Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification

Géraud Faye, Benjamin Icard, Morgane Casanova, Julien Chanson, François Maine, François Bancilhon, Guillaume Gadek, Guillaume Gravier, Paul Égré

Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification

Géraud Faye[1,2], Benjamin Icard[3,4], Morgane Casanova[5], Julien Chanson[6], François Maine[4,7], François Bancilhon[8], Guillaume Gadek[1], Guillaume Gravier[5] and Paul Égré[3]

[1]*Airbus Defence and Space*, France
[2]*Université Paris-Saclay, CentraleSupélec*, MICS, France
[3]*Institut Jean-Nicod*, CNRS, ENS-PSL, EHESS, France
[4]*LIP6*, CNRS, Sorbonne Université, France
[5]*Université de Rennes, CNRS, Inria*, IRISA, France
[6]*Mondeca*, France
[7]*Freedom Partners*, France
[8]*Observatoire des Médias*, France

Abstract

This paper investigates the language of propaganda and its stylistic features. It presents the PPN dataset, standing for Propagandist Pseudo-News, a multisource, multilingual, multimodal dataset composed of news articles extracted from websites identified as propaganda sources by expert agencies. A limited sample from this set was randomly mixed with papers from the regular French press, and their URL masked, to conduct an annotation-experiment by humans, using 11 distinct labels. The results show that human annotators were able to reliably discriminate between the two types of press across each of the labels. We propose different NLP techniques to identify the cues used by the annotators, and to compare them with machine classification. They include the analyzer VAGO to measure discourse vagueness and subjectivity, a TF-IDF to serve as a baseline, and four different classifiers: two RoBERTa-based models, CATS using syntax, and one XGBoost combining syntactic and semantic features.

1 Introduction

In times of warfare as well as in authoritarian regimes, state propaganda is an informational weapon whose aim is to damage the opponents' reputation and to maintain trust in the state's actions (Jowett and O'Donnell, 2019). With the development of the internet and social networks, propaganda has new media to sprawl and to cross borders (Da San Martino et al., 2020a). Current trends on news consumption show an increase in the number of people getting informed on digital device.[1] Internet platforms are a new playground for propagandists, where they can disseminate partisan pieces among news articles and opinions shared on social media.

The rhetorical techniques of propagandists differ and their detection is currently a topic of interest (Da San Martino et al., 2020b; Quaranto and Stanley, 2021). In this paper, we pursue this general line of analysis, by examining the language of propaganda and its stylistic features. More specifically, we propose a comparison between human classification and machine classification of propaganda.

We present the PPN dataset, standing for Propagandist Pseudo-News, a multisource, multilingual, multimodal dataset composed of news articles extracted from websites identified as propaganda sources by Newsguard and Viginum, a French state-backed misinformation and foreign interference surveillance organisation. Composition of the dataset is detailed in Section 2.

To analyse the corpus and deepen our understanding of the language of propaganda, we also conducted a multilabel annotation experiment involving randomly mixing articles from that corpus with a sample of articles from mainstream French newspapers. The experiment is detailed in Section 3, and the results are presented in Section 4, showing that regular press articles and articles from the corpus are recognizably different to annotators, despite sharing topics.

To find the cues characteristic of each corpus, we then used different techniques. In Section 5, we use the expert system VAGO to check on the occurrence of subjective and vagueness markers in either type of corpus, since intentional vagueness (Égré and Icard, 2018) is among recognized techniques of propaganda (Da San Martino et al., 2020b) and its higher prevalence detectable in fake

[1] https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/

news (Guélorget et al., 2021). Then in Section 6, we train machine learning models to detect articles from propagandist sources, three based on text processing and one on stylistic and syntactic features. Explainability capabilities of the models are used to confirm the features learnt by the models and to discuss ways in which they can be improved.

2 The PPN dataset

The proposed PPN dataset is diverse in terms of sources, topics and used languages. The corpus has been extracted from 5 sources (news distribution by source is shown in Table 1), all of which were created after the Russian invasion of Ukraine on February 24, 2022:

- rrn.media: *Reliable Recent News* (previously named *Reliable Russian News*) has the form of a news website publishing articles containing a pro-Russia or anti-Occident stance. The website contains news in 9 languages (Arabic, Chinese, English, French, German, Italian, Russian, Spanish and Ukrainian), which receive a different coverage over time.

- tribunalukraine.info: this website aims at accusing Ukraine of committing war crimes and financially benefiting from the conflict. The writing style is more aggressive than *rrn*, as it aims at damaging Ukraine's reputation. All articles from this source are available in English, French, German, Russian and Spanish.

- waronfakes.com: the counterpart of tribunalukraine, it aims at denying Russian war crimes allegations. It does not publish news articles, but short summaries of allegations, and as such it qualifies as fake news. All *"debunked"* facts are available in Arabic, Chinese, English, French, German and Spanish.

- notrepays.today and lavirgule.news: these French-writing websites publish polarizing news with the aim of damaging trust in Western institutions. Contrarily to the first three sources, which were created at the beginning of the Russian invasion, *notrepays* and *lavirgule* were created one year later, with a related agenda.

