Interpretable Propaganda Detection in News Articles

Seunghak Yu[1*] **Giovanni Da San Martino**[2]
Mitra Mohtarami[3] **James Glass**[3] **Preslav Nakov**[4]
[1] Amazon Alexa AI, Seattle, WA, USA
[2] Department of Mathematics, University of Padova, Italy
[3] MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
[4] Qatar Computing Research Institute, HBKU, Qatar
yuseungh@amazon.com, dasan@math.unipd.it
{mitra, glass}@csail.mit.edu, pnakov@hbku.edu.qa

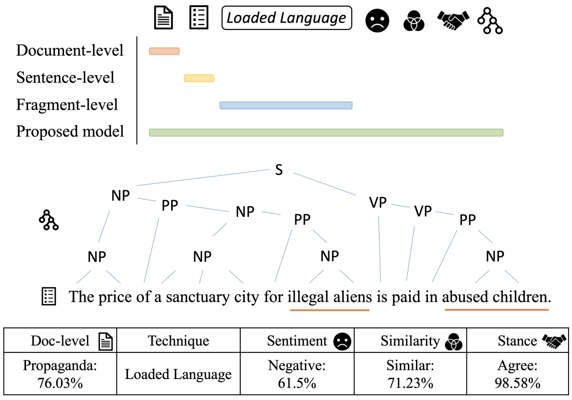
Abstract

Online users today are exposed to misleading and propagandistic news articles and media posts on a daily basis. To counter thus, a number of approaches have been designed aiming to achieve a healthier and safer online news and media consumption. Automatic systems are able to support humans in detecting such content; yet, a major impediment to their broad adoption is that besides being accurate, the decisions of such systems need also to be interpretable in order to be trusted and widely adopted by users. Since misleading and propagandistic content influences readers through the use of a number of deception techniques, we propose to detect and to show the use of such techniques as a way to offer interpretability. In particular, we define qualitatively descriptive features and we analyze their suitability for detecting deception techniques. We further show that our interpretable features can be easily combined with pre-trained language models, yielding state-of-the-art results.

1 Introduction

With the rise of the Internet and social media, there was also a rise of fake (Nguyen et al., 2020), biased (Baly et al., 2020a,b), hyperpartisan (Potthast et al., 2018), and propagandistic content (Da San Martino et al., 2019b). In 2016, news got weaponized, aiming to influence the US Presidential election and the Brexit referendum, making the general public concerned about the dangers of the proliferation of fake news (Howard and Kollanyi, 2016; Faris et al., 2017; Lazer et al., 2018; Vosoughi et al., 2018; Bovet and Makse, 2019).

There ware two reasons for this. First, disinformation disguised as news created the illusion that the information is reliable, and thus people tended to lower their barrier of doubt compared to when information came from other types of sources.



Doc-level	Technique	Sentiment	Similarity	Stance
Propaganda: 76.03%	Loaded Language	Negative: 61.5%	Similar: 71.23%	Agree: 98.58%

Figure 1: Comparison of propaganda prediction interpretability using existing methods. Our proposed method helps users to interpret propaganda predictions across various dimensions, e.g., is there a lot of positive/negative sentiment (can signal the use of *loaded language*, which appeals to emotions), are the target sentence and the document body related to the title, does the sentence agree/disagree with the title, etc. Each symbol in the top bar chart represents an information source for propaganda detection.

Second, the rise of citizen journalism led to the proliferation of various online media, and the veracity of information became an issue. In practice, the effort required to fact-check the news, and its bias and propaganda remained the same or even got more complex, compared to traditional media, since the news was re-edited and passed through other media channels.

Propaganda aims to influence the audience with the aim of advancing a specific agenda (Da San Martino et al., 2020b). Detecting it is tricky and arguably more difficult than finding false information in an article. This is because propagandistic articles are not intended to simply make up a story with objective errors, but instead use a variety of techniques to convince people, such as selectively conveying facts or appealing to emotions (Jowett and O'Donnell, 2012).

