	**.324**
	Coarse-grained classification																	
	Subword						Sentence						Paragraph					
	micro			macro			micro			macro			micro			macro		
model	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
HerBERT	.348	.132	.191	.186	.076	.108	.483	.281	.355	.258	.162	.199	.613	.430	.505	.354	.243	.288
XLM-RoBERTa	.362	.141	.203	.195	.081	.115	.500	.291	.368	.291	.166	.212	.640	.464	.538	.390	.260	.312
XLM-RoBERTa$_{multi}$	.519	.207	**.296**	.469	.164	**.244**	.675	.353	**.463**	.544	.261	**.353**	.808	.471	**.595**	.709	.371	**.487**

Table 3: Evaluation results for Polish for fine- and coarse-grained classification for models trained on *train* dataset and evaluated on the *development* dataset. Best results in terms of F_1 are highlighted in bold.

	Fine-grained classification																	
	Subword						Sentence						Paragraph					
	micro			macro			micro			macro			micro			macro		
model	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
RuRoBERTa	.241	.212	**.226**	.145	.075	.099	.309	.349	.327	.161	.139	.150	.360	.403	.381	.185	.169	.176
XLM-RoBERTa	.241	.095	.136	.217	.064	.099	.323	.150	.205	.220	.084	.122	.478	.221	.302	.295	.130	.181
XLM-RoBERTa$_{multi}$	.367	.161	.223	.314	.136	**.190**	.500	.269	**.350**	.363	.196	**.254**	.569	.329	**.417**	.416	.242	**.306**
	Coarse-grained classification																	
	Subword						Sentence						Paragraph					
	micro			macro			micro			macro			micro			macro		
model	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
RuRoBERTa	.362	.310	**.334**	.228	.156	.185	.460	.486	.473	.298	.282	.290	.516	.535	.525	.346	.318	.331
XLM-RoBERTa	.390	.152	.219	.471	.124	.196	.509	.243	.329	.540	.178	.268	.645	.325	.432	.654	.256	.368
XLM-RoBERTa$_{multi}$	.498	.221	.306	.458	.188	**.267**	.663	.374	**.478**	.541	.291	**.378**	.755	.470	**.580**	.630	.367	**.464**

Table 4: Evaluation results for Russian for fine- and coarse-grained classification for the models trained on *train* dataset and evaluated on the *development* dataset. Best results in terms of F_1 are highlighted in bold.

ulate a different scenario, in which it is assumed that the text fragments that contain persuasion techniques are already identified, and the remaining task is to classify those fragments with the corresponding fine-grained persuasion technique labels. As a matter of fact, we have trained XML-RoBERTa on all *training* data in six languages and evaluated on the task of classifying whether paragraphs and sentences are persuasive or not, and achieved F_1 scores of .823 and .669 respectively when evaluated on the *development* data. This indicates that a reliable binary persuasiveness classifier can be developed. Subsequently, we trained a linear multi-label SVM classifier with 3-5 character n-grams as features using solely the text spans labelled with fine-grained persuasion techniques in Polish/Russian and exploiting the respective *training* datasets and evaluated it on the *development* datasets. The evaluation results of this experiment are provided in Figure 4.4. We can observe that such linguistically-poor models achieve, not fully unexpected, reasonable results (F_1 score) for some classes, e.g., *Name Calling-Labeling* (.85), *Loaded Language* (.51), *Conversation Killer* (.49), *Slogans* (.49) and *Flag Waving* (.40) for Polish, and *Name Calling-Labeling* (.60), *Guilt by Association* (.54), *Doubt* (.46), *Appeal to Time* (.40), *Loaded Language* (.53) for Russian. These results indicate the discriminatory potential of lexical features, as one of the areas to explore in future.