Unlike some previous publications (Heppell et al., 2023), we present the propaganda articles in their original language for analysis, but knowing that several of the sites present translations

Source	Number of documents
rrn	12,427
tribunalukraine	4,975
waronfakes	344
notrepays	480
lavirgule	503

Table 1: PPN articles distribution by source.

Language	Number of documents
Arabic	1,079
Chinese	794
English	3,219
French	4,141
German	3,341
Italian	1,796
Russian	1,435
Spanish	2,485
Ukrainian	439

Table 2: PPN articles distribution by language.

in different languages. We share the collected dataset on the following GitHub repository: https://github.com/hybrinfox/ppn. The distribution of articles by languages is shown in Table 2.

3 Annotated corpus and labels

To understand how propaganda can be perceived and its characteristics, we conducted an annotation experiment on a subset of the French PPN dataset. In order to balance the dataset, we added articles from five French national newspapers of different political orientations, namely lefigaro.fr, lemonde.fr, marianne.fr, liberation.fr and mediapart.fr. The articles were randomly selected among those sources. They had to be published after the beginning of the Ukraine invasion (February 24, 2022) and to contain at least the mention of Russia or Ukraine. An additional filter, based on article length, was applied to limit bias linked to the length of articles. All annotated articles contained between 1,000 and 10,000 characters (shorter articles belong almost exclusively to the propaganda class and longer articles always belong to the regular class). A total of 48 articles were selected for each type of press, with a maximum of 14 and a minimum of 7 articles by source in the alternative press, and a maximum of 15 vs. a minimum of 1 by source in the regular press, and roughly similar distributions across the two types.

Eleven labels were used for the annotations. Figure 1 presents them in the order in which annotators had to mark them, with a summary of their definition. The 11 labels included 5 labels targeting manipulative content proscribed by the deontology

- **Vague:** the information contained in the article is general with few details or specific facts.
- **Subjective:** the article essentially presents opinions and the explicit or implicit subjective viewpoint of its author.
- **Exaggeration:** the article presents information in an exaggerated or excessive manner.
- **Pejorative:** the article primarily aims to vilify individuals or institutions.
- **Descriptive:** the article essentially reports facts or events rather than opinions.
- **Propaganda:** the article gives a biased presentation of the situation and seems to serve above all the interests of a state or organization.
- **Satirical:** the article is intended to make people laugh and is written in a joking tone.
- **Dishonest Title:** the title reports false or artificially inflated information.
- **Adequate Sources:** the article cites its sources sufficiently and accurately.
- **Fake News:** in your opinion, the article deserves to be called "fake news".
- **False Information:** the article contains at least one false information.

Figure 1: Description of the 11 labels used for the annotation task.

of journalism[2] and the Gricean norms of cooperative discourse (Quality in particular, Grice 1975), namely "Dishonest Title", "Fake News", "False Information", "Exaggeration", and "Propaganda". We also included 2 labels "Satirical" and "Pejorative", targeting jocular and adversarial intention; and finally, 4 labels for features susceptible to be applicable to either type of press, with 2 labels targeting the expression of opinion or its absence, namely "Subjective" and "Descriptive", and 2 labels targeting the quality of justification, namely "Vague" and "Adequate Sources". Each label was explicitly defined and accompanied by examples in the annotation manual, except for "Fake News", which was deliberately left up to the annotator to judge without explicit criteria, in order to find out about its best predictors among the other labels. The label "False Information" was presented last, since the annotators were told they had the option to do some research and fact-checking on each topic if necessary, but in order to minimize the risk of the annotators coming across the source of the articles. The labels were binary (1 for "applies" and 0 for "does not apply") and the annotators forced to choose between them (with the option of giving a free commentary). Some of our labels, finally, overlap with the propaganda techniques listed in Da San Martino et al. (2020a), in particular our label "Pejorative" with their "Name calling" and "Doubt", "Exaggeration" with "Exaggeration/Minimization", "Satirical/Pejorative/Subjective" with their "Loaded language", and "Vague" with their "Obfuscation/Intentional vagueness", except that they define vagueness mostly in terms of confusion and unclarity, whereas our definition targets generality/lack of specificity.

[2]See the 1971 Charter of Munich.

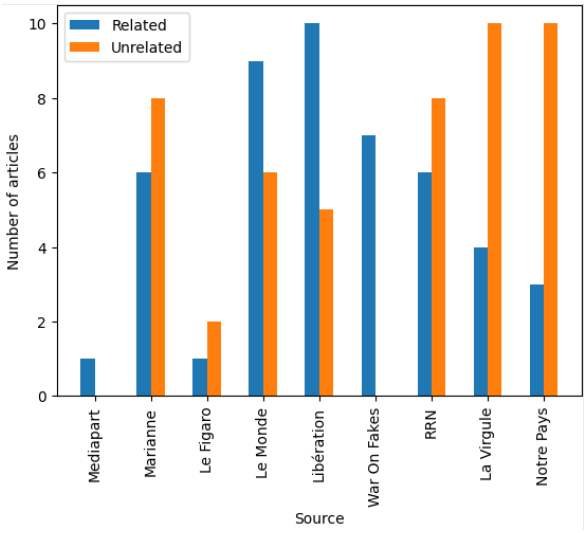


Figure 2: Topic distribution of articles from the annotated corpus.

