INDEPROP: Information-Preserving De-propagandization of News Articles (Student Abstract)

Aaryan Bhagat[*1], **Faraaz Mallick**[*1], **Neel Karia**[*2], **Ayush Kaushal**[*3]

[1] Indian Institute of Technology, Kharagpur
[2] Microsoft Research
[3] The University of Texas at Austin
{aaryan.bhagat, faaazrm}@iitkgp.ac.in, t-neelkaria@microsoft.com, ayushk4@utexas.edu

Abstract

We propose INDEPROP, a novel Natural Language Processing (NLP) application for combating online disinformation by mitigating propaganda from news articles. INDEPROP (**In**formation-Preserving **De**-**prop**agandization) involves fine-grained propaganda detection and its removal while maintaining document level coherence, grammatical correctness and most importantly, preserving the news articles' information content. We curate the first large-scale dataset of its kind consisting of around $1M$ tokens. We also propose a set of automatic evaluation metrics for the same and observe its high correlation with human judgment. Furthermore, we show that fine-tuning the existing propaganda detection systems on our dataset considerably improves their generalization to the test set.

Introduction

Propaganda is the expression of an opinion or an action by individuals or groups deliberately designed to influence the opinions or the actions of other individuals or groups with reference to predetermined ends (Miller 1939). With the rise of digital media, it has become extremely easy to set up independent news agencies and promulgate propaganda. As a step towards combating this, Martino et al. (2019) proposed a fine-grained propaganda detection task and curated its first dataset - QCRI. To make online news safe for public consumption, however, detection is only the first step. With this motivation, we propose the novel task - INDEPROP, where we aim to rewrite news articles to reform the propagandistic spans, while preserving the information content along with grammatical fluency and document coherence. We present a large-scale dataset for enabling research in this direction and demonstrate two of its applications - de-propagandizing text and improving fine-grained propaganda detection.

Dataset

Our dataset is derived from the QCRI dataset (Martino et al. 2019), which contains 451 articles from 49 news outlets labeled for propagandistic spans across eighteen propaganda types. In this section, we explain how we enhanced the QCRI dataset by adding de-propagandized counterparts.

Annotation Strategies

We focus on eighteen propaganda types and conduct a pilot study for analyzing the methods to modify these spans while preserving the content. A few examples of de-propagandization are shown in Table 1. Some of the ideas used to mitigate propagandistic text spans are:

- **Toning down target spans:** In cases of *loaded language*, *exaggeration* etc., we put emphasis on subduing the sentiment of the target spans by replacing strong adjectives and adverbs with their weaker counterparts.
- **Editing target spans:** We modify the target spans in various ways to remove propaganda techniques. E.g., in case of *black-and-white fallacy* and *causal oversimplification*, we edit the span to include the possibility of other alternatives thus removing the dichotomy. In case of *bandwagon*, we remove the part of the target span, which persuades readers to take the side of the masses.
- **Removing target spans:** We remove target spans (e.g. in *straw man*, *red herring*, etc.), or their segments (e.g. in *reductio ad hitlerum*, *appeal to authority*, etc.) to de-propagandize the text.

Annotation Procedure and Dataset Statistics

The annotations for de-propagandization were done by the authors, based on the aforementioned strategies. We explored crowdsourcing as a possible method for annotation, but observed the quality to be mediocre, as seen in the case of the QCRI dataset. To ease the process, we highlighted the propagandistic spans in our annotation interface and provided Masked Language Modelling suggestions. We also experimented with SpanBert for smaller spans, however, it made the tool more difficult to use during the annotation. Annotations were carried out in batches of ten documents per annotator at a stretch. After each batch, these were then verified by three other annotators, to maintain uniformity. The procedure took over 285 hours.

In the resulting INDEPROP dataset, we maintain the train-dev-test split over documents according to the QCRI dataset. Our new large-scale dataset contains 940963 tokens across 39147 sentences. A total of 26.8% of the sentences were edited and 0.11% were dropped. The de-propagandized documents were about 0.32% shorter than their propaganda-containing counterparts.

