Undiplomatic diplomats: building an NLP tool for propaganda detection on Twitter

Keywords: propaganda, natural language processing, digital diplomacy, Twitter, rhetoric

Extended Abstract

According to Sparkes-Vian (2019), propaganda is "an evolving set of techniques and mechanisms which facilitate the propagation of ideas and actions". Unlike fake news, the automated detection of propaganda on social media has not attracted so much attention from journalists, fact-checkers, or scholars. In our view, this hinders the endeavors against manipulative information. The deceiving intent of propaganda may be more subtle and devious than disinformation; its content does not have to be false, and its effects may be only discernible through systematic observation over time.

This study relies on natural language processing (NLP) models to propose a novel tool for propaganda detection. Previous works that tried to automatedly detect online propaganda have rested on San Martino et al (2019), who classify propagandistic segments within news articles depending on the propaganda technique they contain. Our approach remains complementary to their study, as we also seek to detect and categorize propaganda techniques, although we re-orient the study towards Twitter and international politics. Our corpus encompasses tweets in English from authorities of four different international actors: China, Russia, United States, and the European Union. The authorities collected include government accounts, embassies, ambassadors, and other diplomatic profiles like consuls or missions in international organizations.

Twitter provides for governments a direct channel for influencing international audiences. In this endeavor, as it could be observed during the COVID-19 pandemic, some authorities from the US, Russia and China showed an undiplomatic, confrontational behavior that tried to undermine the reputation of their adversaries, even spreading conspirative content, hate speech and fake news. The growing concern on the activity of certain governments led Twitter, in August 2020, to label the accounts of key officials, publicly linking each user with his country. But whereas this decision may have been a good first step to provide useful context on these accounts, our research aims to provide more detailed information on the language of public authorities. On the one hand, our proposed tool seeks to help audiences to detect propagandistic messages and act critically, and on the other hand it can be useful for digital diplomacy and international relations' studies.

Our selection of propaganda techniques was inspired by San Martino et al (2019). We grouped various of the techniques they used into four clusters to improve efficiency, decrease complexity and try to get better computational results. To synthesize the task, we first consider that most of the techniques involve either emotional language or logical fallacies. Therefore, as suggested by Miles (2019) our grouping is inspired by Aristotle's principles of rhetoric: ethos (appealing to the authority and credibility of the person speaking), pathos (appealing to emotions) and logos (appealing to reasoning and logical arguments). Besides, we also considered that diplomatic propaganda is used to improve existing relationships with allies, to facilitate favorable relationships with neutrals and to maintain or increase the desired hostilities

with enemies. Thus, when clustering techniques we further divided the techniques corresponding to the logos principle depending on the target of the rhetoric. Following Bjola & Pamment (2019), we distinguish between propaganda with "constructive" results, in which the receiver is influenced to voluntarily make decisions favorable to the controlling side, or "destructive" results, aspiring to affect a rival's reputation. Table 3 provides below a detailed list with the 14 techniques considered. These techniques were finally clustered in the following four groups:

- Appeal to Commonality (logos, constructive)
- Discrediting the opponent (logos, destructive)
- Loaded Language (pathos)
- Appeal to authority (ethos)

For the annotation process, we collected a dataset with 12,000 tweets in English published by 106 Chinese, 114 Russian, 186 European and 216 American authorities between January 1st, 2020, and March 11th, 2021, coinciding this last date with the first anniversary of the declaration of the COVID-19 pandemic by the World Health Organization. The annotation process consisted of, first, marking the technique (or techniques) identified in the tweet, and then, annotating which of the four clusters it corresponds to accordingly. If a cluster was annotated, that tweet was then marked as propagandistic. We split the data into 80% training and 20% test.

The NLP task that this presentation will expose consist thus of a twofold exercise. Given a tweet, our model must first identify whether it contains propaganda or not; and second, classify the propagandistic tweet within a specific cluster. For that purpose, we will exploratorily apply baselines given by the Roberta-large (Liu, 2019) and Deberta (He et al, 2021). The quantitative results of this study will shed light on the complexity and manageability of this problem in computational terms, and it will also set the bases for an unexplored line of research. The qualitative results inferred from the annotated dataset (see tables 1 and 2) disclose that China and the US were the countries whose authorities diffused a higher proportion of propagandistic tweets: 28% and 23%, respectively. In the case of the US, most of these tweets came from top officials of the Trump Administration, including the president. Concerning the propagandistic techniques employed, Loaded Language was the most annotated category for Russian, American and European authorities. Discrediting the opponent also stood out in the tweets of Chinese, American and Russian diplomats, which seems to suggest a recurrent confrontational rhetoric by some of their representatives. Appeal to commonality techniques frequently appeared in tweets linked to the history and values of these countries.

