Exploring the Usability of Persuasion Techniques for Downstream Misinformation-related Classification Tasks

Nikolaos Nikolaidis, Jakub Piskorski, Nicolas Stefanovitch
Athens University of Economics and Business, Polish Academy of Sciences, EC Joint Research Centre
Athens (Greece), Warsaw (Poland), Ispra (Italy)
nnikon@aueb.gr, jpiskorski@gmail.com, nicolas.stefanovitch@ec.europa.eu

Abstract

We systematically explore the predictive power of features derived from Persuasion Techniques detected in texts, for solving different tasks of interest for media analysis; notably: detecting mis/disinformation, fake news, propaganda, partisan news and conspiracy theories. Firstly, we propose a set of meaningful features, aiming to capture the persuasiveness of a text. Secondly, we assess the discriminatory power of these features in different text classification tasks on 8 selected datasets from the literature using two metrics. We also evaluate the per-task discriminatory power of each Persuasion Technique and report on different insights. We find out that most of these features have a noticeable potential to distinguish conspiracy theories, hyperpartisan news and propaganda, while we observed mixed results in the context of fake news detection.

1. Introduction

In order to perform media landscape analysis, the use of automated tools is a prerequisite, given the vast quantity of available information on the web. As online media is increasingly being used to share citizen opinions, it is crucial for a media analyst to be equipped with the right tools to identify potentially harmful content. Such content includes, for instance, disinformation campaigns, conspiracy theories and extremely biased news texts. In this direction, we propose the use of tools able to analyze the use of persuasive patterns.

The analysis or Persuasion Techniques in text is a recent and active field (Da San Martino et al., 2020b), as several datasets have been released in recent years allowing for training of Persuasion Technique classifiers. However, a systematic study of the discriminatory power of the use of Persuasion Techniques to other, adjacent tasks has not attracted a lot of attention. Our work constitutes a preliminary study on the practical utility of detecting Persuasion Techniques to indirectly solve a range of diverse media analysis tasks, with a particular focus on the broader domain of mis-/disinformation detection. More specifically, we are interested in answering the following research questions:

- **(RQ1):** Do Persuasion Techniques detected in texts exhibit discriminatory power and for which specific binary Misinformation-related Classification Tasks?

- **(RQ2):** Which of the Persuasion Technique-derived features yields the highest discriminatory power?

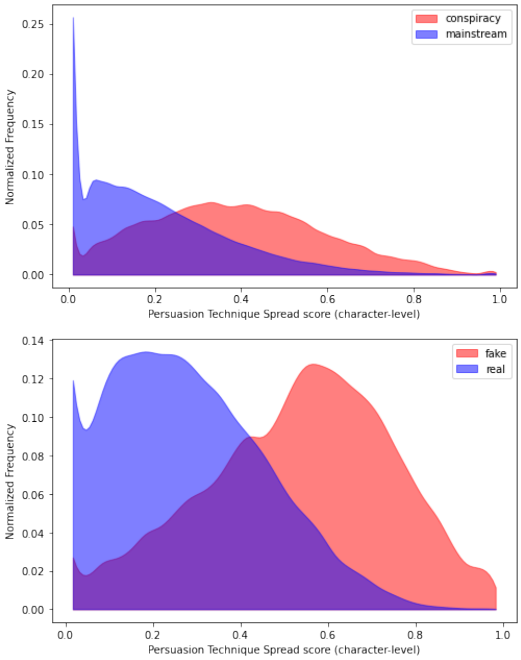

Figure 1: Normalized histograms of the spr_c persuasiveness metric, color-coded for each class, across the two label-sets of the LOCO (above) and FANG-COVID (bellow) corpora.

- **(RQ3):** What is the contribution of each Persuasion Technique to the overall discriminatory power of the specific features?

In order to tackle these research questions, we assess to which extent the simple presence of Persuasion Techniques in online news, measured by their frequency histograms and by the derived features that we introduce, is sufficient to discriminate between different types of content. As a motivating example we report in Figure 1 the histograms

LREC-COLING 2024, pages 6992–7006
20-25 May, 2024.

of Persuasion Technique density, color-coded for each class. It can be observed that each class exhibits a notable difference in shape and mode, suggesting that we could exploit Persuasion Technique information alone to discriminate between the two classes. The derived features can further be combined to provide a measure of the overall 'persuasiveness' of an article, which could assist media analysts in their understanding of the media landscape across various tasks.

More specifically, we carried out experiments and analysis using 8 datasets from the literature, all on the binary classification tasks of detection of: hyperpartisan news, fake news, propaganda and conspiracy theories. Moreover, we also performed additional experiments aimed at assessing the impact of the occurrence of named entities in the Persuasion Technique context and studying the correlation of the features among themselves.

The rest of this paper is organized as follows. First, in Section 2 we report on related work. Section 3 presents the Persuasion Technique model and the underlying taxonomy. Next, in Section 4 we introduce the datasets exploited for our study. Subsequently, in Section 5 we report on the experiments. Finally, we end up with some conclusions and future outlook in Section 6.

2. Related Work

As online misinformation continues to pose a significant societal threat, various approaches have been proposed to increase the resilience of the societies to such threats, such as debunking, prebunking (Lewandowsky and Van Der Linden, 2021) and strengthening media literacy of citizens (Guess et al., 2020; Roozenbeek et al., 2022; Commission et al., 2022). Previous work exploring the drivers behind the effectiveness of such content (Ecker et al., 2022), highlighted (among others) the significance of emotive information, cognitive fallacies and rhetorical devices.

Research on automated detection of specific Persuasion Techniques in text has been reported in Habernal et al. (2017, 2018), which focused on 5 fallacies. A more fine-grained analysis was done by Da San Martino et al. (2019), who created a corpus of English news articles labelled with 18 propaganda techniques at span and sentence level, and reported on experiments of using machine learning solutions to detect them. Yu et al. (2021) addressed the topic of interpretable propaganda detection. Detection of use of Persuasion Techniques in memes was addressed in Dimitrov et al. (2021a), whereas the relationship between propaganda and coordination was studied in Hristakieva et al. (2022a). COVID-19 related propaganda in social media was explored in 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. Andrusyak (2019) studied the use of Persuasion Techniques in the Russian news in the context of the Russian military intervention in Ukraine in 2014, and explored NLP-based models for their automated detection. 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 evaluates 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.

Various shared tasks related to the detection of Persuasion Techniques in news (Da San Martino et al., 2020a; Piskorski et al., 2023b), social media (Alam et al., 2022) and memes (Dimitrov et al., 2021b) were organized in the recent years.

The exploitation of Persuasion Technique-based features for disinformation-related task has not been studied to a very large extent. For instance, Alam et al. (2021) show the potential of using propagandistic score based on Persuasion Technique detection for the tasks of identifying messages containing factual claims, determining their check-worthiness, factuality, and potential to do harm, etc. Hristakieva et al. (2022b) study the use of propaganda-based metrics for the tasks of identifying harmful politically-oriented communities and harmless communities of grassroots activists, etc. The two aforementioned studies neither explore specific Persuasion Techniques nor their intra-document properties, which is in the main focus of our contribution. Somewhat related to our work is the study in Wiegmann et al. (2022) focusing on exploring various lexical, syntactic, semantic and stylometric features for the task of debater persuasiveness prediction, and work presented in Khalid and Srinivasan (2022) which investigates whether sensorial style features can be used to identify text genre, not strictly related to propaganda and disinformation though.