After the annotation experiment, an additional analysis of the topics was conducted to ensure that regular articles were roughly about the same topics as propaganda articles, in order to validate the experiment results. To this end, we labeled the articles depending on whether they were directly about the armed conflict (labeled *Related*) or about other topics such as economic sanctions or politics (labeled *Unrelated*). The articles' distribution is shown in Figure 2.

Every source, with the exception of *waronfakes* and *mediapart*, had articles in both classes. *mediapart* had only one article meeting our filtering conditions, and *waronfakes* aims at denying war crimes allegations, so it is logical that it only contains articles directly about the armed conflict. Unexpectedly, the sample from *lavirgule* and *notrepays* contained more articles not directly linked to the conflict. Those articles seem to aim at polarizing the public debate not only on the war in Ukraine, but on other topics as well, including French politics. Overall, the annotated dataset is balanced,

with 27 *Related* regular articles, 21 *Unrelated* regular articles, 20 *Related* propaganda articles, and 28 *Unrelated* propaganda articles.

4 Analysis of the annotations

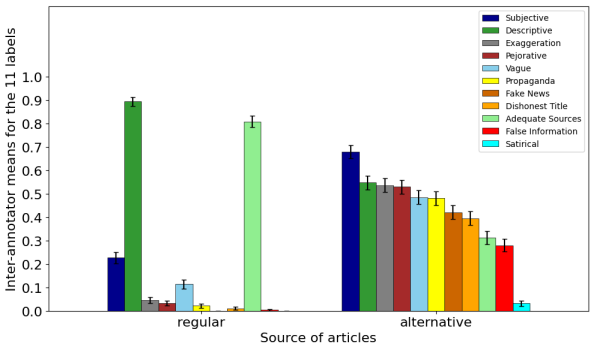
The 6 annotators included the designers of the experiment. Only one of them had briefly seen the texts prior to annotating, in order to upload them on the form used for the annotation task, but without verifying their content. The articles were presented in a common random order for all participants. To avoid bias by source, the URL was removed, in contrast to other datasets (viz. ISOT, Ahmed et al. 2018 or Horne and Adali 2017).

One article happened to contain mostly video links, leaving a meta-content description of the journal's policies on cookies: it could not be annotated, and was removed, leaving a total of 48 alternative vs. 47 regular articles for analysis. Among those, five articles (4 regular, 1 alternative) happened to bear an indication of their source by self-citation in their content. Eleven articles were also truncated because they were behind a paywall (ending on the necessity to subscribe in order to access content). We kept them for analysis, but knowing that they might introduce a confound. Importantly, however, post-hoc analyses made after exclusion of those 16 articles show the same main contrasts as reported below.

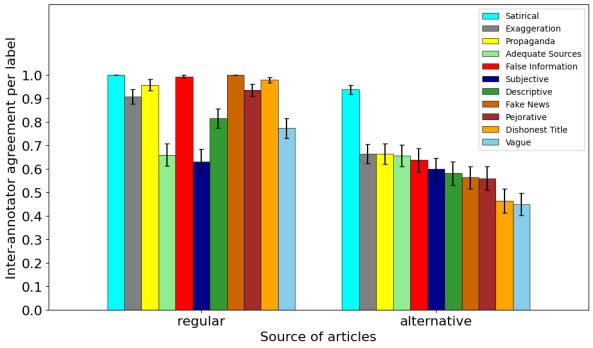
The combined dataset presents individual annotations grouped by annotator, instead of aggregate data (as PolitiFact and GossipCop, Shu et al. 2018), dropping personal commentaries on the articles to secure anonymity.

In order to assess the quality of the annotations, we calculated the inter-rater agreement based on the percentage of agreement between annotators, rescaled to 0 in a case of equal split between annotators (3:3), and to 1 in case of unanimity (6:0). That is, for each document, we computed the proportion x of 1-answers, rescaled by the function returning the value $|2x - 1|$. For example, a value of 0.4 indicates that 70% of the raters go in the same direction, while a value of .6 or above indicates 80% of agreement or more.

As shown in Figure 3b, for both the regular and the alternative press articles, all labels reached a mean value above .4, indicative of moderate to high agreement. The agreement between annotators increases systematically from the alternative to the regular corpus, meaning that for each label, the agreement is higher in the regular corpus, compared to the alternative corpus.



(a) Mean inter-annotator scores per label.



(b) Mean inter-annotator agreement per label.

Figure 3: Mean scores and agreement by label (error bars=standard error of the mean).

