From the Stage to the Audience: Propaganda on Reddit

Oana Balalau, Roxana Horincar

From the Stage to the Audience: Propaganda on Reddit

Oana Balalau
Inria, Institut Polytechnique de Paris
France
oana.balalau@inria.fr

Roxana Horincar
Thales Research & Technology
France
roxana.horincar@thalesgroup.com

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 analyze the impact of propaganda on six major political forums on Reddit that target a diverse audience in two countries, the US and the UK. We focus on three research questions: *i*) who is posting propaganda? *ii*) how does propaganda differ across the political spectrum? and *iii*) how is propaganda received on political forums?

1 Introduction

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. Our work is the first study that leverages a high quality annotated dataset of propaganda techniques (Da San Martino et al., 2019) to understand the impact of propaganda on online conversations.

In this paper, we perform an in-depth and long-term analysis of propaganda on online forums. We focus on six subreddits from two English speaking countries, the US and the UK, for one year. We select a popular subreddit for political news with no party affiliation and two subreddits dedicated to each country's dominant parties. In the US, the two main parties are the Democrat and the Republican party. The Democrat party is center-left; however, it contains several factions with ideologies varying from the center to the left. The Republican Party is a center-right party and has shifted in recent years towards national conservatism. In the UK, the most popular parties are the Labour Party and the Conservative Party. Similarly to the US, these parties represent the center-left and the center-right. The Labour Party has social democratic and socialist factions, while the Conservative Party has many factions, such as one-nation conservatism, liberal conservatism, or social conservatism. In recent years, both countries passed through significant political turmoil, such as Donald Trump's election in the US and the referendum on leaving the EU in the UK. However, a recent opinion piece in the Washington Post highlights an essential difference between the political discourse in the two countries. The journalist believes that the division between the left and the right in America is driven by the different interpretations the two parties give to the words "rights", "liberty" or "freedom", which have a strong moral imperative. This difference is not present in the UK, hence political parties there might find it easier to reach common grounds.

Our contribution to the study of propaganda in online discussions is in investigating 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 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.

2 Related Work

Analysis of political discussions. (Roozenbeek and Salvador Palau, 2017) explore the role of online communities in elections and how different types of new events impact their dynamics. In (Soliman et al., 2019), the authors analyze political communities (subreddits on Reddit), comparing them to the content posted, their relationships to other subreddits, and the distribution of attention received in these subcommunities. They compare left-leaning with right-leaning communities, with significant differences emerging, such as higher use of derogatory language in the right-leaning communities, stronger connectivity between the US and the European right-leaning communities, and more substantial focus on media sources reflecting their political leaning in the left-leaning subreddits. In (Guimaraes et al., 2019), the authors identify different conversation patterns that refine the notion of controversy into disputes, disruptions, and discrepancies and perform a systematic analysis of discussion threads based on essential facets of a conversation like users, sentiments, and topics. (An et al., 2019) proposes an analytical template to explore the nature of political discussions by studying the interaction and linguistic patterns within and between politically homogeneous and heterogeneous communication spaces on Reddit. (Carman et al., 2018) analyzes the effects of vote manipulation on article visibility and user engagement by comparing political threads on Reddit whose visibility is artificially increased.

Propaganda detection. Previous works on propaganda have focused on proposing datasets to foster further research, including document-level annotations (Rashkin et al., 2017; Barrón-Cedeño et al., 2019) and fragment level annotations (Da San Martino et al., 2019). Efforts for constructing annotated datasets have also been made in other European languages different from English (Kmetty et al., 2020; Baisa et al., 2019). Automatic propaganda detection approaches are almost always proposed alongside new corpora. (Rashkin et al., 2017) defines a four-class text classification task that detects propaganda, satire, hoaxes, and real news, while (Barrón-Cedeño et al., 2019) uses a binary classification to detect propaganda and non-propaganda articles. (Da San Martino et al., 2019, 2020) perform fine-grained analysis of texts by detecting all fragments that contain propaganda techniques, as well as their type. In (Kellner et al., 2020), the authors quantify the influence of trolls on Twitter that contribute to the propaganda spread during political elections in online communities. Studies on the use of propaganda have also helped understand how terrorist organizations share their ideology and attract new members (Al-Rawi and Groshek, 2020; Bisgin et al., 2019). (Martino et al., 2020) reviews the state of the art of computational propaganda detection from both an NLP and a network analysis perspective, arguing on the need to combine these communities' efforts.