[*] Work conducted while the author was at MIT CSAIL.

Proceedings of Recent Advances in Natural Language Processing, pages 1597–1605
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_179

While many techniques are ethically questionable, we can think of propaganda techniques as rhetorical expressions that effectively convey the author's opinion (O'Shaughnessy, 2004). Due to these characteristics, propagandistic articles are often produced primarily for political purposes (but are also very common in commercial advertisement), which directly affect our lives, and are commonly found even in major news media outlets, which are generally considered credible.

The importance of detecting propaganda in the news has been recently emphasized, and research is being conducted from various perspectives (Rashkin et al., 2017; Barrón-Cedeno et al., 2019a; Da San Martino et al., 2019b). However, while previous work has done reasonable job at detecting propaganda, it has largely ignored the question of why the content is propagandistic, i.e., there is a lack of interpretability of the system decisions, and in many cases, there is a lack of interpretability of the model as well, i.e., it is hard to understand what the model actually does even for its creators.

Interpretability is indispensable if propaganda detection systems are to be trusted and accepted by the users. According to the confirmation bias theory (Nickerson, 1998), people easily accept new information that is consistent with their beliefs, but are less likely to do so when it contradicts what they already know. Thus, even if a model can correctly predict which news is propagandistic, if it fails to explain the reason for that, people are more likely to reject the results and to stick to what they want to believe. In order to address this issue, we propose a new formulation of the propaganda detection task and a model that can explain the prediction results. Figure 1 compares the coverage of the explanations for pre-existing methods vs. our proposal.

Our contributions can be summarized as follows:

- We study how a number of information sources relate to the presence and the absence of propaganda in a piece of text.

- Based on this, we propose a general framework for interpretable propaganda detection.

- We demonstrate that our framework is complementary to and can be combined with large-scale pre-trained transformers, yielding sizable improvements over the state of the art.

2 Task Setup

Given a document $\boldsymbol{d}$ that consists of n sentences $\boldsymbol{d} = \{d_i\}_{i=1}^n$, each sentence should be classified as belonging to one of 18 propaganda techniques or as being non-propaganda. The exact definition of propaganda can be subtly different depending on the social environment and the individual's growth background, and thus it is not surprising that the propaganda techniques defined in the literature differ (Miller, 1939; Jowett and O'Donnell, 2012; Hobbs and McGee, 2014; Torok, 2015; Weston, 2018). The techniques we use in this paper are shown in Table 1. Da San Martino et al. (2019b) derived the propaganda techniques from the literature: they selected 18 techniques and manually annotated 451 news articles with a total of 20,110 sentences. This dataset[1] has fragment-level labels that can span over multiple sentences and can overlap with other labeled spans.

This granular labeling went beyond our scope and we had to restructure the data. First, we divided the data into sentences. Second, in order to reduce the complexity of the task, we changed the multi-label setup to a multi-class one by ignoring duplicate labels and only allowing one technique per sentence (the first one), breaking ties at random. As a result, we obtained 20,111 sentences labeled with a non-propaganda class or with one of 18 propaganda techniques. Based on this data, we built a system for predicting the use of propaganda techniques at the sentence level, and we provided the semantic and the structural information related to propaganda techniques as the basis of the results.

3 Proposed Method

Our method can detect the propaganda for each sentence in a document, and can explain what propaganda technique was used with interpretable semantic and syntactic features. We further propose novel features conceived in the study of human behavioral characteristics. More detail below.

3.1 People Do Not Read Full Articles

Behavior studies have shown that people read less than 50% of the articles they find online, and often stop reading after the first few sentences, or even after the title (Manjoo, 2013). Indeed, we found that 77.5% of our articles use propaganda techniques in the first five sentences, 65% do so in the first three sentences, and 31.07% do so in the title.