Regarding the labels themselves, Figure 3a shows a strong contrast between the two types of corpora. Except for the label "Satirical", which is almost never used in either type of corpus, the other 10 labels are used in very distinct proportions in either type of corpora (paired t-tests between the two corpora by label are all significant at the $\alpha = .01$ significance level). While each of the 10 remaining labels is applied to some extent in the alternative corpus, two labels are conspicuously never applied in the case of the regular corpus, namely: "False Information" and "Fake News". The labels "Descriptive" and "Adequate Sources", used for both types of corpora, are used in much higher proportion in the regular case. The labels "Subjective" and "Vague", while occurring for the regular corpus, are much less prevalent in the regular corpus. Finally, all other labels, in particular "Exaggeration", "Propaganda", "Pejorative", "Dishonest Title", are applied only marginally in the regular corpus.

The correlation matrix of the labels is displayed in Figure 4. The label "Satirical" is not corre-

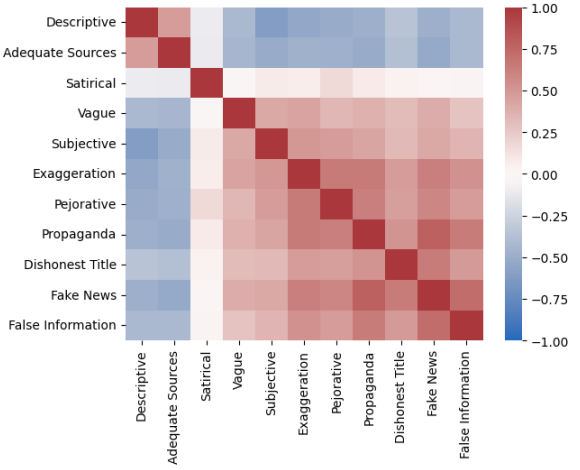


Figure 4: Correlation matrix of the 11 labels used for human annotations.

lated to other labels, due to its low frequency in the annotations (about 1.5% of annotations), and is left out in the remaining of the analysis. Two main groups of labels emerge from the matrix: the labels "Descriptive" and "Adequate Sources" are strongly correlated with each other and inversely with the others, and the remaining labels, including "Vague", "Subjective", etc., are positively correlated to various degrees. Our main label of interest, "Propaganda", correlates most strongly with "Fake News", "Pejorative", and "Exaggeration".

In summary, the annotators were able to reliably discriminate between the two corpora, across each of the dimensions selected by a specific label, and moreover the strong correlation between the labels "Propaganda" and "Exaggeration" legitimizes an analysis in terms of stylistic cues.

5 Analysis with the VAGO tools

To see what textual features might explain the difference between the two classes, we used the lexical database and analyzer VAGO (Icard et al., 2022). For a given text, VAGO calculates three scores: a score of vagueness, a score of opinion, and a score of relative detail (compared to vagueness). To calculate the vagueness score of a text, the system checks for the occurrence of vague expressions, subcategorized into four types: generality V_G ("*some*", "*or*"), approximation V_A ("*about*", "*almost*"), one-dimensional vagueness V_D ("*old*", "*many*"), and multi-dimensional vagueness V_C ("*good*", "*effective*"). For opinion, VAGO checks for the occurrence of implicit markers of subjectivity (all expressions of type V_D and V_C, including evaluative adjectives and pejorative terms), as well as explicit markers (first-person pronouns, exclamation marks). For detail, finally, the system compares the ratio of named entities to vague terms.

While VAGO does not incorporate any world-knowledge, previous studies on larger corpora have shown that the VAGO scores of vagueness and opinion were positively correlated with the label "biased" in news articles (Guélorget et al., 2021; Icard et al., 2023), and that the score of detail-vs-vagueness was negatively correlated with the label "Satirical" (Icard et al., 2023). Hence, we asked if the VAGO scores of vagueness, opinion, and detail might be good predictors of the human annotations, and in particular of labels such as "Exaggeration", "Pejorative", "Propaganda" and "Dishonest Title".

To investigate this question, we calculated the correlation between the VAGO scores for each article of the corpus and the mean inter-annotator scores for all of the 10 labels ("Satirical" left aside). As shown in Table 3, the labels "Subjective", "Exaggeration" and "Pejorative" turned out to be positively correlated to the VAGO scores of vagueness and opinion, and negatively correlated to the scores of detail-vs-vagueness. Consistent with these results, the scores of vagueness and opinion were also negatively correlated with labels "Descriptive" and "Adequate Sources". By contrast, labels "Propaganda", "Dishonest Title", "Fake News" and "False Information" turned out to be positively correlated to the scores of vagueness only. All these correlations are weak to moderate, but they replicate results found in previous studies, with an even higher order of magnitude in the labels "Subjective" and "Descriptive" connected to VAGO's opinion score, as presented in Figure 5.

Human annotations of the label "Vague" did not correlate with VAGO scores of either vagueness or detail, however, contrary to expectations. We conjecture that this could be due to a discrepancy between the definition given of the label, which targets generality vagueness, and the fact that the VAGO vagueness score is based on more types of vagueness, in particular the semantic vagueness of one-dimensional and multi-dimensional adjectives, which represent 96% of the VAGO lexicon.