3. Persuasion Technique Detection Model

3.1. Taxonomy

For our study we exploit the two-tier Persuasion Technique taxonomy introduced in *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., 2023b). At the top level, there are 6 coarse-grained types of Persuasion Techniques: *Attack on reputation*, *Justification*, *Simplification*, *Distraction*, *Call*, and *Manipulative wording*. These six main types are

Language	micro F_1	rank	macro F_1	rank
English	.37604	#1	.18303	#2
French	.45354	#2	.35535	#1
German	.48495	#3	.31053	#1
Italian	.54319	#3	.35832	#1
Polish	.41241	#3	.28155	#1
Russian	.36948	#3	.23006	#1

Table 1: Persuasion Technique model performance on SemEval 2023 Task 3 official test dataset per language with potential rank vis-a-vis the systems' scores reported on the leaderboard.

further subdivided into 23 fine-grained techniques. Figure 2 presents the entire taxonomy organized into the six main categories, whereas the full definitions and examples are provided in Piskorski et al. (2023c) and Piskorski et al. (2023a).

3.2. Model Training and Performance

We used our Persuasion Technique multi-lingual token-level multi-label classifier, trained on the corpus presented in Piskorski et al. (2023c), capable of processing arbitrarily long text using sliding window chunking with 50% overlap. Our model's encoder, based on XML-Roberta (Conneau et al., 2019), produces 23 sigmoid outputs, one for each (subword) token of the input, and then uses a tunable threshold (by default 0.5, with a sensitive preset of 0.35) in order to output 23 labels for each token. In the post-processing stage, we aggregate using max-pooling the results to sentence-level, where we perform all of our analyses and compute our features. For the experiments conducted, we used the sensitive preset, although in preliminary experiments we did not see much variance in the results between those presets.

Our model achieves state-of-the-art results on the SemEval 2023 Task 3 test dataset, both in terms of micro and macro F_1 scores. The evaluation results for 6 languages are provided in Table 1.

4. Datasets

In order to evaluate and compare the discriminatory power of persuasion classifiers, we surveyed publicly available datasets about tasks for which Persuasion Techniques might potentially exhibit discriminatory power and also tasks in the broader area of misinformation, unreliable or harmful content detection, that we believed could potentially correlate with persuasiveness. We consulted various surveys on fact-checking (Guo et al., 2022) and fake-news datasets (D'Ulizia et al., 2021), in order to compile a list of potential datasets that we later refined. We set the following criteria for inclusion:

- Dataset should contain long-form text (thus excluding tweets/microblogs)
- Text should be available in the dataset itself, thus excluding datasets with URLs only
- Text should not be heavily distorted (e.g. noise added for copyright reasons)
- The task should depend primarily on the information contained in the text itself (and not metadata such as engagement or comments)
- The dataset should not resort to some external oracle/ground-truth (e.g. supporting evidence in some Fact-checking datasets) other than the text itself
- We prioritized datasets with gold labels (per document annotation) over those with only silver labels (per domain/source) and datasets with binary classification tasks
- We aimed at including also datasets in languages other than English

After the survey and data inspection we chose to run experiments on 8 datasets shown in Table 2. In addition, we also included a soon-to-be-released dataset with 35K articles (we used only the train set with 23k articles) each annotated using a four-level hierarchical codebook with common misinformation narratives regarding COVID-19 (Kotseva et al., 2023). The first level annotations (Super-Narrative level), includes 12 labels (such as 'Conspiracy theories, 'Vaccine-related narratives', 'Distrust towards media' and others) with the task of misinformation narrative detection and classification. As we will see later, we had to "binarize" this dataset using a selection of those labels.

The selected datasets include: binary-classification tasks related to (a) the detection of conspiracy theories (LOCO, COVID-DISINFO), (b) hyperpartisan news (HND - 2 variants, FNC bias label), (c) fake news (Spanish FNC, Kaggle FN, FakeNewsCorpus, FANG-COVID) and (d) propaganda (QProp). Their statistics and references are provided in Table 2. Regarding HND corpus we used both gold and silver versions of the corpus. Most of the corpora, except for COVID-DISINFO and QProp, are relatively balanced. In the case of FakeNewsCorpus, we used a stratified sample of 10k articles per class.

In two cases (COVID-DISINFO and FakeNewsCorpus), the annotations followed a multi-class format, and thus we had to make a mapping to binary labels. Since this is a preliminary study, we narrowed our scope to binary classification in order to provide easily understandable metrics. We intend to extend the study to multi-class and multi-label settings in future work. In the first case (COVID-DISINFO), we used the class "Conspiracy theories" as positive and the class

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

DISTRACTION
- Strawman
- Red Herring
- Whataboutism

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

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

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

CALL
- Slogans
- Conversation Killer
- Appeal to Time

Figure 2: Two-tier Persuasion Technique taxonomy, color-coded with the appropriate colors used in plots.

”Not-Apply” signaling non-misinformative articles as negative, discarding all other classes, effectively transforming it into a conspiracy detection task. In the second case (FakeNewsCorpus), we matched the labels, *political*, *satire* and *reliable* to negative and *junksci*, *bias*, *clickbait*, *conspiracy*, *fake*, *rumor*, *hate* and *unreliable* for positive classes respectively.

In our collection, COVID-DISINFO is the only multi-lingual corpus, covering mostly European languages (EN, DE, FR, IT, BG, PL, RU and others).

We carried out very minimal text preprocessing; specifically, we removed all data that contained fewer than 10 characters, and where there was a title and a main text, we concatenated the two.

The majority of surveyed datasets consist primarily of Twitter or short text such as individual claims, and most of the annotations we encountered where using distant labeling on domain level through the use of domain reputation lists. As described elsewhere (Zhou et al., 2021), given that silver-standard datasets are of lower quality and tend to yield contradicting labels due to the fact that sources tend to mix reliable and unreliable articles, we aimed to avoid using them in our study. The only dataset with exclusively distant labels was FakeNewsCorpus. Furthermore, most datasets present several biases, such as the number of sentences or the text length. As we will see later, there is a strong difference in the average number of sentences in each article per class, and thus, unexpectedly, the number of sentences has discriminatory power on its own, which must be taken into account to avoid biasing the results.

5. Experiments

We conducted a series of experiments to measure the discriminatory power of Persuasion Techniques for the selected tasks. We computed 9 different features derived from the sentence-level aggregated output of our Persuasion Technique classifier. Then, for each feature, we computed the resulting normalized frequency histogram for both classes and we measured the difference between the two histograms using two distance metrics. Please note that, as this is a preliminary study, we are not interested in building a classifier; instead we are interested in assessing the discriminatory power of individual features, based on the prevalence of Persuasion Techniques, across different tasks and datasets.

5.1. Persuasion Techniques Features

We first introduce some basic notations. Let $PT = \{t_1, t_2, \ldots, t_n\}$ denote the set of all Persuasion Techniques in the taxonomy. Let D be a text document. $|D|$ denotes the number of sentences and $|D|_{char}$ the number of characters in D. We denote with $density(D)$ the number of occurrences of Persuasion Techniques in document D. Analogously, $densityCo(D)$ denotes the proportion of Persuasion Technique instances that co-occur with other techniques in document D. Let us further denote with $sent(D)$ and $char(D)$ the set of sentences and characters in D that contain at least one instance of a Persuasion Technique. Next, let $sentPos(D)$ be the set of all sentence indexes (counting from 1 to N) in which Persuasion Techniques appear. Finally, let $div(D)$ be the number of Persuasion Technique types found in document D.