References

- Bjola, C. & Pamment, J. (2019) Countering online propaganda and extremism. The dark side of digital diplomacy. Routledge.
- He, P. Liu, X. Gao, J. Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. https://arxiv.org/abs/2006.03654
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [Cs]. http://arxiv.org/abs/1907.11692
- Miles, C, (2019). "Rhetorical Methods and Metaphor in Viral Propaganda" in Baines, P., O'Shaughnessy, N., Snow, N. (eds) The SAGE Handbook of Propaganda. SAGE.
- San Martino, G., et al (2019). Fine-Grained Analysis of Propaganda in News Article. In Proc. 2019 EMNLP-IJCNLP, 5636–5646, Hong Kong, China. ACL.
- Sparkes-Vian, C. (2019). Digital Propaganda: The Tyranny of Ignorance. Critical Sociology, 45(3), 393-409. https://doi.org/10.1177/0896920517754241

Table 1: Number of annotated tweets that contained propaganda techniques associated with each group. *No propaganda* column shows the number of tweets that did not contain any technique.

Country	Appeal to Commonality	Discrediting the opponent	Loaded Language	Appeal to Authority	No propaganda
China	229	448	333	2	2778
Russia	209	246	261	0	2996
USA	289	428	449	3	3074
EU	79	11	187	0	3307

Table 2: Top propaganda spreaders by number of propagandistic tweets (Prop.). *Prop. Ratio* indicates the percentage of each user's annotated tweets that contained at least one propaganda technique:

China	**Prop.**	**Prop. Ratio**	**Russia**	**Prop.**	**Prop. Ratio**
@spokespersonchn	228	40%	@mfa_russia	144	22%
@mfa_china	114	32%	@dpol_un	104	53%
@zlj517	93	35%	@russia	96	31%
USA	Prop.	Prop. Ratio	**EU**	**Prop.**	**Prop. Ratio**
@realdonaldtrump	245	75%	@josepborrellf	38	12%
@secpompeo	214	32%	@vonderleyen	35	11%
@whitehouse45	112	35%	@eu_commission	29	6%

Table 3: List of propaganda techniques considered for this study and their assigned group

Group	Technique	Example
Appeal to commonality	Ad populum/Ad Antiquitatem	*The leadership of the #CPC is the choice of history*
	Flag Waving	*The USA is the greatest country in the history of civilization*
Discrediting the opponent	Name Calling / Labelling	*I called the politicization of the China Virus by the Radical Left Democrats a Hoax*
	Undiplomatic assertiveness / whataboutism	*Western elites are readily shouting the name of #Navalny-the-crook, but shyfully forget to mention their political prisoner Julian #Assange.*
	Reductio ad hitlerum	*This is reminiscent of the ugly history of McCarthyism. Sadly*
	Doubt	*What's behind the closure of the biolab at Fort Detrick?*
	Propaganda Slinging	*This is a piece of fake news, launched in the media space by the #US intelligence community.*
	Scapegoating	*It is not China but the US that is militarizing and stoking tensions in the South China Sea*
	Fear Appeal	*Those harming our core interests will meet countermeasures & get severe punishment from history*
	Absurdity appeal	*Is there anything left that US hasn't sanctioned yet? Maybe the dolphins that were swimming alongside the tankers? US looks increasingly pathetic and ridiculous*

	Demonization	*Biden is a corrupt globalist sellout who never missed a chance to stab American workers in the back*
Loaded Language	Loaded Language	*Democrat leaders have made their radical disregard for the U.S. Constitution*
Appeal to Authority	Appeal to false authority	*A voice of a Pakistani student's wife tells real situation about the coronavirus in China. Trust the Chinese Government. No panic!*
	Bandwagoning	*#Germany took strong action today against Hizballah. We call on #EU member states to follow suit in holding Hizballah accountable*