Despite that, what Table 3 shows is that the VAGO scores track the clustering of labels found in Figure 4: the polarity of the correlations for the labels "Descriptive" and "Adequate sources" is inverse to

that of the other labels. In summary, VAGO scores are correlated with the separating features of the alternative vs. regular press, but they explain only part of the variance in the annotations. In the next section, we examine classification models properly in order to get further insights.

Label	vague	opinion	detail
Vague	0.163	0.188	−0.180
Subjective	0.344*	0.384**	−0.238
Exaggeration	0.282	0.222	−0.225
Pejorative	0.289	0.222	−0.265
Descriptive	−0.371**	−0.367**	0.228
Propaganda	0.249	0.165	−0.152
Dishonest Title	0.257	0.164	−0.206
Adequate Sources	−0.210	−0.210	0.130
Fake News	0.233	0.178	−0.148
False Information	0.214	0.140	−0.099

Table 3: Pearson correlations between the labels' mean scores and the VAGO scores (* and ** indicate p-value $< .05$ and $< .01$), with Bonferroni correction.

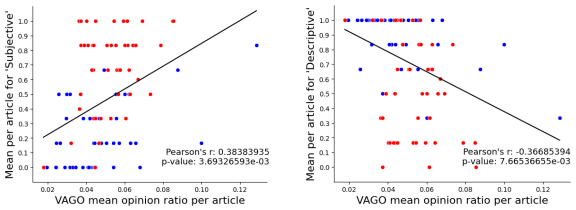


Figure 5: Pearson correlations between the VAGO mean opinion score per article and the mean scores for labels "Subjective" (left) and "Descriptive" (right). Blue data points correspond to regular articles while red data points correspond to alternative press articles.

6 Machine learning for propaganda detection

Propaganda detection (Da San Martino et al., 2020b) from texts can be a difficult task depending on the form of the content. Classifying sentences (Mapes et al., 2019) is harder, even for large language models (LLMs) such as BERT (Devlin et al., 2018). In this section, a methodology for training a propaganda detection model is explained and evaluated. Smaller models with explainability capabilities were also trained in order to identify which parts of the articles the model considers when taking its decision.

6.1 Dataset for detecting propaganda related to the conflict

In order to train a model that could be used to identify propaganda articles, it is required to also collect regular press articles on a related topic. Here, we present the larger corpus of regular press from which the French subset of the previous section was drawn. This larger corpus also contains English articles, since the classification model is supposed to handle classification in French and in English.

English regular articles were collected from 11 reliable news outlets, with constraints of date (being post Ukraine invasion), length (between 1,000 and 10,000 characters), and topic (mention Russia and Ukraine). English regular articles were collected using news-please (Hamborg et al., 2017) before being filtered. The articles distribution by source is given in Table 4. The wider set of French regular articles was collected in the same way, but with a more limited choice of sources, their distribution is given in Table 5.

Source	Number of articles
apnews.com	520
cbsnews.com	63
dailymail.co.uk	43
cnn.com	10
usatoday.com	10
forbes.com	42
foxnews.com	5
nbcnews.com	10
nytimes.com	4
theguardian.com	185
washingtonpost.com	12
Total	1,004

Table 4: English language regular articles distribution by source.

Source	Number of articles
lefigaro.fr	3
lemonde.fr	449
liberation.fr	386
marianne.net	523
mediapart.fr	6
Total	1,367

Table 5: French regular articles distribution by source.

6.2 Models

Five models were chosen for propaganda detection, two in English and three in French. The English[3] and French[4] models are available on Huggingface-hub and can be freely downloaded and tested.

The first English model used for classification is a RoBERTa-base model (Liu et al., 2019) with a classification layer using the last hidden state. For practicality, we load pre-trained English RoBERTa

[3] https://huggingface.co/hybrinfox/ukraine-operation_propaganda-detection-EN

[4] https://huggingface.co/hybrinfox/ukraine-operation_propaganda-detection-FR

weights and fine-tune the model using the HuggingFace transformers library.

The first French model combines the "CamemBERT-base" version (Martin et al., 2019) based on the RoBERTa architecture (Liu et al., 2019) (*Batch Size*=10, *Learning Rate*=1e-05, *Epochs*=5) with one classification layer and a BCE loss function to detect whether the articles of our French larger dataset counts as propaganda or not.

The second French model is an XGBoost (Chen and Guestrin, 2016) (Extreme Gradient Boosting) model. It is a scalable, distributed gradient-boosted decision tree. Contrarily to the other three models which process texts directly, XGBoost only takes numerical values as input. In our case it takes the following parameters: the length of the sentence, the three VAGO scores (vagueness, opinion, detail), the sentiment of the sentence, positive or negative (using the HuggingFace sentiment classification model "Monsia/camembert-fr-covid-tweet-sentiment-classification"), the number of verbs, adjectives, adverbs and nouns present in the sentence and the number of occurrences of dependencies between the words (using the spaCy python library for Natural Language Processing).[5] The sentence features are then aggregated by an operator. Several aggregation operators were tested and gave similar results so the sum operator was chosen.