We now introduce the following six basic Persuasion Technique-based attributes, based on the intuitive assumption that: (a) the more Persuasion Technique instances in the document ($dens_p$ – density), (b) the more evenly they are spread (spr_p), (c) the more types of Persuasion Techniques are present (div_p – diversity), (d) the more instances that overlap with others there are ($comp_p$ – complexity), (e) the closer they appear to the beginning of the text (pos_p – position), and (f) the more rare techniques are used ($rarity_p$ – rarity) vis–vis the usage of Persuasion Techniques in a given reference dataset, the more likely it is the text is persuasive.

$$dens_p(D) = \min(\frac{density(D)}{|D|}, 1) \tag{1}$$

Dataset	Task	Size	Balance	Language
LOCO (Miani et al., 2021)	conspiracy theory detection	96,743	32.9%	EN
HND (Kiesel et al., 2019)	hyperpartisan news detection	1,273 gold	63%	EN
		754,000 silver	50%	EN
COVID-DISINFO	conspiracy theory	23,509	21%	MULTI
Spanish FNC (Gómez-Adorno et al., 2021)	fake news detection	971	49.4%	ES
Kaggle FN 2018 (Lifferth, 2018)	fake news detection	20,800	50%	EN
FakeNewsCorpus (Szpakowski, 2020)	fake news detection	96,000 (subsampled)	50%	EN
QProp (Barrón-Cedeño et al., 2019)	propaganda detection	51,294	11.2%	EN
(*) FANG-COVID (Mattern et al., 2021)	fake news detection	41,242	32%	DE

Table 2: The datasets characteristics. The value in the ”balance” column indicates the proportion of target class in the dataset.

$$spr_p(D) = \frac{sent(D)}{|D|} \tag{2}$$

$$div_p(D) = \min(\frac{div(D)}{q}, 1) \tag{3}$$

, where q is set to 6 based on empirical observations on the online news domain.

$$comp_p(D) = \frac{densityCo(D)}{density(D)} \tag{4}$$

$$pos_p(D) = \frac{Median(sentPos(D))}{|D|} \tag{5}$$

$$rarity_p(D) = \sum_{t \in PT} freq_t(D) * IDF(t) \tag{6}$$

where $freq_t(D)$ denotes the number of occurrences of Persuasion Technique t in document D, and $IDF(t)$ is the inverse document frequency of technique t. In order to compute the IDF scores, we used the dev set of SemEval 2023 Shared Task 3 and applied smoothing. Additionally, we define the character-level density and spread:

$$dens_c(D) = \min(\frac{density(D)}{|D|_c}, 1) \tag{7}$$

$$spr_c(D) = \frac{c(D)}{|D|_c} \tag{8}$$

On top of some of the attributes we define a compound metric $P(D)$, namely, the *persuasiveness* attribute as follows.

$$P(D) = dens_p(D) \cdot spr_p(D) \cdot div_p(D) \cdot comp_p(D) \tag{9}$$

The intuition behind this compound score is as follow: The more Persuasion Techniques used (density), the more evenly they are spread across the document (spread), the more types of Persuasion Techniques are present (diversity), and the more instances that overlap with others there are (complexity), the more likely the text is persuasive.

Additionally, we also explore two variants of each of the above introduced attributes: (a) one in which the most prevalent Persuasion Techniques, namely *Name-Calling and Loaded Language* are discarded, and (b) one in which we count Persuasion Techniques only for sentences that contain named entities[1].

Here, the main drive behind considering only sentences containing named entities is an assumption that such sentences are more likely to be more persuasive than others. For named-entity recognition we exploited the StanfordNLP tool (Manning et al., 2014).

Finally, let #$sent$ denote the number of sentences in the document $|D|$. As discussed above, many corpora present big differences in sentence count between the labels. In two cases, the positive class had, on average, almost double the sentences than the negative class and thus the number of sentences could have a discriminatory power on its own. In order to account for this bias, we use the #$sent$ as a control variable to account how much better we can discriminate from such a trivial metric. Please note that all our metrics except $rarity$, are adjusted to the length of the document measured either in sentences or characters.

5.2. Discriminatory Power Metrics

In order to gain insight into the discriminative power of the various Persuasion Technique-based features we use the $absDistance$ measure (Piskorski et al., 2008). For a given Persuasion Technique-based feature histogram h, let $\{p_h\}$ and $\{n_h\}$ denote the sequences of heights of the bars for ‘positive’ and ‘negative’ instances respectively, noting I the set of bins, the *absDistance* is then defined as follows:

$$absDistance(h) = \sum_{i \in I} |n_i^h - p_i^h|/2 \tag{10}$$

The metric has a straightforward geometric interpretation, i.e., it can be interpreted as the fraction of the total area under the histogram curves corresponding to the absolute difference between

[1] We consider only the following named entity types: *PERSON, ORGANIZATION, DATE, COUNTRY, IDEOLOGY, CITY, STATE_OR_PROVINCE, LOCATION, NATIONALITY, CRIMINAL_CHARGE, RELIGION.*

COVID DISINFO		FANG COVID		FakeNewsCoprus		HND gold		HND silver		Kaggle FN		LOCO		QProp		spanish FNC	
feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score
$dens_p$	0.503	$dens_c$	0.524	$\#sent$	0.286	div_p	0.492	spr_c	0.292	$dens_p$	0.363	$dens_c$	0.436	div_p	0.496	$dens_c$	0.314
$dens_c$	0.503	spr_c	0.512	div_p	0.109	P	0.484	$dens_p$	0.281	$dens_c$	0.361	div_p	0.419	P	0.462	$dens_p$	0.308
spr_c	0.502	$dens_p$	0.506	$dens_p$	0.106	spr_p	0.441	spr_p	0.28	spr_p	0.359	$dens_p$	0.418	$dens_c$	0.461	spr_c	0.302
spr_p	0.501	spr_p	0.487	$dens_c$	0.103	$comp_p$	0.438	$dens_c$	0.276	spr_c	0.35	spr_c	0.414	$dens_p$	0.451	spr_p	0.296
div_p	0.501	P	0.451	spr_p	0.103	$dens_p$	0.42	div_p	0.253	$\#sent$	0.31	P	0.407	$comp_p$	0.445	$\#sent$	0.273
pos_p	0.475	$\#sent$	0.244	spr_c	0.103	$dens_c$	0.395	P	0.227	P	0.204	spr_p	0.396	spr_c	0.439	pos_p	0.201
$comp_p$	0.4	div_p	0.188	pos_p	0.087	spr_c	0.372	$comp_p$	0.213	$comp_p$	0.154	$comp_p$	0.362	spr_p	0.431	$comp_p$	0.198
$\#sent$	0.372	$comp_p$	0.179	$comp_p$	0.031	$\#sent$	0.367	pos_p	0.166	div_p	0.14	pos_p	0.237	$\#sent$	0.324	div_p	0.192
P	0.259	pos_p	0.161	P	0.014	pos_p	0.26	$\#sent$	0.09	pos_p	0.129	$\#sent$	0.123	pos_p	0.282	P	0.145

Table 3: Ranking of the features with respect to their discriminatory power for the various tasks using $absDistance$ metric.