[*] Equal contribution

Type	Original	De-Propagandized
Exaggeration	But the dramatic act of extending his hand...	But the act of extending his hand...
Bandwagon	Kritarch Patti Saris like many other Federal judges does not like...	Kritarch Patti Saris does not like...

Table 1: De-propagandization

Models	BLEU-S	BLEU-R	METEOR	GLEU
T5-Small	55.18	52.50	55.07	53.14
T5-Base	68.28	64.85	70.79	70.25
Bart-Base	91.63	88.22	91.00	88.64
References	90.08	100.0	99.99	100.0

Table 2: Test performance on the INDEPROP dataset

Propaganda Mitigation Systems

We adopt some of the state of the art methods of text style-transfer, like BART (Lewis et al. 2020), as baselines. We also consider a pre-trained language model, T5 (Raffel et al. 2019), for conditional generation from the propagandistic source sentences. We use the metric GLEU, which has been widely adopted for evaluating style-transfer and grammatical error correction tasks. For a correction candidate C with a corresponding source S and reference R, it is formulated as follows:

$$GLEU(C,R,S) = BP \cdot \exp(\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \log p_i^*)$$

$$p_i^* = \frac{\sum_{n \in \{C \cap R\}} c_{C,R}(n) - \sum_{n \in \{C \cap S\}} max[0, c_{C,S}(n) - c_{C,R}(n)]}{\sum_{n \in C} c(n)}$$

Here *BP* is a normalizing factor for $|C|$ and $|R|$, and $c_{A,B}()$ denotes the number of matching n-grams in A and B. We also measure our results using the BLEU (Papineni et al. 2002) and METEOR (Banerjee and Lavie 2005) metrics. METEOR is a unigram matching based metric; BLEU-S and BLEU-R measure similarity to source and annotated documents respectively.

Table 2 shows the results of the models considered. Bart-Base outperforms all other models on the all metrics. For this model, both BLEU-S and BLUE-R scores are high. This indicates that the output is not too far from the source as well as the annotated documents, indicating the preservation of semantic information.

Enhancing Propaganda Detection

Here, we demonstrate the usefulness of our dataset for improving fine-grained propaganda detection on the QCRI dataset. Specifically, while training a detection model on the QCRI dataset, we sample contrastive de-propagandized examples from the INDEPROP dataset and then compare it with vanilla training. Table 3 shows the results of these experiments on the test set across five state of the art detection models {Bert, Bert-Joint, Bert-Granu, MGN-ReLU, MGN-Sigmoid} (Martino et al. 2019). We observe that using our dataset leads to performance gains across all the models. This demonstrates that our dataset can also be leveraged to augment the QCRI dataset for improved propaganda detection.

Models	F_1 Score	
	QCRI Only	QCRI+INDEPROP
Bert	21.11	21.42
Bert-Joint	21.16	22.97
Bert-Granu	20.45	22.03
MGN-ReLU	22.72	22.87
MGN-Sig	22.67	22.78

Table 3: Performance on QCRI Detection task on test set

Conclusion

We propose the novel NLP task of INDEPROP as the next step towards making online information safe for public consumption. In order to accelerate research in this direction, we curate a dataset of $\approx 1M$ tokens. We propose systems for the novel task and a set of evaluation metrics. Finally, we illustrate the value of our dataset, by improving fine-grained detection task, through augmentations. Our work paves the way for pursuing research on improving both propaganda detection and de-propagandization systems, as well as to study its implications with reference to free speech.

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Martino, G. D. S.; Yu, S.; Barrón-Cedeño, A.; Petrov, R.; and Nakov, P. 2019. Fine-Grained Analysis of Propaganda in News Articles. arXiv:1910.02517.

Miller, C. R. 1939. The Techniques of Propaganda. From "How to Detect and Analyze Propaganda," an address given at Town Hall. *The Center for learning*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.