[1] http://propaganda.math.unipd.it/

Techniques	Definition
Name Calling	give an object an insulting label
Repetition	inject the same message over and over
Slogans	use a brief and memorable phrase
Appeal to Fear	plant fear against other alternatives
Doubt	questioning the credibility
Exaggeration	exaggerate or minimize something
Flag-Waving	appeal to patriotism
LL	appeal to emotions or stereotypes
RtoH	the disgusted group likes the idea
Bandwagon	appeal to popularity
CO	assume a simple cause for the outcome
OIC	use obscure expressions to confuse
AA	use authority's support as evidence
B&W Fallacy	present only two options among many
TC	discourage meaningful discussion
Red Herring	introduce irrelevant material to distract
Straw Men	refute a nonexistent argument
Whataboutism	discredit an opponent's position

Table 1: List of propaganda techniques and brief definitions. LL: Loaded Language, RtoH: Reduction to Hitlerum, CO: Casual Oversimplification, OIC: Obfuscation, Intentional vagueness, Confusion, AA: Appeal to Authority, TC: Thought-terminating Clichés.

We used three types of features ($\boldsymbol{f}^{rp}$, $\boldsymbol{f}^{sim}$, $\boldsymbol{f}^{stn}$) to account for these observations, which we describe below.

3.1.1 Relative Position of the Sentence

We define the relative position of a sentence as $\boldsymbol{f}_i^{rp} = i/n$, where i is the sequence number of the sentence, and n is the total number of sentences in the article.

3.1.2 Topic Similarity and Stance with Respect to the Title

The title of an article typically contains the topic and also the author's view of that topic. Thus, we hypothesize that propaganda should also focus on the topic expressed in the title.

We represent the relationship between the target sentence and the title by measuring the semantic similarity $\boldsymbol{f}_i^{sim}$ between them as the cosine between the sentence-BERT representations ($\boldsymbol{\phi}(x)$) (Reimers and Gurevych, 2019) of the target sentence d_i and of the title d_1.

$$\boldsymbol{f}_i^{sim} = \frac{\boldsymbol{\phi}(d_1) \cdot \boldsymbol{\phi}(d_i)}{|\boldsymbol{\phi}(d_1)||\boldsymbol{\phi}(d_i)|} \tag{1}$$

We further model the stance of a target sentence with respect to the title $\boldsymbol{f}_i^{stn}$ using a distribution over five classes: *related*, *unrelated*, *agree*, *disagree*, and *discuss*. For this, we use a BERT model (Fang et al., 2019) fine-tuned on the Fake News Challenge dataset (Hanselowski et al., 2018).

Level	Phrases
Clause	S, SBAR, SBARQ, SINV, SQ
Phrase	ADJP, ADVP, CONJP, FRAG, INTJ, LST, NAC, NP, NX, PP, PRN, PRT, QP, RRC, UCP, VP, WHADJP, WHAVP, WHADVP, WHNP, WHPP, X

Table 2: The syntactic labels we used as features.

The class *unrelated* indicates that the sentence is not related to the claim made in the title, while *agree* and *disagree* refer to the sentence agreeing/disagreeing with the title, and finally *discuss* is assigned when the topic is the same as that in the title, but there is no stance. We further introduce the *related* class as the union of *agree*, *disagree*, and *discuss*. We use as features the binary classification labels and also the probabilities for these five classes.

3.2 Syntactic and Semantic Information

Some propaganda techniques have specific structural or semantic characteristics. For example, *Loaded Language* can be configured to elicit an emotional response, usually using an emotional noun phrase. To model this, we define the following three features: $\boldsymbol{f}^{dp}$, $\boldsymbol{f}^{sent}$, and $\boldsymbol{f}^{doc}$.

3.2.1 Syntactic Information

We used a syntactic parser to extract structural features about the target sentence $\boldsymbol{f}_i^{dp}$. Our hypothesis is that such information could help to discover techniques that have specific structural characteristics such as *Doubt* and *Black and White Fallacy*. We considered a total of 27 clause-level classes and phrase-level labels, including the *unknown* class. The set is shown in Table 2.