COVID DISINFO		FANG COVID		FakeNewsCoprus		HND gold		HND silver		Kaggle FN		LOCO		QProp		spanish FNC	
feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score
div_p	0.396	spr_c	0.428	$\#sent$	0.278	div_p	0.405	spr_c	0.244	$\#sent$	0.306	$dens_c$	0.366	div_p	0.415	spr_p	0.277
$dens_p$	0.38	$dens_c$	0.42	div_p	0.096	P	0.395	$dens_p$	0.24	$dens_p$	0.301	$dens_p$	0.355	$comp_p$	0.379	spr_c	0.276
spr_p	0.38	$dens_p$	0.419	$dens_p$	0.083	spr_p	0.387	$dens_c$	0.238	spr_p	0.3	div_p	0.345	$dens_c$	0.378	$dens_p$	0.271
$dens_c$	0.378	spr_p	0.407	spr_p	0.082	$dens_p$	0.374	spr_p	0.237	spr_c	0.292	spr_c	0.343	$dens_p$	0.374	$dens_c$	0.271
spr_c	0.373	P	0.37	spr_c	0.082	$comp_p$	0.358	div_p	0.209	$dens_c$	0.29	spr_p	0.335	spr_p	0.363	$\#sent$	0.248
pos_p	0.35	$\#sent$	0.2	$dens_c$	0.081	$dens_c$	0.351	P	0.208	P	0.193	P	0.327	spr_c	0.362	div_p	0.197
$comp_p$	0.317	$comp_p$	0.192	pos_p	0.072	$\#sent$	0.331	$comp_p$	0.172	pos_p	0.135	$comp_p$	0.305	P	0.356	pos_p	0.18
$\#sent$	0.291	div_p	0.152	$comp_p$	0.031	spr_c	0.328	pos_p	0.14	$comp_p$	0.125	pos_p	0.217	$\#sent$	0.293	$comp_p$	0.166
P	0.255	pos_p	0.137	P	0.025	pos_p	0.259	$\#sent$	0.086	div_p	0.124	$\#sent$	0.122	pos_p	0.269	P	0.165

Table 4: Ranking of the features with respect to their discriminatory power for the various tasks using $JensenShannon$ metric.

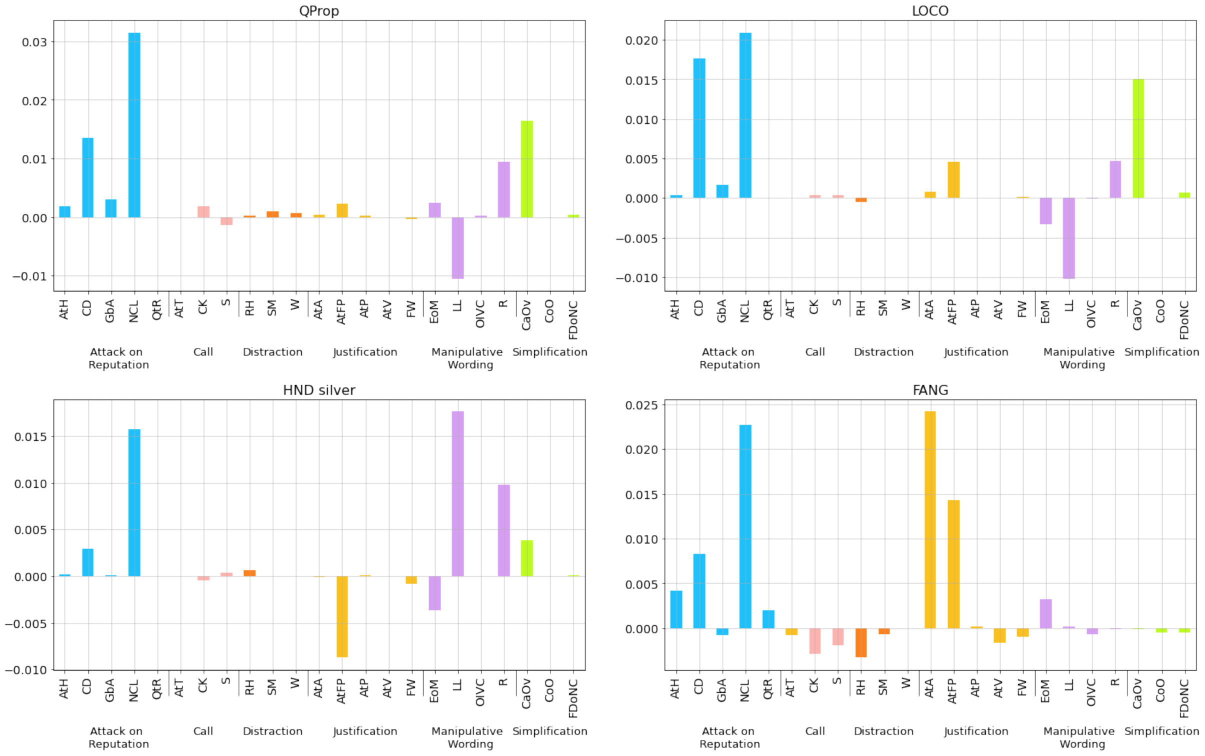

Figure 3: Results of the leave-one-out experiment on Persuasion Technique discrimination power. The higher the bar, the more negative impact the exclusion of the Persuasion Technique had. We report the scores on the $dens_p$ feature using the $absDistance$ metric. The Persuasion Technique labels are referred with their fine-grained initials as explicited in Table 7 in Figure 2 and color coded with the appropriate coarse-grained family. The coarse labels are repeated at the bottom of the axis.

them. The area under each histogram is equal to 1, hence the sum is divided by 2. The higher values of $absDistance$ indicate better discriminative power of the feature being considered.

In order to avoid bias and cross-examine the results, we add one more entropy-based metric for measuring discriminatory power, specifically, $JensenShannon$ divergence (Lin, 1991).

COVID DISINFO		FANG COVID		FakeNewsCoprus		HND gold		HND silver		Kaggle FN		LOCO		QProp		spanish FNC	
feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score
spr_p	0.483	$dens_c$	0.496	$\#sent$	0.286	div_p	0.437	spr_c	0.237	$\#sent$	0.31	$dens_c$	0.404	div_p	0.468	spr_p	0.275
$dens_p$	0.482	spr_c	0.488	spr_c	0.076	$comp_p$	0.409	$dens_p$	0.227	spr_p	0.249	div_p	0.404	$comp_p$	0.449	$\#sent$	0.273
div_p	0.482	$dens_p$	0.482	$dens_c$	0.075	$dens_p$	0.378	spr_p	0.227	$dens_p$	0.246	$dens_p$	0.393	$dens_c$	0.419	$dens_p$	0.272
spr_c	0.481	spr_p	0.464	div_p	0.067	$\#sent$	0.367	$dens_c$	0.222	spr_c	0.24	spr_c	0.392	$dens_p$	0.417	spr_c	0.26
$dens_c$	0.48	P_n	0.409	$dens_p$	0.066	spr_c	0.363	div_p	0.221	$dens_c$	0.236	spr_p	0.389	spr_c	0.415	$dens_c$	0.249
pos_p	0.447	$\#sent$	0.244	spr_p	0.066	spr_p	0.349	$comp_p$	0.187	div_p	0.146	$comp_p$	0.359	spr_p	0.414	pos_p	0.234
$\#sent$	0.376	div_p	0.191	pos_p	0.043	$dens_c$	0.333	pos_p	0.183	pos_p	0.129	pos_p	0.277	pos_p	0.338	div_p	0.204
$comp_p$	0.357	$comp_p$	0.146	$comp_p$	0.02	pos_p	0.263	$\#sent$	0.09	$comp_p$	0.098	P	0.162	$\#sent$	0.324	$comp_p$	0.107
P	0.119	pos_p	0.141	P	0.018	P	0.217	P	0.056	P	0.097	$\#sent$	0.123	P	0.155	P	0.098

Table 5: Ranking of the features when discarding the two most frequent Persuasion Techniques (*Loaded Language*, *Name Calling or Labeling* for the various tasks using $absDistance$ metric.