Models applicable to both languages were tested. The first is the neurosymbolic model CATS (Faye et al., 2023). It does not use *a priori* knowledge on the language except for the English syntax. It is lighter than RoBERTa, and has explainability capabilities that will be useful to identify what the model considers a marker of propaganda. It can also be used for other languages and results for a French version have also been reported. The second one is TF-IDF, with which the texts are vectorized after removing stopwords and lemmatizing the remaining words. This representation is then processed by a random forest, predicting the class of the article.

The datasets for each language were initially split between training, validation and test using a 80/10/10 ratio with no overlap. The models were chosen on the best validation score and the reported results are on the test set, which was never used during the training procedure.

[5] https://universaldependencies.org/u/dep/all.html#al-u-dep/nmod

6.3 Results

Language	Models	Test accuracy
English	RoBERTa	0.997
	CATS - EN	0.953
	TF-IDF - EN	0.985
French	CamemBERT	0.997
	CATS - FR	0.946
	XGBoost	0.921
	TF-IDF - FR	0.963

Table 6: Test accuracies for Ukraine invasion propaganda detection models.

The models' performances on their test sets are reported in Table 6. Propaganda detection on this specific topic is easily achieved by LLMs, and even by shallow models like CATS or XGBoost. The performance of CATS is slightly lower than RoBERTa's, but this is expected since it contains only 0.6 million parameters, about 200 times fewer parameters than RoBERTa-base with its 125 million parameters. XGBoost's performance is even lower, but the model processes high-level features of the texts, lacking other features that other models can use.

6.4 Identified markers of propaganda

The interest of training a smaller model like CATS on the texts is to identify which markers are learnt by this machine learning model. To this end, each token's contribution to the final decision is aggregated by sentence, enabling us to recover the most salient sentences from propaganda articles. These sentences contain more markers of propaganda and can help us understand what the model is tracking when classifying articles between propaganda and regular.

A representative example is given in Figure 6. In this example, the first underlined sentence is a case of laudatory exaggeration; the second one is pejorative, and the third is again pejorative, with even a racist insinuation. Other sentences in the text contain propagandist cues, however, making the selection hard to directly interpret. For comparison, we run VAGO on the text. In this case, the scores of vagueness, opinion, and detail were 0.13, 0.08 and 0.42, respectively. The underlined items correspond to vague and subjective markers found by VAGO. VAGO detects several adjectives used pejoratively ("Old [Joe]", "trivial" and "simple" in particular). It misses out on others ("smug", "round lost"), and on more complex syntactic markers ("even" in "even a child", "by the way" to introduce a derogatory and covertly racist remark). But

it identifies several subjective adjectives reflecting the implicit viewpoint of the writer.

> “Round Lost Joe Biden made a rant in Warsaw about the ” unity of the West ” and the” power of democracy ”. But in his own country , Vladimir Putin was more believable . The American president’ s speech in Poland was not intended as a direct response to the Russian leader , who addressed the Federal Assembly the day before - and the entire world as well. Biden ’ s national security adviser Jake Sullivan claimed it was ” not a rhetorical contest with anybody ” . But the 80 - year - old politician ’ s smug stand - up proved otherwise: he tried to confront his opponent from Moscow - and appeared to yield to him. Old Joe was satisfied with a 20 - minute monologue on the lawn of the Royal Castle - by comparison , Vladimir Putin spoke for 1 hour 45 minutes. Biden’ s entire message was made up of high-pitched quotations - especially for the applause he prepared : ” Democracies have become stronger , not weaker. Autocracies have grown weaker , not stronger.” Quite trivial and as simple as possible - so that even a child would get the point . By the way , there were a lot of children at the President’ s speech , and of different races too. And all of them had Ukrainian flags - in the best traditions of American propaganda.”

Figure 6: Example of an article classified as propaganda by CATS. The sentences contributing the most to the propaganda class according to CATS are highlighted in red while the VAGO vocabulary is underlined.

6.5 Explainability of the XGBoost model

We used the SHAP tool (Lundberg and Lee, 2017) to analyze which features were the most useful for the XGBoost classification. The results are reported in Figure 7. We observe that overall syntactic features bear more weight than other features in the detection of propaganda, with the number of punctuation marks (PUNCTUATION) having greater impact than the length of sentences (LENGTH_SENT), the number of clausal modifiers (ACL), of nominal subjects (NSUBJ) and of sentences (ROOT) all receiving similar weight.

In Figure 8, we observe that the frequency percentage of punctuation compared to other tokens is significantly higher in regular articles than in propaganda articles ($p = 8.31 \times 10^{-240}$). We observed more precisely which type of punctuation was more represented in regular versus propaganda articles. Compared to other tokens, we

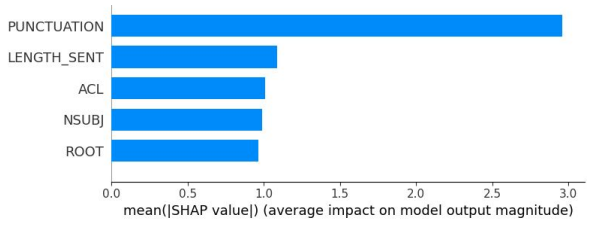


Figure 7: SHAP explainability of the XGBoost model for propaganda classification. Only the top 5 syntactic features are displayed.