3.2.2 Sentiment of the Sentence

The sentiment of the sentence $\boldsymbol{f}_i^{sent}$ is another important feature for detecting propaganda. This is because many propagandistic articles try to convince the readers by appealing to their emotions and prejudices. Thus, we extract the sentiment using a sentiment analyzer trained on social media data (Hutto and Gilbert, 2014), which gives a probability distribution over the following three classes: *positive*, *neutral*, and *negative*. It further outputs *compound*, which is a one-dimensional normalized, weighted composite score. We use all four scores as features.

3.2.3 Document-Level Prediction

If the document is likely to be propagandistic, then each of its sentences is more likely to contain propaganda. To model this, we use as a feature $\boldsymbol{f}^{doc}$ the score of the document-level propaganda classifier Proppy (Barrón-Cedeno et al., 2019a). Note that Proppy is trained on articles labeled using media-level labels, i.e., using distant supervision. Therefore, all articles from a propagandistic source are considered to be propagandistic.

4 Experimental Results

In this section, we present our experimental setup for interpretable propaganda detection and the evaluation results from our experiments. Specifically, we perform three sets of experiments: (*i*) in Section 4.1, we quantitatively analyze the effectiveness of the features we proposed in Section 3; (*ii*) in Sections 4.2 and 4.3, we compare our feature-based model to the state-of-the-art model described in (Da San Martino et al., 2019b) using the experimental setup from that paper; (*iii*) in Section 4.4, we analyze the performance of our model with respect to each of the 18 propaganda techniques.

4.1 Quantitative Analysis of the Proposed Features

Figure 2 shows the absolute value of the covariance between each of our features $\boldsymbol{f}$ and each of the 18 propaganda techniques $\boldsymbol{T}$. We calculated the values of the features on the training and on the development datasets, and we standardized their values. Then, we formulated this as a problem of calculating the covariance between continuous and Bernoulli random variables as follows: $cov(\boldsymbol{f}, \boldsymbol{T}) = p \cdot (1-p) \cdot (E[\boldsymbol{f}|\boldsymbol{T} = 1] - E[\boldsymbol{f}|\boldsymbol{T} = 0])$.

The total number of sentences used is 16,137 (for the training and for the development datasets, combined), among which there are 4,584 propagandistic sentences. In Figure 2, the vertical axis represents the proposed features, and the horizontal axis shows the individual propaganda techniques and the total number of instances thereof. Each square shows an absolute value of the covariance between some feature and some propaganda technique. We show absolute values in order to ignore the direction of the relationship, and we apply a threshold of 0.001 in order to remove the negligible relations from the figure.

Although the most frequent propaganda techniques appear in less than 10% of the examples, they do show qualitatively meaningful associations. Indeed, we do not expect a feature to correlate with multiple techniques, as they are fundamentally different. We believe that having features that strongly correlate with one technique might be an advancement towards detecting that technique.

We can see that the structural information ($\boldsymbol{f}^{dp}$) and the sentiment of a sentence ($\boldsymbol{f}^{sent}$) are closely associated with certain propaganda techniques. For example, *Loaded Language* has a strong correlation with features identifying words bearing either a positive or a negative sentiment. This makes sense as the authors are more likely to use emotional words rather than neutral ones, and *Loaded Language* aims to elicit an emotional response. Similarly, *Doubt* has high correlation with certain syntactic categories.

There are a number of interesting observations about the other features. For example, the relative position of sentences ($\boldsymbol{f}^{rp}$) is associated with more than half of the propaganda techniques. Moreover, the similarity to the title ($\boldsymbol{f}^{sim}$) and the stance with respect to the title ($\boldsymbol{f}^{stn}$) are strongly correlated with the likelihood that the target sentence is propagandistic. The features that indicate whether a sentence is related to the subject of the title are complementary, and thus the covariances are the same when absolute values are considered.