	Polish			Russian		
technique	P	R	F_1	P	R	F_1
Name Calling-Labeling	.78	.44	**.56**	.79	.52	**.63**
Guilt by Association	.38	.19	.26	.23	.15	.18
Doubt	.49	.30	**.37**	.48	.28	**.35**
Appeal to Hypocrisy	.37	.20	.26	.44	.18	.25
Questioning the Reputation	.55	.08	.14	.60	.14	.22
Flag Waving	.25	.44	**.32**	.16	.28	.20
Appeal to Authority	.42	.15	.22	.46	.19	.27
Appeal to Values	.47	.15	.22	.52	.14	.22
Appeal to Popularity	.67	.13	.21	.47	.15	.23
Appeal to Fear-Prejudice	.49	.45	**.47**	.40	.53	**.46**
Causal Oversimplification	.20	.14	.16	.34	.26	.29
Conseq. Oversimplification	.23	.06	.09	.17	.06	.09
False Dilemma-No Choice	.49	.21	.29	.47	.21	.29
Straw Man	.16	.03	.05	.15	.04	.07
Red Herring	.00	.00	.00	.00	.00	.00
Whataboutism	.00	.00	.00	.00	.00	.00
Slogans	.63	.22	**.33**	.56	.33	**.41**
Conversation Killer	.44	.08	.14	.57	.15	.23
Appeal to Time	.69	.29	**.41**	.36	.26	**.30**
Loaded Language	.55	.39	**.46**	.56	.38	**.46**
Obfusc.-Vagueness-Confusion	.00	.00	.00	.00	.00	.00
Exaggeration-Minimisation	.39	.27	**.32**	.49	.17	.25
Repetition	.16	.13	.14	.17	.10	.13

Table 5: Evaluation results per class for Polish and Russian for fine-grained classification at sentence level using XLM-RoBERTa$_{multi}$ trained on the *train* dataset and evaluated on the *development* dataset. Results with F_1 score above .3 are shown in bold.

4.5 Error Analysis

We conducted some error analysis of the XLM-RoBERTa$_{multi}$ model, trained on the *train* dataset

model	P	R	micro F_1	macro F_1
			Russian	
ruRoBERTa	.271	.175	.212	.134
XLM-RoBERTa	341	.204	**.255**	.146
XLM-RoBERTa$_{multi}$	.379	.176	.240	**.211**
			Polish	
HerBERT	.343	.219	**.267**	.156
XLM-RoBERTa	.323	.150	.205	.122
XLM-RoBERTa$_{multi}$	.392	.199	.264	**.199**

Table 6: Evaluation results on the *test* dataset at sentence level for models trained and validated on *train* and *development* datasets respectively. The best F_1 scores are highlighted in bold.

	Polish			Russian		
technique	P	R	F_1	P	R	F_1
Name Calling-Labeling	.78	.95	**.85**	.56	.64	**.60**
Guilt by Association	.30	.24	.26	.62	.48	**.54**
Doubt	.28	.39	.33	.41	.52	**.46**
Appeal to Hypocrisy	.34	.43	.38	.18	.13	.15
Questioning the Reputation	.23	.21	.22	.25	.33	.28
Flag Waving	.39	.41	**.40**	.19	.12	.15
Appeal to Authority	.35	.18	.24	.00	.00	.00
Appeal to Values	.31	.33	.32	.28	.17	.21
Appeal to Popularity	.31	.17	.22	.33	.10	.15
Appeal to Fear-Prejudice	.32	.31	.31	.36	.21	.27
Causal Oversimplification	1.00	.12	.21	.00	.00	.00
Conseq. Oversimplification	.25	.03	.05	.17	.12	.14
False Dilemma-No Choice	.50	.10	.17	.44	.26	.33
Strawman	.25	.14	.18	.15	.06	.09
Red Herring	.50	.11	.17	.00	.00	.00
Whataboutism	.00	.00	.00	.00	.00	.00
Slogans	.68	.39	**.49**	.40	.25	.31
Conversation Killer	.50	.49	**.49**	.23	.20	.21
Appeal to Time	.33	.10	.15	.61	.30	**.40**
Loaded Language	.49	.54	**.51**	.49	.58	**.53**
Obfusc.-Vagueness-Confusion	.38	.13	.19	.00	.00	.00
Exaggeration-Minimisation	.24	.17	.20	.30	.25	.27
Repetition	.27	.13	.18	.24	.19	.21
micro average	.47	.49	.48	.40	.43	.41
macro average	.39	.26	.28	.27	.21	.23
weighted average	.45	.49	.46	.38	.43	.40

Table 7: Evaluation of text-span multi-label SVM classifier for Polish and Russian trained and evaluated using *training* and *development* dataset resp. The best performing classes in terms of F_1 score (above .40) are highlighted in bold.

and evaluated on the *development* one, and noticed that some of the False Positives (FP) seemed correct. To get a better understanding, we analyzed in detail a sample of 10 random FPs for Russian, results are reported Figure 3. As we can see from the results, Recall scores are lower than Precision which indicates that the challenge of the model is the number of False Negatives.