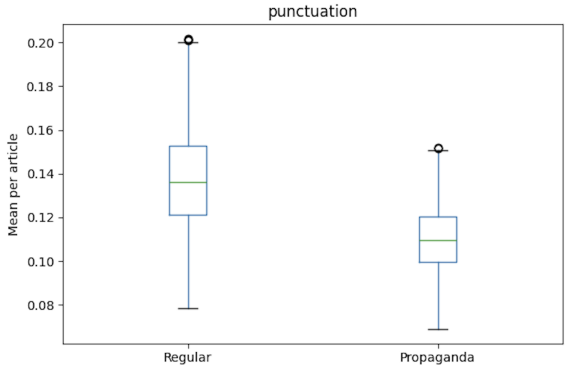


Figure 8: Percentage frequency distributions of “punct” dependence in regular articles vs. propaganda articles.

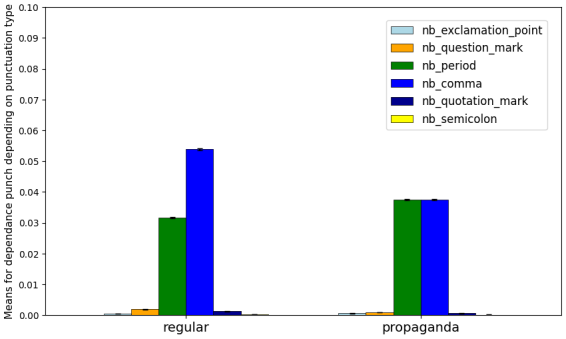


Figure 9: Relative weight of punctuation marks in either article type.

observed that propaganda articles contain significantly more periods ($p = 2.09 \times 10^{-78}$), but fewer question marks ($p = 2.12 \times 10^{-32}$), fewer commas ($p = 1.47 \times 10^{-290}$) and fewer quotation marks ($p = 1.12 \times 10^{-06}$) than regular articles (see Figure 9). Since propaganda articles happened to be significantly shorter than regular articles ($p = 7.12 \times 10^{-26}$), the data was normalized by the length of the article, corresponding to the total number of tokens in the article.

Looking at the VAGO-N scores on the corpora, we observe that, besides punctuation, the VAGO-N mean score of detail vs vagueness per article is significantly higher for regular articles than for propaganda articles ($p = 2.66 \times 10^{-44}$, with Bonferroni correction). By contrast, the differences between the VAGO-N scores of vagueness and opinion are no longer significant after Bonferroni correction.

6.6 Potential biases of machine learning models

The near perfect accuracy of the models reported in Table 6 concerning Large Language Models raises questions about the shallowness of the learnt features and about potential biases in the dataset.

Regarding the first aspect, the high performance of models such as TF-IDF and CATS shows that these simpler models can also detect propaganda when trained on a large dataset. The deeper models, as a result of their higher complexity, can achieve better scores, very close to 100%.

The high accuracy of TF-IDF, which uses only lexical features, manifests a clear distinction between the language of regular articles versus propaganda articles when they deal with the topic of Ukraine operation. While the models are performing well on this specific topic, there is no guarantee that they would perform equally well on other propaganda topics.

We analyzed the terms whose TF-IDF scores differ significantly between the two classes in the French corpus. Among the terms more prevalent in the propaganda corpus compared to the regular corpus, we find terms like "état" (*state*), "pays" (*country*), "unis" (*united*), "déclaré" (*declared*), "ue" (*EU*), "zelensky", "biden", "kiev","allemagne" (*Germany*), "armes" (*weapons*). By contrast, terms like "lire" (*read*), "russe" (*Russian*), "poutine", "kyiv", "invasion", "vladimir", "guerre" (*war*), "jeudi" (*thursday*), "mars" (*march*) and "lundi" (*monday*) are more prevalent in the regular corpus. We notice that "Kiev/Kyiv" is not spelled the same way depending on the corpus. The name "Zelensky" is cited more in propaganda articles, whereas "Putin" is cited more in the regular articles of the corpus. Finally, the regular corpus contains more markers of precise time indications than the propaganda corpus, consistently with the higher VAGO score of detail.

7 Conclusion and perspectives

In this paper, we introduced PPN, a multilingual propaganda dataset, and we conducted an experiment to investigate the basis on which human annotators, and then classification algorithms, can discriminate propagandist articles from non-propagandist articles on a specific topic. The annotations reveal that exaggeration, combined with lesser descriptive content, and absence of adequate sources, are prevalent in assessments of propagandist press. The VAGO analyzer confirmed that the use of vague markers is significantly correlated with those features. Further analyses based on different families of classifiers revealed further syntactic cues, pertaining in particular to punctuation, but also to the lexicon.

Further work is needed to refine this analysis. Machine learning models, while efficient at detecting topic-specific propaganda, still have room for improvement regarding explainability and generalization to other topics. If some alignment has been observed with what humans attend to when judging an article, there is still no guarantee that language models process the text as humans would. The use of propaganda technique classifiers to identify manipulative articles yields more explainability, but at the cost of performance, especially for topic-specific propaganda.