4.2 Comparison to Existing Approaches

Table 3 shows a performance comparison for our model vs. existing models on the sentence-level propaganda detection dataset (Da San Martino et al., 2019b). This dataset consists of 451 manually annotated articles, collected from various media sources, and a total of 20,111 sentences. Unlike the experimental setting in the previous sections, the task here is a binary classification one: given a sentence, the goal is to predict whether it contains *at least one* of the 18 techniques or not. For the performance comparison, we used BERT (Devlin et al., 2019), which we fine-tuned for sentence-level classification using the Multi-Granularity Network (MGN) (Da San Martino et al., 2019b) architecture on top of the [CLS] tokens (trained end-to-end), as this model improves the performance for both tasks by controlling the word-level prediction using information from the sentence-level prediction and vice versa.

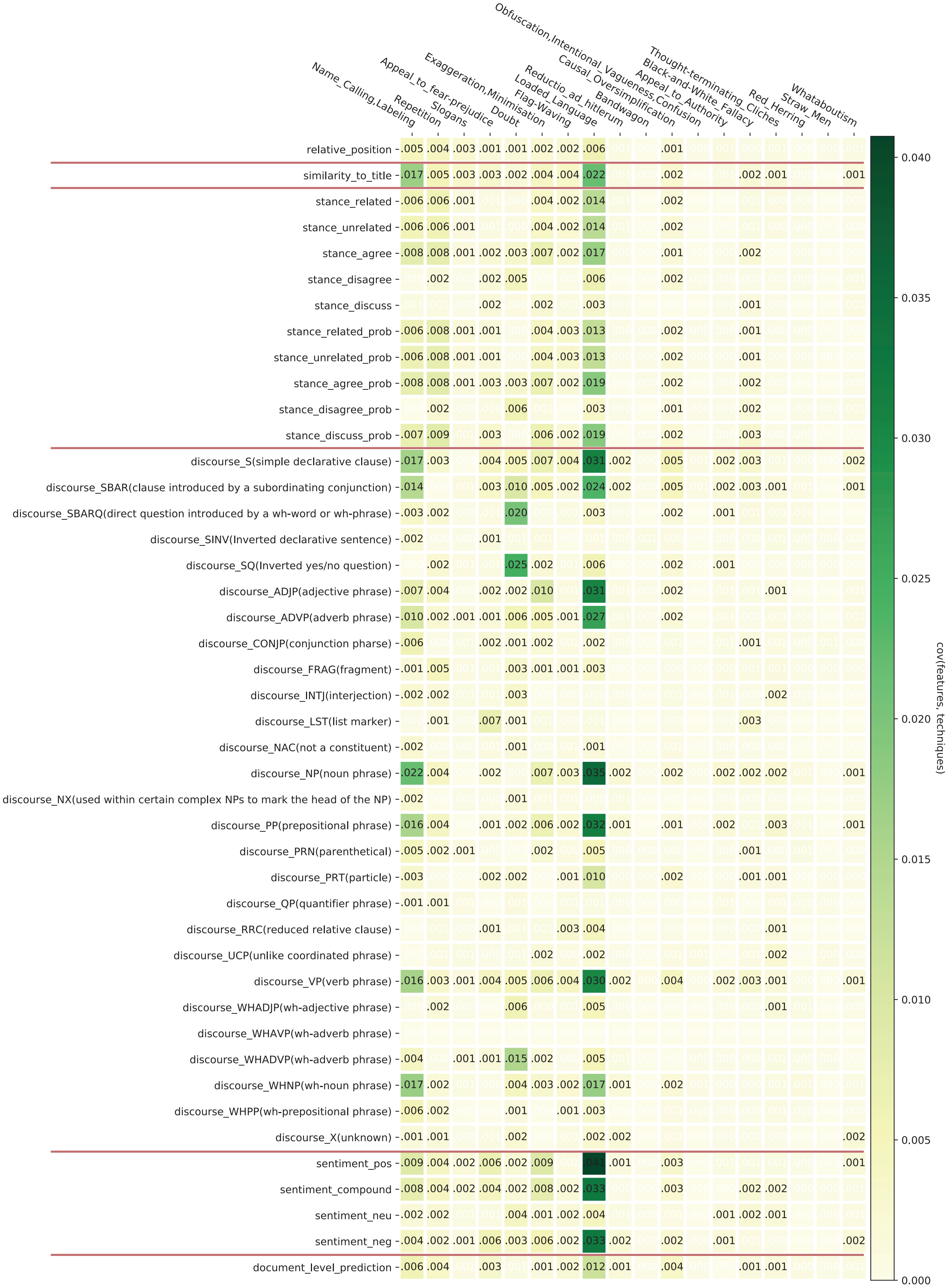


Figure 2: Covariance matrix between the 18 propaganda techniques and the proposed features.

Model	P	R	F1
fine-tuned BERT[1]	**63.20**	53.16	57.74
MGN[1]	60.41	61.58	60.98
Proposed	40.97	73.27	52.55
Proposed w/ emb	49.41	80.87	61.34
Proposed w/ emb - $\boldsymbol{f}^{stn}$	49.59	**81.44**	**61.64**

Table 3: Comparison of our method to pre-existing propaganda detection models at the sentence level for binary classification (*propaganda* vs. *non-propaganda*). The models flagged with [1] are described in (Da San Martino et al., 2019b).

Ablations	Precision	Recall	F1
All	40.97	73.27	52.55
- $\boldsymbol{f}^{rp}$	40.87	73.17	52.45
- $\boldsymbol{f}^{sim}$	40.85	70.87	51.83
- $\boldsymbol{f}^{stn}$	40.07	69.62	50.86
- $\boldsymbol{f}^{dp}$	37.85	61.54	46.87
- $\boldsymbol{f}^{sent}$	30.53	77.69	43.83

Table 4: Ablation study for our model on binary propaganda detection at the sentence level.

We followed the original data split when training and testing the model, which is 14,137/2,006/3,967 for training/development/testing. We trained a Support Vector Machine (SVM) model[2] using the above-mentioned features and we optimized the values of the hyper-parameters on the development dataset using grid search. We used an RBF kernel with gamma={1e-3, **1e-4**} and C={10,**100**}.