Interestingly, we can see that around half of the False Positives are actually correct detections of persuasion techniques, and 2 of the others are arguable and have at least the coarse-grained category correct. Our intuition is that an important part of the FPs could actually be correct, however we do not measure it here precisely as it would require an important annotation effort, and this is left for future work. This is to be expected in a task with an inherently significant amount of subjectivity such as persuasion technique detection.

We further noticed that, confusion in fine-grained labels seems to happen within the same coarse-grained category (e.g. *Appeal to Hypocrisy* is confused with *Questioning the reputation*, both under *Attack on Reputation* category). This is coherent with the fact that we observed in Tables 3 4, a strong increase on most micro scores when moving from fine to coarse-grained evaluation.

5 Conclusions and Future Work

In this paper we reported on some preliminary experiments on the detection of persuasion techniques in online news in Polish and Russian, using a taxonomy of 23 persuasion techniques, and considering different evaluations scenarios: fine- versus coarse-grained classification, the text-structure level at which the labels are detected (subword, sentence, or paragraph). The comparison of mono- and multi-lingual-trained state-of-the-art transformed-based models revealed the superiority of the latter in most evaluation settings, however, given the complexity of the task, there is significant space for improvement.

In our future research we envisage to: (a) enlarge the pool of transformer-based models for inclusion in the evaluation to get a more complete picture of the phenomena observed so far, (b) explore whether and how to exploit data augmentation (Feng et al., 2021) to boost the performance of the low-populated persuasion technique classes, and (c) investigate different pre-trained models for the task, like models fine-tuned on multi-lingual

Питались они в кафе[Casting Doubt]Not Correct: and not a technique
Но болгарское правительство удивило своих граждан [CASTING DOUBT] Correct

Единый подход воспитания и образования [APPEAL TO VALUES] Not Correct

В то же время, отмечает Селиванов, в ВСУ осознают, что значительная часть населения Украины не будет поддерживать страну [FLAG WAIVING] Almost Correct: Would have been correct without the negation, otherwise it is both Casting Doubt and Appeal to Popularity

развернутая США и их союзниками пропагандистская кампания о «российской агрессии» против Украины преследует провокационные цели, тем самым поощряя власти в Киеве к саботажу Минских соглашений [CASTING DOUBT] Correct

Есть Миша Кавелашвили, который всегда был верен принципам и был бойцом [APPEAL TO VALUES] Correct

Ранее прокуратура Санкт-Петербурга направила в суд иск о признании блокады Ленинграда геноцидом [LOADED LANGUAGE] Correct

На них денег в казне вечно не хватает [QUESTIONING REPUTATION] Correct

Схожим образом высказалась премьер Новой Зеландии Джасинда Ардерн [CASTING DOUBT] Not correct: and not a technique

сделать ставку на дальнейший развал России, то есть Российской Федерации [CAUSAL OVERSIMPLIFICATION] Almost Correct: it is rather an instance of False Dilemma

Figure 3: Analysis of 10 randomly sampled examples of False Positives in Russian.

QA (Artetxe et al., 2017) or NLI (Williams et al., 2018) corpora to investigate their performance on thought coherent classes (like *Simplification* or *Distraction* families).

Limitations

The results reported in this paper are to a certain degree limited since the range of state-of-the-art mono- and multilingual models explored is by far not complete. Therefore, the main findings of the paper should be considered as of preliminary nature. We envisage to carry out more comprehensive explorations both in terms of models, architectures and languages in future. It is also important to emphasize that the underlying dataset used for the sake of carrying out the experiments exhibits some data scarcity problems, which might have led to some partially poor results, and which constitutes another aspect to be addressed in future research.

References

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.

Bohdan Andrusyak. 2019. Principle-Guided Propaganda Analysis - Case Study on Russian Military Intervention in Ukraine. https://diglib.tugraz.at/download.php?id=6144a2c5719c6&location=browse.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. 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).

Claire Bonial, Austin Blodgett, Taylor Hudson, Stephanie M. Lukin, Jeffrey Micher, Douglas Summers-Stay, Peter Sutor, and Clare Voss. 2022. The search for agreement on logical fallacy annotation of an infodemic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4430–4438, Marseille, France. European Language Resources Association.

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.

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

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, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *IJCAI*.

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

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

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.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, 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*.

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.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. 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.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. 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, Toronto, Canada.

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 *EMNLP*.

Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. 2022. Robust and explainable identification of logical fallacies in natural language arguments.

Jarosz K. Stepaniuk K. 2021. Persuasive linguistic tricks in social media marketing communication-The memetic approach. *PLoS One*, 16(7).

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. 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.