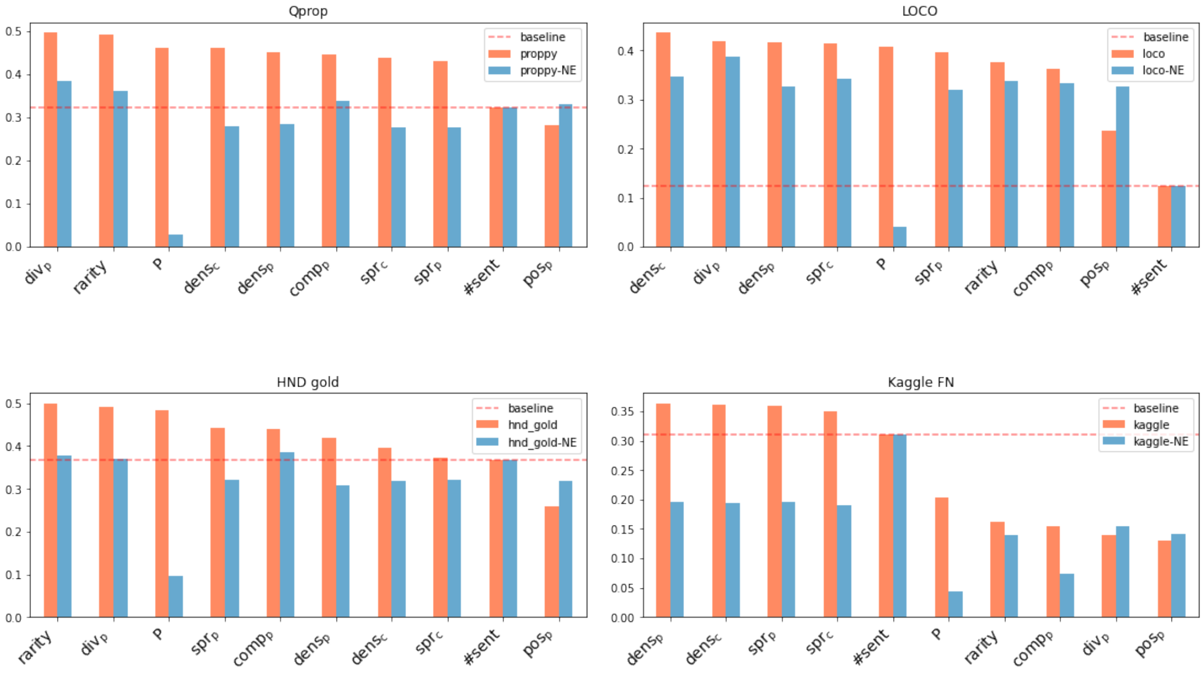

Figure 4: Comparison of discriminatory power when computing features considering the entire text of the documents versus only considering sentences with at least one relevant Named-Entities type.

5.3. Basic Features

Table 3 and 4 provides the ranking of all basic features with respect to their discriminatory power for the various tasks using the $absDistance$ metric and $JensenShannon$ respectively, and number of bins set to 20. We picked this number after several empirical observations, as it was large enough to avoid aliasing. We did not notice important variations with different bin-sizes.

First, we see that in the case of FakeNews-Corpus, our features fail to discriminate the binary mapping better than the baseline ($\#sent$). In Conspiracy theory detection tasks (LOCO and COVID DISINFO) we see that the discriminatory power of most features is noticeably above the baseline. Similarly, for the tasks of Hyperpartisan News detection and Propaganda Detection (HND and QProp respectively) almost all the metric exhibit discriminatory power with important margins. These patterns are evident in both discriminatory power metrics ($absDistance$ and $JensenShannon$). In the tasks of FakeNews detection (without ground-truth) such as Kaggle FN challenge, Spanish FNC, FakeNews corpus and FANG, the results are mixed. In the case of FANG COVID, the discrimination is quite evident (double that of the baseline), however in all other cases the gain is either marginal (Spanish FNC), worse than the baseline (FakeNewsCorpus) or inconclusive, as in the case of Kaggle, where the two metrics contradict each other (with the highest score still marginally above control variable $\#sent$).

In general we see that $dens_p$, spr_p and div_p, present the three most stable metrics across corpora w.r.t. discriminatory power, and this also appears to be somewhat natural, i.e., the more techniques are used, the more diverse they are, and the more evenly they are spread, the more likely the text is manipulative. On the other end, $comp_p$, and, in particular, pos_p do exhibit lower discriminatory power across corpora, scoring lower or similar to the baseline ($\#sent$).

Regarding the more mixed and inconclusive results on the discriminatory power of the features for the task of detecting fake news, we hypothesize that Persuasion Techniques do not necessarily play a key role in the process of determining the factuality of a given piece of text, but rather as an instrument to spot text fragments potentially worth further investigating. Furthermore, the discrepancies in the discriminatory power for the various fake news-related corpora explored in our study might have potentially resulted from a bias of how the various corpora have been created, which would require further study to assess.

5.4. Features with NER Filtering

In Figure 8, we present a comparison of the discriminatory power of all features vis-a-vis their corresponding versions that apply NER filtering (by considering only sentences that contain at least one occurrence of a named entity in the selected types), for 4 datasets. We can observe that for all 4 datasets consistently, applying NER filtering deteriorates the discriminatory power of all features. This is validated by both discriminatory power metrics. Thus, we conclude that the task-relevant persuasive content is present also in sentences without specific Named Entities. We also conducted a variation of the above filtering, where for each sentence, we also include the preceding and the following sentences. The results of this variant are presented in Annex A.4

5.5. Individual Persuasion Technique Contribution

Table 5 provides the discrimination scores if we exclude the two most frequent Persuasion Techniques (*Loaded Language* and *Name Calling or Labeling*.) We can clearly see that the discriminatory power deteriorates across all the datasets. This is not an unexpected finding, since the two Persuasion Techniques are the most frequent and classified very reliably. Additionally, we wanted to capture the contribution of each Persuasion Technique in the discriminatory power of the features. In order to establish this, we performed the following experiment, we calculated how much the exclusion of each of the 23 Persuasion Techniques affects the overall discrimination metrics for each feature for 4 selected corpora. In Figure 3, we report the results, on four datasets (QProp, LOCO, HND, FANG-COVID), using the *absDist* metric on the $dens_p$ feature. The results on the other metrics follow the same general patterns. *Name Calling or Labeling* has a high impact and, surprisingly, the very similar technique *Loaded Language* seems to contribute differently for each task. *Causal Oversimplification* and Casting Doubt seem to contribute positively in all cases as well with *Repetition* having positive or neutral impact. Interestingly, in the case of FANG corpus, *Appeal to Authority* and *Appeal to Fear-Prejudice* seem to play a significant role. We believe that this is the case because FANG dataset contains articles regarding COVID-19 where both fear and the presence of (often fake) experts were very noticeable in the relevant discourse around the disease. Thus, we hypothesize that when it comes to the utility of each Persuasion Technique class there, there is a different profile for each task. It should be noted that these results can be influenced by the classifier's performance on each individual class. In this case however, it is interesting to observe that the techniques with noticeable impact seem to be the ones for which the classifier tends to obtain better performance results (e.g *Name-Calling or Labeling*) and also we can see that for such techniques the impact can be opposite across datasets (e.g. Loaded Language). Thus, we can deduce that, indeed, in the context of our classifier (Piskorski et al., 2023c), discrimination performance can be considered conditioned on individual techniques.

5.6. Correlation of features

Lastly, we wanted to investigate the correlation of the different computed features. We measured $Pearson$ correlation of all the feature pairs for all the corpora. In Figure 5 we present the heat map for feature correlation computed for LOCO dataset. We can see that, as expected, sentence and character features correlate and there is a strong correlation between spr_p and $dens_p$ features and relatively low correlation with pos_p. The heatmaps for feature correlation for other datasets are provided in Figure 6 in Annex A.2. What is interesting is the difference in correlation of pos_p with the other features across different datasets (ranging from 0.11 up to 0.67), suggesting a different usage patterns on different datasets.

6. Conclusions

In this paper, we presented a systematic exploration of Persuasion Technique-derived features and whether they yield sufficient discriminatory power for online media monitoring tasks related to misinformation detection.