We can see in Table 3 that our proposed model, which is based on interpretable features, performs relatively well when compared to fine-tuned BERT without direct semantic information about the target sentence. While our model is not state-of-the-art by itself, we managed to outperform the existing models and to improve over the state of the art by simply adding to it sentence embeddings as features (Reimers and Gurevych, 2019), which were not fine-tuned on propaganda data. However, when the stance of the sentence and the embedding of the sentence are used together, performance decreases. This may be due to the two techniques based on semantic similarity being somewhat inconsistent.

4.3 Ablation Study

Next, we performed an ablation study of the binary (propaganda vs. non-propaganda) model discussed in Section 4.2. The results are presented in Table 4. The values in the last row of the table, i.e., - $\boldsymbol{f}^{sent}$, are obtained by applying the document-level classifier, i.e., the feature $\boldsymbol{f}^{doc}$, to all sentences. We can see that the structural information about the sentence ($\boldsymbol{f}^{dp}$) is the best feature for this task. This is due to the nature of some propaganda techniques that must have a specific sentence structure, such as *Doubt*. In addition, as described above, since there are many techniques related to inducing emotional responses in the readers, it can be understood that the sentiment of a sentence may be a good feature, e.g., for *Loaded Language*. These results are consistent with our findings in Section 4.1 above. Moreover, the novel features we devised based on a human behavioral study for propaganda detection ($\boldsymbol{f}^{rp}$, $\boldsymbol{f}^{sim}$, $\boldsymbol{f}^{stn}$) improved the performance further. Overall, we can see in the table that all features contributed to the performance improvement.

4.4 Detecting the 18 Propaganda Techniques

For the experiments described in the following, we revert back to the task formulation in Section 2, but we perform a more detailed analysis of the outcome of the model: for a given article, the system must predict whether each sentence uses propaganda techniques, and if so, which of the 18 techniques in Table 1 it uses.

