# Detecting Propaganda in Online Media

**Giovanni Da San Martino**
Qatar Computing Research Institute, HBKU
Doha, Qatar.
gmartino@hbku.edu.qa

## Abstract

Many recent political events, like the 2016 U.S.A. or the 2018 Brazilian elections have raised the attention of institutions and the general public on the role of internet and social media in influencing the outcome of these events. Understandably, the aspect of such campaigns that has caught more attention is the spread of fake news. However, faking facts is only one of the many devices that can be used to reach the ultimate goal of persuading and influencing readers' opinions, i.e. to set a propaganda campaign: many psychological and rhetorical techniques are known in literature. This talk aims at giving an overview of Natural Language Processing studies on automatic detection of the use of propaganda in texts.

## 1 Detecting the Use of Propaganda in Online Media

Propaganda aims at influencing people's mindset with the purpose of advancing a specific agenda. In the previous century, massive propaganda campaigns were mostly prerogative of governments and large institutions and were mostly directed to their own citizens and used to consolidate power or support possibly unpopular political decisions, e.g. starting a war.

The advent of the Internet has affected profoundly the way political agendas and ideological messages are spread to large audiences. Social media and messaging apps may be exploited not only by large institutions and governments, but also by small organisations or individuals to reach an audience of unprecedented size. This, as observed in Bolsover and Howard (2017), has opened new scenarios for propaganda campaigns: "It has allowed cross-border computational propaganda and interference in domestic political processes by foreign states. The anonymity of the Internet has allowed state produced propaganda to be presented as if it were not produced by state actors". Examples of the novel types of propaganda campaigns allegedly happened during the 2016 US, the 2018 Brazilian and the 2018 Mexican presidential elections (Muller, 2018; Tardáguila et al., 2018; Glowacki et al., 2018), the 2016 UK–European Union Referendum (Bastos and Mercea, 2017). Events like the infamous "pizzagate"[1] demonstrated the real-life consequences of fake news also for individuals and brought the attention of the public on the problem.

However, in this talk we argue that faking news is only one of the means to reach the ultimate goal of persuading someone and advocate for a broader view on the problem. Indeed there exist a number of rhetorical techniques, mostly logical fallacies, and techniques appealing to the emotions of the audience (Torok, 2015; Weston, 2000). Logical fallacies are usually hard to spot since the argumentation, at first sight, might seem correct and objective. However, a careful analysis shows that the conclusion can not be drawn from the premise without the misuse of logical rules. Another set of techniques makes use of emotional language to induce the audience to agree with the speaker only on the basis of the emotional bond that is being created, provoking the suspension of any rational analysis of the argumentation. All of these techniques are intended to go unnoticed to achieve maximum effect (Miller, 1939). Studies have shown that even educated users can be fooled and have their thoughts driven towards some end. As a result, malicious propaganda news outlets are potentially able to achieve large-scale impact.

This talk discusses the detection of propaganda at media source and document level (Rashkin et al., 2017; Barrón-Cedeño et al., 2019). One of

[1] https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory

the main issues affecting the research on propaganda detection is the extreme scarcity of quality gold labels. We discuss the resources available and how they affect the research. Rashkin et al. (2017) deal with a document categorisation problem in which propaganda is one out of three other categories: trusted articles, satire, hoax. In order to obtain gold labels for a document, they use distant supervision: the categorisation of a media source performed by expert journalists[2] is transferred to all the articles published by that media source. The noise introduced by this labelling process calls for specific techniques to avoid learning algorithms to model the source instead of the category (Barrón-Cedeño et al., 2019).

Works related to the detection of the use of propaganda in text fragments, i.e. the detection of logical fallacies and techniques appealing to the emotions, are then discussed (Habernal et al., 2017, 2018a,b). Given the difficulty of obtaining quality fragment-level annotations, and therefore the limited size of the datasets available, deep learning models, which are currently successfully on a number of Natural Language Processing problems, cannot be straightforwardly applied. We discuss the possibility of using learning algorithms able to use external knowledge or techniques such as transfer learning.

## References

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: Organizing news coverage on the basis of their propagandistic content. *Information Processing and Management*.

Marco T. Bastos and Dan Mercea. 2017. The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, page 0894439317734157.

Gillian Bolsover and Philip Howard. 2017. Computational Propaganda and Political Big Data: Moving Toward a More Critical Research Agenda. *Big Data*, 5(4):273–276.

Monika Glowacki, Vidya Narayanan, Sam Maynard, Gustavo Hirsch, Bence Kollanyi, Lisa-Maria Neudert, Phil Howard, Thomas Lederer, and Vlad Barash. 2018. News and political information consumption in mexico: Mapping the 2018 mexican presidential election on twitter and facebook. Technical Report COMPROP DATA MEMO 2018.2, Oxford University, Oxford, UK.

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*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018a. Adapting Serious Game for Fallacious Argumentation to German: Pitfalls, Insights, and Best Practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3329–3335.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Robert Muller. 2018. Indictment of Internet Research Agency.

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 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP 17, pages 2931–2937, Copenhagen, Denmark.

Cristina Tardáguila, Fabrício Benevenuto, and Pablo Ortellado. 2018. Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It. `https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html`.

Robyn Torok. 2015. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming. 2015:58–65.

Anthony Weston. 2000. *A Rulebook for Arguments*.

[2] `https://mediabiasfactcheck.com/`