The main findings of this paper can be summarized as follows:

Regarding **(RQ1)**, we conclude that Persuasion Technique-derived features exhibit noticeable discriminatory power in the tasks related to Conspiracy theory detection, Hyperpartisan news detection and Propaganda detection. In the case of Fake News detection, the results are inconclusive, exhibiting discriminatory power in the context of COVID-19 misinfomation (FANG) but failing to reliably discriminate in the case of

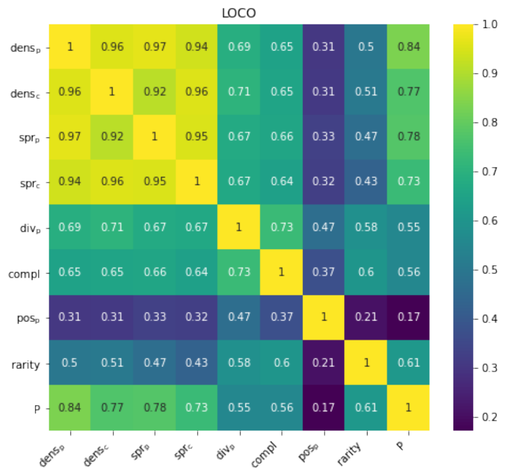

Figure 5: Pearson correlation heat map for the different computed features on the LOCO corpus.

more generic datasets (FakeNewsCorpus, Kaggle, Spanish FNC).

Regarding **(RQ2)**, we conclude that Persuasion Technique diversity (div_p), density ($dens_p$), and spread (spr_p) and their character level counterparts are the most reliable attributes across corpora. Still, there is some variation from corpus to corpus. Also, the relatively large difference in correlation of pos_p feature with frequency-derived features (div_p, $dens_p$, spr_p) across datasets, hints at the potential of different usage patterns of Persuasion Techniques.

As regards **(RQ3)**, we conclude that the most frequent techniques (*'Loaded Language'* and *'Name Calling or Labeling'*) contribute noticeably to the discriminatory power. Furthermore, we can conclude that the exact contribution and specific histogram profile depends on the given task and potentially also on the corpus. We provide some preliminary insights on such contribution, when using the results of our classifier results as proxy for Persuasion Technique frequency.

We believe the reported findings constitute valuable material for researchers working on mis/disinformation, propaganda and conspiracy theory detection. Considering ways to extend the presented work, we envisage to: (a) extend this analysis to the case of multi-class and multi-label tasks; (b) compare the performance of a baseline model (e.g., logistic regressor, SVM) trained using the presented features vis-a-vis state-of-the-art models for selected tasks; (c) explore how the presented features could be used in a ranking setting (e.g., evaluate them using *precision@top-k*) for each task; (d) study how to combine the features with modern text classification models in order to measure how their performance is affected; and lastly (e) apply conformal prediction methodology (Angelopoulos and Bates, 2022) using the derived features.

7. Acknowledgements

The work presented in this paper is financed by the Joint Research Centre of the European Commission.

Limitations The significance of the findings presented in this paper is limited in various ways. First, the Persuasion Technique detection model, although performing strong on the test data of the SemEval 2023 Task 3, does not capture all techniques equally well, and was trained using data potentially containing some biases according to (Piskorski et al., 2023c).

Furthermore, although we exploited 9 different corpora for the experiments, the pool of potential resources in this regard has by far not been exhausted. The quality of the datasets used is also another point of cautiousness (Zhou et al., 2021) as they might not fully reflect data encountered in real-world scenarios and have biases on their own.

In the presented work we have only studied binary classification tasks, while in the domain of disinformation and fact-checking there are interesting multi-class tasks, which were not in the scope of this study. The study on NER filtering was limited as there are also settings which we did not explore, e.g., inclusion of broader context.

When assessing the individual Persuasion Technique contribution to the discriminatory power of the given tasks, the reader should consider that these results can be partially influenced by the different classification performance on each individual technique.

Ethics Statement The Persuasion Technique detection and classification model exploited and the reported findings could be of use for media analyst practitioners, however, they could also be used for some adversarial activities. Hence, we recommend to exploit the results and findings presented in this paper in a cautious manner.

It is important to note that the use of deep learning-based language models might require a lot of computational power and the use of GPUs/TPUs for training purposes, which for very large models contributes to global warming (Strubell et al., 2019). For the sake of carrying out the presented research we have only trained a single model of 0.5B parameters that was then used for making predictions for all corpora.

8. Bibliographical 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.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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`.

Anastasios N. Angelopoulos and Stephen Bates. 2022. A gentle introduction to conformal prediction and distribution-free uncertainty quantification.

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

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.

European Commission, Sport Directorate-General for Education, Youth, and Culture. 2022. *Final report of the Commission expert group on tackling disinformation and promoting digital literacy through education and training – Final report*. Publications Office of the European Union.

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, 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, Yu Seunghak, Roberto Di Pietro, Preslav Nakov, et al. 2020b. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,*{*IJCAI-20*}, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. 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.

Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. 2021. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518.

Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. ArXiv:2108.11896 [cs].

Helena Gómez-Adorno, Juan Pablo Posadas-Durán, Gemma Bel Enguix, and Claudia Porto Capetillo. 2021. Overview of fakedes at iberlef 2021: Fake news detection in spanish shared task. *Procesamiento del Lenguaje Natural*, 67(0):223–231.

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. 2022a. 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.

Kristina Hristakieva, Stefano Cresci, Giovanni Da San Martino, Mauro Conti, and Preslav Nakov. 2022b. The spread of propaganda by coordinated communities on social media. In *Proceedings of the 14th ACM Web Science Conference 2022*, WebSci '22, page 191–201, New York, NY, USA. Association for Computing Machinery.

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.

Osama Khalid and Padmini Srinivasan. 2022. Smells like teen spirit: An exploration of sensorial style in literary genres. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 55–64, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Bonka Kotseva, Irene Vianini, Nikolaos Nikolaidis, Nicolò Faggiani, Kristina Potapova, Caroline Gasparro, Yaniv Steiner, Jessica Scornavacche, Guillaume Jacquet, Vlad Dragu, et al. 2023. Trend analysis of covid-19 mis/disinformation narratives–a 3-year study. *Plos one*, 18(11):e0291423.

Stephan Lewandowsky and Sander Van Der Linden. 2021. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, 32(2):348–384.

William Lifferth. 2018. Fake news. `https://kaggle.com/competitions/fake-news`.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Justus Mattern, Yu Qiao, Elma Kerz, Daniel Wiechmann, and Markus Strohmaier. 2021. FANG-COVID: A new large-scale benchmark dataset for fake news detection in German. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 78–91, Dominican Republic. Association for Computational Linguistics.

Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2021. Loco: The 88-million-word language of conspiracy corpus. *Behavior Research Methods*, 54.

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, 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, Marcin Sydow, and Dawid Weiss. 2008. Exploring linguistic features for web spam detection: a preliminary study. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*, AIRWeb '08, page 25–28, New York, NY, USA. Association for Computing Machinery.

Jon Roozenbeek, Sander Van Der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34):eabo6254.

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.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *CoRR*, abs/1906.02243.

Maciej Szpakowski. 2020. Fake news corpus. `https://github.com/several27/FakeNewsCorpus`.

Matti Wiegmann, Khalid Al Khatib, Vishal Khanna, and Benno Stein. 2022. Analyzing persuasion strategies of debaters on social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6897–6905, Gyeongju, Republic of Korea. International Committee on 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.

Xiang Zhou, Heba Elfardy, Christos Christodoulopoulos, Thomas Butler, and Mohit Bansal. 2021. Hidden biases in unreliable news detection datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2482–2492, Online. Association for Computational Linguistics.