In addition to that, while the given scores are very high, they were obtained for the task of *topic-specific* propaganda detection, which is an easier task than general propaganda detection. However, topic-specific models still have use and can prevent the spread of disinformation in cases of conflict similar to the one used here.

While only a model for English and French propaganda detection on the Ukraine invasion is provided here, we encourage the community to use the parts of the dataset corresponding to their native language to train more classifiers. Collaborations could be considered to train a multilingual model, based on the dataset and collected regular articles from the other languages of the dataset. The same goes for annotation experiments on the way propaganda is perceived by readers, as propaganda strategies may change by languages and by target audience.

Last, in this paper we see that symbolic AI tools explain part of the classifications operated by humans as well as by classifiers. We see two ways in which explainability can be further improved: firstly, by continuing to enrich tools like VAGO with lexical and even syntactic units highlighted by classifiers or by annotators in this task; secondly by considering more labels in order to improve the quality of annotations and identify more stylistic features. We introduced a label for "Pejorative" speech, we may also have introduced a dual label "Laudatory", to identify cases of glorification also typical of state propaganda, and to refine the category of "Exaggeration". Similarly, we may want to better control the positive and negative connotations of the labels, for instance by using labels such as "Precise" rather than "Vague", or "Objective" instead of "Subjective".

Limitations

Annotation experiments were only run on a subset of the French data. While an additional manual verification of the data quality has been done for English articles, other languages have not been manually reviewed. There may be parsing errors for some languages, and further analysis from native speakers of other languages may be required before using these parts of the dataset.

Experiments on propaganda detection were only run on two examples of Romance and Germanic languages. While language models for these types of languages are common, there is no guarantee that performant language models exist for all proposed languages from the dataset.

Ethics statement

This article deals with the topic of propaganda and proposes a dataset to help improve propaganda detection. Proposing and sharing propaganda detection methods is crucial to keep the information space clean and safe to use for everyone.

Human exposition to propaganda should be contained. To this end, we ensured that all annotators were performing the annotation task voluntarily, with a content warning, and the possibility to stop the experiment at any time.

We encourage future works on the dataset to be conducted cautiously and on limited parts of the global dataset.

Acknowledgements

We thank two anonymous reviewers for helpful comments and feedback. This work was supported by the programs HYBRINFOX (ANR-21-ASIA-0003), FRONTCOG (ANR-17-EURE-0017), and PLEXUS (Marie Skłodowska-Curie Action, Horizon Europe Research and Innovation Programme, grant n°101086295). PE thanks Monash University for hosting during the writing of this paper.

Declaration of contribution

All the authors contributed to the design, annotations, analysis and discussion of the results. GF, BI, MC, and PE wrote the paper, which all authors read and revised together. First authorship is equally shared between GF, BI and MC. Correspondence: geraud.faye@centralesupelec.fr, benjamin.icard@ens.fr, morgane.casanova@irisa.fr, paul.egre@ens.psl.eu.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *SECURITY AND PRIVACY*, 1(1):e9.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

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 Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. *CoRR*, abs/2007.08024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

Paul Égré and Benjamin Icard. 2018. Lying and vagueness. In J. Meibauer, editor, *Oxford Handbook of Lying*. OUP.

Géraud Faye, Wassila Ouerdane, Guillaume Gadek, Souhir Gahbiche, and Sylvain Gatepaille. 2023. A novel hybrid approach for text encoding: Cognitive attention to syntax model to detect online misinformation. *Data & Knowledge Engineering*, 148:102230.

Paul Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Paul Guélorget, Benjamin Icard, Guillaume Gadek, Souhir Gahbiche, Sylvain Gatepaille, Ghislain Atemezing, and Paul Égré. 2021. Combining vagueness detection with deep learning to identify fake news. In *IEEE 24th International Conference on Information Fusion (FUSION)*, pages 1–8.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Freddy Heppell, Kalina Bontcheva, and Carolina Scarton. 2023. Analysing state-backed propaganda websites: a new dataset and linguistic study.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Benjamin Icard, Ghislain Atemezing, and Paul Égré. 2022. VAGO: un outil en ligne de mesure du vague et de la subjectivité. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (PFIA 2022)*, pages 68–71.

Benjamin Icard, Vincent Claveau, Ghislain Atemezing, and Paul Égré. 2023. Measuring vagueness and subjectivity in texts: from symbolic to neural VAGO. In *IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2023)*.

Garth S Jowett and Victoria O'Donnell. 2019. *Propaganda & persuasion*. Sage publications. 7th edition.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Norman Mapes, Anna White, Radhika Medury, and Sumeet Dua. 2019. Divisive language and propaganda detection using multi-head attention transformers with deep learning BERT-based language models for binary classification. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 103–106, Hong Kong, China. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a tasty French language model. *arXiv preprint*.

Anne Quaranto and Jason Stanley. 2021. Propaganda. In Justin Khoo and Rachel Katharine Sterken, editors, *The Routledge Handbook of Social and Political Philosophy of Language*, pages 125–146.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint*.