Table 5 shows the performance of our model on this task. We can see in the rightmost column that some techniques appear only in a very limited number of examples, which explains the very low results for them, e.g., for *Red Herring* and *Straw Man*. In an attempt to counterbalance the lack of gold labels for some of the techniques, we used sentence embeddings with the proposed features to capture more semantic information. Since this task is more challenging than the binary classification problem, we can intuitively expect a performance reduction, resulting in a weighted average F1 score of 42.88. However, this formulation of the problem has the advantage of providing more granular predictions, thus enriching the propaganda detection results.

[2] Ran on Intel Xeon E5-1620 CPU @ 3.60GHz x 4; 16GiB DDR3 RAM @ 1600MHz.

Techniques	**P**	**R**	**F1**	**#**
Non-propaganda	94.37	36.62	52.77	2,927
Name Calling	14.16	21.92	17.20	146
Repetition	4.60	5.59	5.05	143
Slogans	3.75	20.69	6.35	29
Appeal to F.	12.99	38.37	19.41	86
Doubt	5.97	34.85	10.20	66
Exaggeration	6.06	20.90	9.40	67
Flag-Waving	10.98	44.62	17.63	65
Loaded L.	32.80	20.13	24.95	303
Reduction	8.00	22.22	11.76	9
Bandwagon	0.00	0.00	0.00	3
Casual O.	4.03	27.27	7.02	22
O, I, C	0.00	0.00	0.00	5
Appeal to A.	1.32	13.04	2.39	23
B&W fallacy	0.89	4.55	1.49	22
T. clichés	3.67	44.44	6.78	18
Red Herring	0.00	0.00	0.00	11
Straw Men	0.00	0.00	0.00	1
Whataboutism	2.54	14.29	4.32	21
weighted avg	73.59	32.80	**42.88**	3,967

Table 5: Performance of our proposed method for the task of detecting the 18 propaganda techniques, as evaluated at the sentence level.

5 Related Work

Research on propaganda detection has focused on analyzing textual content (Barrón-Cedeno et al., 2019b; Rashkin et al., 2017; Da San Martino et al., 2019b,a; Yu et al., 2019; Da San Martino et al., 2020b). Rashkin et al. (2017) developed the `TSHP-17` corpus, which uses document-level annotation with four classes: *trusted*, *satire*, *hoax*, and *propaganda*. They trained a model using word n-gram representation and reported that the model performed well only on articles from sources that the system was trained on. Barrón-Cedeno et al. (2019b) developed the `QProp` corpus with two labels: *propaganda* vs. *non-propaganda*. They also experimented on `TSHP-17` and `QProp` corpora, where for the `TSHP-17` corpus, they binarized the labels: *propaganda vs.* any of the other three categories. Similarly, Habernal et al. (2017, 2018) developed a corpus with 1.3k arguments annotated with five fallacies, including *ad hominem*, *red herring*, and *irrelevant authority*, which directly relate to propaganda techniques. Moreover, Saleh et al. (2019) studied the connection between hyperpartisanship and propaganda.

A more fine-grained propaganda analysis was proposed by Da San Martino et al. (2019b), who developed a corpus of news articles annotated with 18 propaganda techniques which was used in two shared tasks: at SemEval-2020 (Da San Martino et al., 2020a) and at NLP4IF-2020 (Da San Martino et al., 2019a). Subsequently, the Prta system was released (Da San Martino et al., 2020c), and improved models were proposed, addressing the limitations of transformers (Chernyavskiy et al., 2021). The Prta system was used to perform a study of COVID-19 disinformation and associated propaganda techniques in Bulgaria (Nakov et al., 2021a) and Qatar (Nakov et al., 2021b). Finally, multimodal content was explored in memes using 22 fine-grained propaganda techniques (Dimitrov et al., 2021a), which was also used in a SemEval-2021 shared task (Dimitrov et al., 2021b).