A. Appendix

A.1. NER-filtered results

We re-run the histogram and discriminatory power computations limiting the computation to considering only sentences that include instances of one of the following Named entity types: *PERSON, ORGANIZATION, DATE, COUNTRY, IDEOLOGY, CITY, STATE_OR_PROVINCE, LOCATION, NATIONALITY, CRIMINAL_CHARGE, RELIGION.*. The results are provided in Table 6.

A.2. Correlation between features

We calculated the Pearson correlation for all pairs of features. We report the results for all 9 datasets in Figure 6.

A.3. Individual Persuasion Technique contribution

We performed a leave-one-out test for each Persuasion Technique label and measured the impact on the discriminatory power. Figure 7 presents the plots for spr_p feature. The results look very similar to Figure 3 for the majority of the Persuasion Techniques.

HND gold				Kaggle				LOCO				QProp			
full		NER only		full		NER only		full		NER only		full		NER only	
feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score	feat	score
div_p	0.492	$comp_p$	0.385	$dens_p$	0.363	$\#sent$	0.31	$dens_c$	0.436	div_p	0.388	div_p	0.496	div_p	0.384
P	0.484	div_p	0.371	$dens_c$	0.361	$dens_p$	0.197	div_p	0.419	$dens_c$	0.346	P	0.462	$comp_p$	0.338
spr_p	0.441	$\#sent$	0.367	spr_p	0.359	spr_p	0.197	$dens_p$	0.418	spr_c	0.342	$dens_c$	0.461	pos_p	0.33
$comp_p$	0.438	spr_p	0.321	spr_c	0.35	$dens_c$	0.194	spr_c	0.414	$comp_p$	0.333	$dens_p$	0.451	$\#sent$	0.324
$dens_p$	0.42	spr_c	0.32	$\#sent$	0.31	spr_c	0.189	P	0.407	$dens_p$	0.327	$comp_p$	0.445	$dens_p$	0.284
$dens_c$	0.395	$dens_c$	0.319	P	0.204	div_p	0.155	spr_p	0.396	pos_p	0.326	spr_c	0.439	$dens_c$	0.28
spr_c	0.372	pos_p	0.319	$comp_p$	0.154	pos_p	0.141	$comp_p$	0.362	spr_p	0.319	spr_p	0.431	spr_p	0.276
$\#sent$	0.367	$dens_p$	0.307	div_p	0.14	$comp_p$	0.075	pos_p	0.237	$\#sent$	0.123	$\#sent$	0.324	spr_c	0.276
pos_p	0.26	P	0.097	pos_p	0.129	P	0.044	$\#sent$	0.123	P	0.041	pos_p	0.282	P	0.028

Table 6: Ranking of the features with respect their discriminatory power for the various tasks. Side-by-side comparison of full data (full) versus the subsets containing at least one relevant Named-Entity (NER only). The results were computed using the $absDistance$ metric.

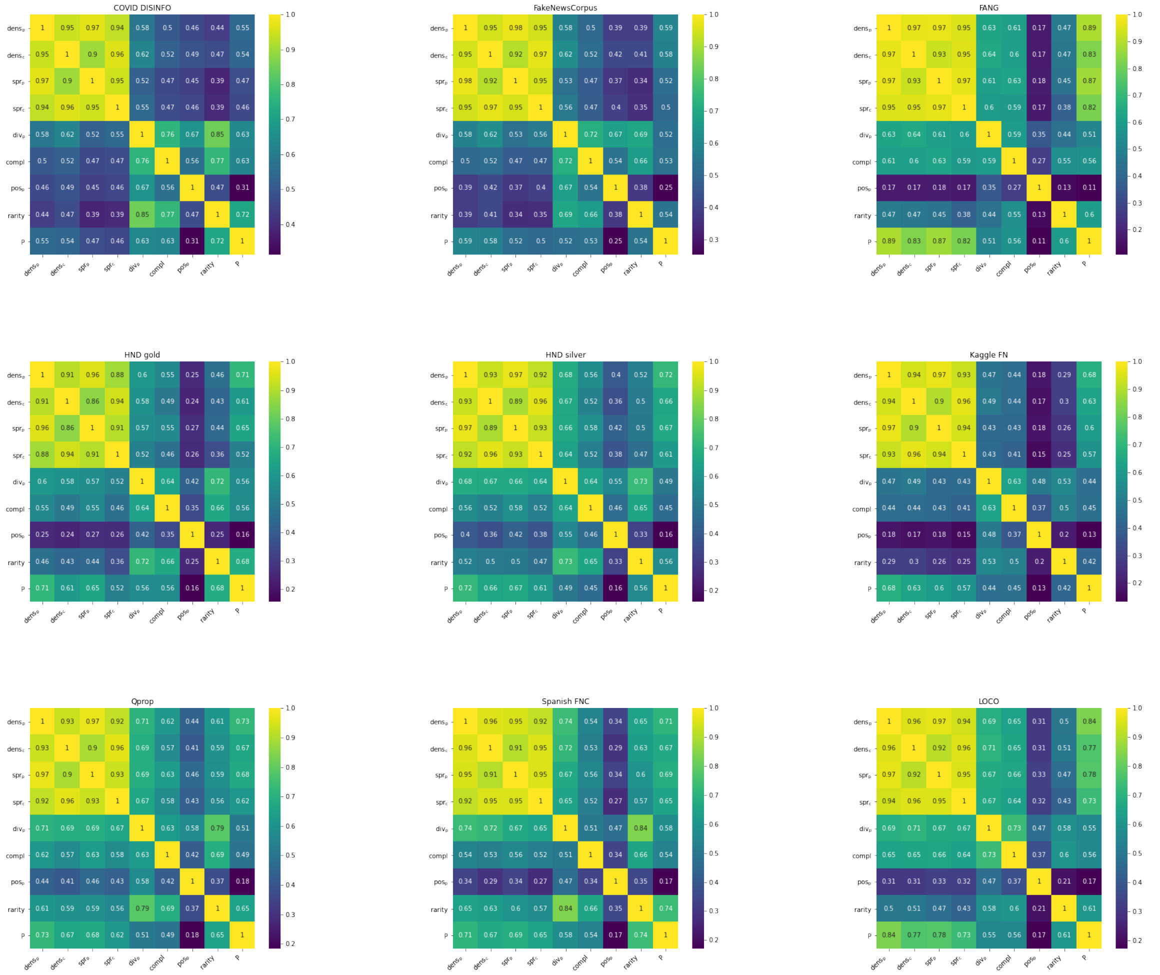

Figure 6: The Pearson correlation between the computed features for the rest of the datasets.

A.4. NER filtering with adjacent sentences

We performed a variation of the NER filtering as described in A.1 where we include also the neighbouring sentences (preceding and following) to the metrics calculation.

