Detecting propaganda and its impact on users

Oana Balalau[1], Théo Galizzi[1], Oana Goga[2], Roxana Horincar[3], and Ioana Manolescu[1]

[1]Inria, Institut Polytechnique de Paris
[2]CNRS, Institut Polytechnique de Paris
[3]Thales Research

Abstract

Political discussions revolve around ideological conflicts that often split the audience into two opposing parties. Both parties try to win the argument by bringing forward information. However, often this information is misleading, and its dissemination employs propaganda techniques. In this work, we present an investigation on the impact of propaganda on English speaking forums, and our ongoing research on providing tools for propaganda detection in French texts.

Context. Propaganda, translated from Latin as "things that must be disseminated", represents information intended to persuade an audience to accept a particular idea or cause by using specific strategies or stirring up emotions. In this short paper, we present past [1], present and future work leveraging high quality annotated datasets of propaganda techniques [3] to understand the impact of propaganda on online conversations. Our ongoing work focuses on providing a tool for propaganda classification in French texts.

Propaganda techniques. While the definition of propaganda has reached consensus in the literature, the complete list of techniques considered propagandist are still under discussion, Wikipedia[1] mentioning 68 of them. We adhere to the hypothesis previously made by [2, 3] that argues that *propaganda is a communication technique that does not depend on the document topic and its topic-specific vocabulary and for which representations based on writing style, readability, and stylistic features generalize better than word-level based representations.* [3] chooses to investigate a curated list of eighteen propaganda techniques found in journalistic articles that can be judged intrinsically, without the need to retrieve supporting information from external resources. Many of these techniques are also fallacies (arguments where the evidence does not support the claims), since propagandists use arguments that are sometimes convincing and not necessarily valid. The other techniques employ emotional language or use rhetorical, psychological, and disinformation strategies to present an idea. We leverage the list of techniques proposed in [3], illustrated in Table 1.

Propaganda detection. The dataset introduced in [3] consists of news articles manually annotated with propaganda techniques. Based on this dataset, we define two classification tasks: *i*) propaganda identification, which predicts if a sentence contains any propaganda techniques and *ii*) propaganda technique identification, which given a sentence containing propaganda, predicts the type of technique. The best classification results we obtained in [1] are of 60.98% F1 score on propaganda identification, and of 23.95% macro F1 on propaganda technique identification using transformer models [5].

Contributions. The contributions we brought [1] to the study of propaganda on English speaking forums address the following research questions: *i*) Who is posting propaganda? *ii*) How does propaganda differ across the political spectrum or different countries? and *iii*) How is propaganda received on political forums? We believe we are the first to investigate these important questions in forums with different political leaning. For the first question, we find that media sources' political bias is a strong indicator of the tendency of using propaganda and that a smaller community of users is disproportionately spreading propagandistic

[1]https://en.wikipedia.org/wiki/Propaganda_techniques, visited October 2020

Appeal to authority (fallacy) cites an expert's opinion to support an argument, without any other supporting evidence.
Appeal to fear or prejudice (fallacy) supports a claim by increasing fear towards an alternative, possibly based on preconceived judgments.
Bandwagon (argumentum ad populum fallacy) persuades the audience that a claim is true because many people believe so.
Black and white fallacy presents only two choices out of many available, with the choice on the agenda as being the better one.
Causal oversimplification (fallacy of the single cause) assumes only one cause for a complex issue out of many possible ones.
Flag waving (fallacy) exploits strong patriotic feelings for a group or idea to justify an action or a claim.
Name calling or labeling uses names, labels, or euphemisms to construct a good/bad image of a group or idea that is to be supported/denounced.
Red herring (fallacy) presents an irrelevant, although possible convincing argument to divert the attention from the matter at hand.
Reductio ad Hitlerum (fallacy) persuades the target audience to disapprove of a claim by associating it with a group widely held in contempt.
Straw man (fallacy) addresses and refutes a superficially similar claim instead of the real one.
Whataboutism (fallacy) discredits the opponent's claim by accusing them of hypocrisy without directly addressing the original argument.
Doubt questions the credibility of an idea by disseminating negative information about it.
Exaggeration / minimization makes the reality look more meaningful / more insignificant than it is.
Loaded language uses words and phrases with substantial emotional implications.
Obfuscation, intentional vagueness, confusion (ambiguity fallacy) deliberately employs vague generalities leaving the audience to draw its interpretations.
Slogans make use of brief and striking phrases to deliver the intended message.

Table 1: Propaganda techniques.

articles. Regarding the second question, we find that forums dedicated to less popular parties in a country are more likely to post biased news and that cultural differences might dictate which propaganda techniques are employed. Finally, we find that if a submission or comment has more propaganda content, it might receive more user engagement, measured either as the number of comments or as upvotes and downvotes.

Current research: multilingual propaganda detection. We currently investigate multilingual classification models that leverage the English labelled dataset for identifying propaganda in French. Observe that some propaganda categories might need new labelled data, such as the "loaded language" category, where words with a high emotional load might differ according to culture. However, we believe that the majority of other classes can be easily matched to French text. Our long-term goal is the analysis of propaganda in French media, and in paid political ads, whose detection is far from trivial [4]. This would entail automatically analysing many newspapers and advertisements and to identify trends. We also hope to identify exploitable elements, that is, general guidelines allowing readers to more reasily detect bias.

References

[1] Oana Balalau and Roxana Horincar. From the stage to the audience: Propaganda on Reddit. In *EACL*, 2021.

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

[3] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news article. In *EMNLP*, pages 5636–5646, 2019.

[4] Vera Sosnovik and Oana Goga. Understanding the complexity of detecting political ads. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *WWW*, pages 2002–2013. ACM / IW3C2, 2021.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.