6 Conclusion and Future Work

We proposed a model for interpretable propaganda detection, which can explain which sentence in an input news article is propagandistic by pointing out the propaganda techniques used, and why the model has predicted it to be propagandistic. To this end, we devised novel features motivated by human behavior studies, quantitatively deduced the relationship between semantic or syntactic features and propaganda techniques, and selected the features that were important for detecting propaganda techniques. Finally, we showed that our proposed method can be combined with a pre-trained language model to yield new state-of-the-art results.

In future work, we plan to expand the dataset by creating a platform to guide annotators. The dataset will be updated continuously and released for research purposes.[3] We also plan to release an interpretable online system, with the aim to foster a healthier and safer online news environment.

Acknowledgements

This research is part of the Tanbih mega-project,[4] which aims to limit the impact of "fake news", propaganda, and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking. It is developed in collaboration between the Qatar Computing Research Institute, HBKU and the MIT Computer Science and Artificial Intelligence Laboratory.

[3] http://propaganda.qcri.org/

[4] http://tanbih.qcri.org/

References

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020a. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 4982–4991.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020b. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 3364–3374.

Alberto Barrón-Cedeno, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019a. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI '19, pages 9847–9848.

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

Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications*, 10(1):7.

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

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

Giovanni Da San Martino, Alberto Barron-Cedeno, and Preslav Nakov. 2019a. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the 2nd Workshop on NLP for Internet Freedom (NLP4IF): Censorship, Disinformation, and Propaganda*, NLP4IF '19, pages 162–170, Hong Kong, China.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence*, IJCAI-PRICAI '20, pages 4826–4832, Yokohama, Japan.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeno, and Preslav Nakov. 2020c. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the Annual Meeting of Association for Computational Linguistics*, ACL '20, pages 287–293.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 5636–5646, Hong Kong, China.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA.

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, pages 70–98.

Wei Fang, Moin Nadeem, Mitra Mohtarami, and James Glass. 2019. Neural multi-task learning for stance prediction. In *Proceedings of the Second Workshop on Fact Extraction and VERification*, FEVER '19, pages 13–19, Hong Kong, China.

Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. *Berkman Klein Center Research Publication*, 6.

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

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018. Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC '18, pages 3329–3335, Miyazaki, Japan.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 1859–1874, Santa Fe, New Mexico, USA.

Renee Hobbs and Sandra McGee. 2014. Teaching about propaganda: An examination of the historical roots of media literacy. *Journal of Media Literacy Education*, 6(2):5.

Philip N Howard and Bence Kollanyi. 2016. Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU referendum. *Available at SSRN 2798311*.

Clayton Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Garth S Jowett and Victoria O'Donnell. 2012. What is propaganda, and how does it differ from persuasion. *Propaganda & Persuasion*, pages 1–48.

David M.J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Farhad Manjoo. 2013. You won't finish this article: Why people online don't read to the end.

Clyde R Miller. 1939. The techniques of propaganda. From "how to detect and analyze propaganda," an address given at town hall. *The Center for learning*.

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.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, pages 1165–1174.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

Nicholas J O'Shaughnessy. 2004. *Politics and propaganda: Weapons of mass seduction*. Manchester University Press.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pages 231–240, Melbourne, Australia.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 3982–3992, Hong Kong, China.

Abdelrhman Saleh, Ramy Baly, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mitra Mohtarami, Preslav Nakov, and James Glass. 2019. Team QCRI-MIT at SemEval-2019 task 4: Propaganda analysis meets hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, SemEval '19, pages 1041–1046, Minneapolis, Minnesota, USA.

Robyn Torok. 2015. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming. In *Proceedings of the Australian Security and Intelligence Conference*, pages 58–65, Perth, Australia.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Anthony Weston. 2018. *A rulebook for arguments*. Hackett Publishing.

Seunghak Yu, Giovanni Da San Martino, and Preslav Nakov. 2019. Experiments in detecting persuasion techniques in the news. In *Proceedings of the NeurIPS 2019 Joint Workshop on AI for Social Good*, NeurIPS '19, Vancouver, Canada.