A.5. Persuasion Technique classifier performance per class

For the sake of completeness we provide in Figure 7 model performance figures for the detection of each specific Persuasion Technique type taken from the Semeval 2023 task 3 description paper (Piskorski et al., 2023b). However, these

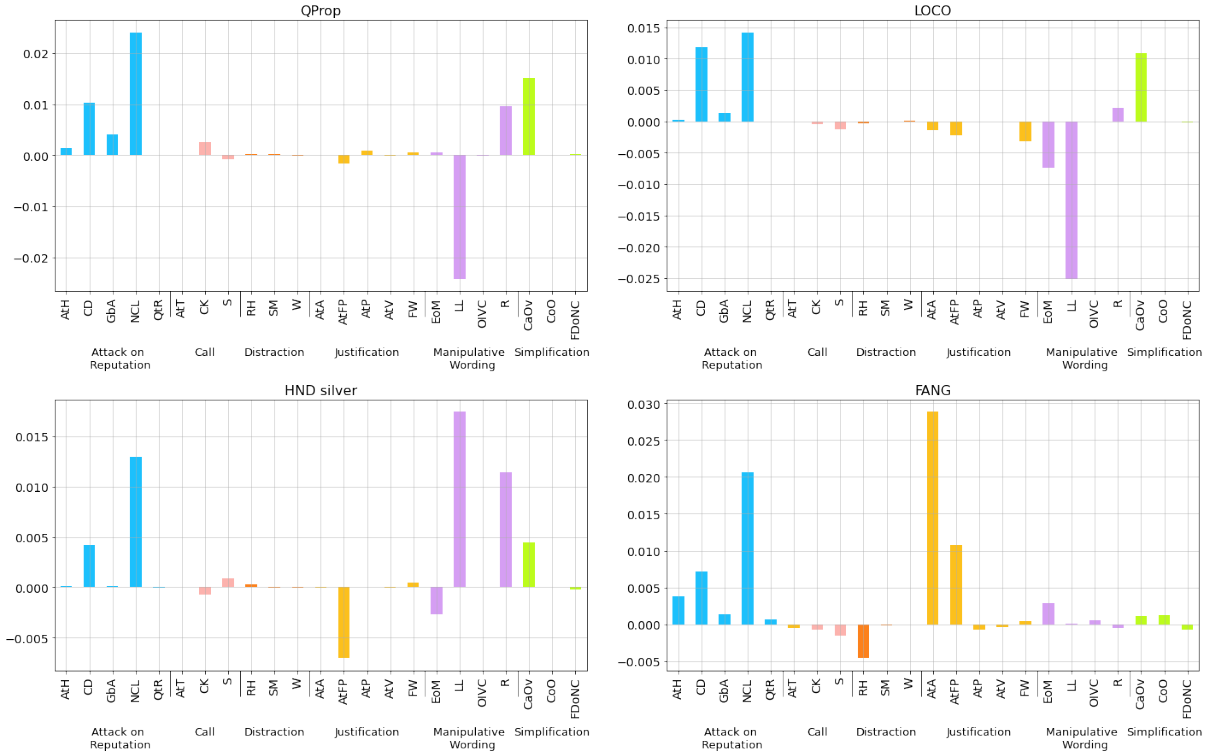

Figure 7: Results of the leave-one-out experiment on Persuasion Technique discrimination power. The higher the bar, the more negative impact the exclusion of the Persuasion Technique had. We report the scores on the $sprp_p$ feature using the $absDistance$ metric.

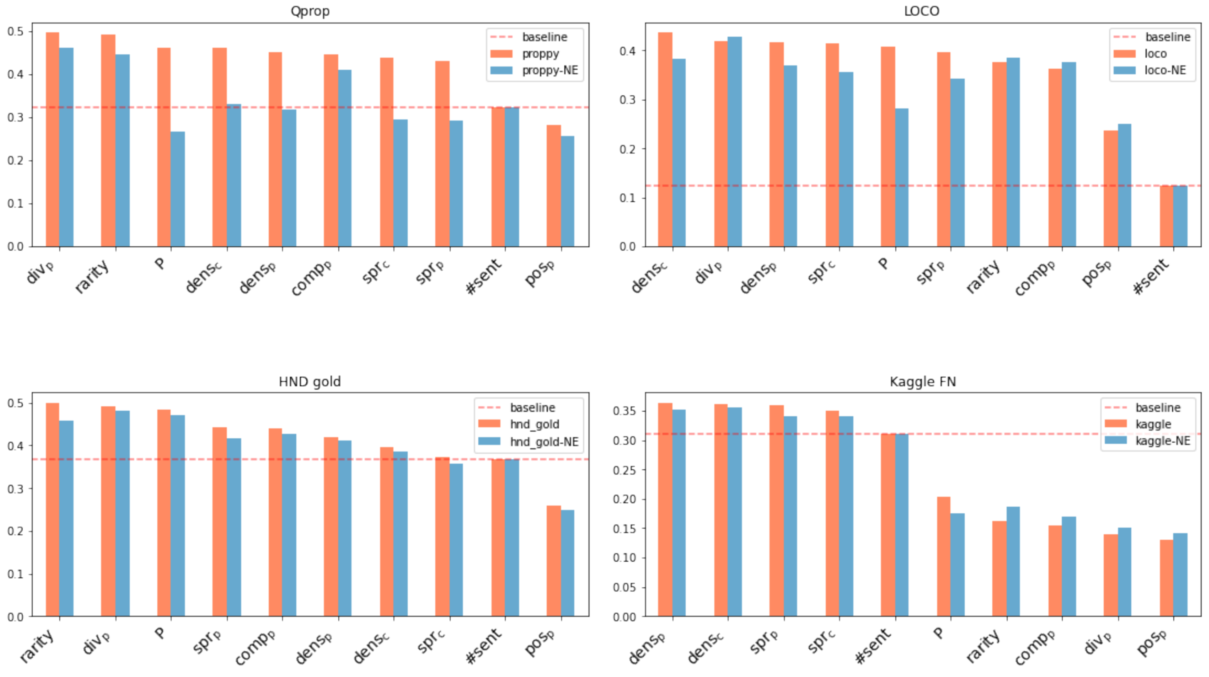

Figure 8: Comparison of discriminatory power when computing features considering the entire text of the documents versus only considering the sentences with at least one relevant Named-Entities type, as well as their neighbours (one preceding and one following sentence).

figures should be considered as indicative only, given a slightly different evaluation set-up.

Technique	Abbrev.	Prec.	Rec.	F1	Support	%
Attack on Reputation		.418	.316	.357	14,814	39.8
Name Calling-Labeling	NCL	.633	.444	.522	5,935	15.9
Guilt by Association	GbA	.449	.273	.339	679	1.8
Casting Doubt	CD	.404	.308	.349	4,922	13.2
Appeal to Hypocrisy	AtH	.277	.316	.295	1,013	2.7
Questioning the Reputation	QtR	.326	.241	.277	2,265	6.1
Justification		.389	.25	.298	4,461	12.0
Flag Waving	FW	.41	.321	.36	772	2.1
Appeal to Authority	AtA	.336	.19	.242	796	2.1
Appeal to Popularity	AtP	.373	.145	.209	378	1.0
Appeal to Values	AtV	.443	.232	.305	728	2.0
Appeal to Fear-Prejudice	AtFP	.384	.36	.371	1,787	4.8
Distraction		.106	.043	.046	837	2.2
Straw Man	SM	.068	.095	.079	414	1.1
Red Herring	RH	.0	.0	.0	253	0.7
Whataboutism	W	.25	.034	.06	170	0.5
Simplification		.293	.176	.211	1,625	4.4
Causal Oversimplification	CaOv	.157	.179	.167	685	1.8
False Dilemma or No Choice	FDoNC	.317	.2	.245	543	1.5
Consequential Oversimplification	CoO	.406	.15	.219	397	1.1
Call		.383	.243	.295	2,004	5.4
Slogans	S	.43	.314	.363	794	2.1
Conversation Killer	CK	.271	.181	.217	1,040	2.8
Appeal to Time	AtT	.448	.232	.306	170	0.5
Manipulative Wording		.302	.168	.204	13,502	36.3
Loaded Language	LL	.596	.423	.495	9,857	26.5
Obfuscation Intentional Vagueness-Confusion	OIVC	.133	.015	.026	440	1.2
Exaggeration or Minimisation	EoM	.246	.181	.209	1916	5.1
Repetition	R	.233	.052	.085	1,289	3.5
Total					**37,243**	**100**

Table 7: Statistics about the token-level performance on fine-grained Persuasion Technique detection, evaluated on the Semeval 2023 task 3 (Piskorski et al., 2023b) test dataset. As the granularity of the results is token-level, the scores are noticeably lower than for sentence-level aggregation configuration that is used in this paper. Taken from Piskorski et al. (2023c)