Bot detection in political discussions. Research on political discussions has mostly focused on specialized topics such as adversarial debates between two parties, like election campaigns and referendums. (Rizoiu et al., 2018; Davis et al., 2016) use machine learning approaches to study social bots' influence in the diffusion of tweets containing partisan hashtags surrounding a political debate. (Hurtado et al., 2019) studies political discussions on Reddit and uses graph-based methods to reveal a fully connected community of users who exhibit a bot-like behavior. (Costa et al., 2015) introduces a generative model based on users' temporal activity patterns to study abnormal posting behavior both on Twitter and Reddit data.

Journalistic efforts in studying online content. There have been some relevant initiatives by communities of expert journalists or volunteers to raise awareness of different online news issues by evaluating the content published by news outlets and social media. For instance, Media Bias/Fact Check (MBFC) is an independent organization that analyzes media in terms of their factual reporting, bias, and propagandist content, among other aspects. Full Fact, an independent fact-checking organization in the UK, provides free tools, information, and advice for checking claims by politicians and the media. Similar initiatives have been taken by US News and World Report and the European Union.

3 Propaganda Techniques

Propaganda is a communication technique primarily used to influence public opinion towards an a-priory established agenda.

According to the Institute for Propaganda Analysis, propaganda had its definition pinned in 1938 as being "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" (for Propaganda Analysis, 1938).

In the past century, spreading propaganda required controlling traditional journalism media, such as newsprint, TV, and radio stations. It represented a form of communication that only large institutions and governments could afford. With the recent rise of the Internet and its use as online mass media, "computational propaganda" appeared (Bolsover and Howard, 2017) as a social and technical phenomenon that made propaganda campaigns easily accessible to a wide variety of small organizations and individuals that targeted audiences of unprecedented size. Recent striking examples include the propaganda allegedly set to influence the 2016 US presidential elections (Mueller, 2018) and the 2016 Brexit referendum (Howard and Kollanyi, 2016).

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 (Barrón-Cedeño et al., 2019; Da San Martino et al., 2019) 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*. (Da San Martino et al., 2019) 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 since propagandists use arguments that are sometimes convincing and not necessarily valid. A fallacy is an argument where the evidence does not support the claim that is put forward. The other techniques employ emotional language or use rhetorical, psychological, and disinformation strategies to present an idea.

We leverage the list of eighteen propaganda techniques proposed by (Da San Martino et al., 2019).

- *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 or minimization* makes the reality look more meaningful or more insignificant than it is.
- *Loaded language* uses words and phrases with substantial emotional implications.
- *Obfuscation, intentional vagueness, confusion (ambiguity fallacy)* deliberately employs

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

vague generalities leaving the audience to draw its interpretations.

- *Repetition* repeatedly uses the same symbol or idea to make it unforgettable.
- *Slogans* make use of brief and striking phrases to deliver the intended message.
- *Thought terminating cliches* take advantage of short, generic phrases that divert the attention or seem to offer simple answers to complex problems to stop an argument from proceeding further.

4 Reddit Dataset

We select six subreddits: Politics, Democrats, Republican, UKPolitics, LabourUK, and Tories. Politics is a subreddit for "current and explicitly political U.S. news.". The subreddit does not claim any political affiliation. The Democrats subreddit description contains "We are here to get Democrats elected up and down the ballot.", and it is a partisan subreddit. Republican is "a partisan subreddit" and the place where "Republicans discuss issues with other Republicans", hence it is a subreddit for people supporting the US Republican party. UKPolitics is a forum for "political news and debate concerning the United Kingdom" and does not claim any political affiliation. LabourUK is a subreddit that discusses breaking news concerning the British Labour Party. Finally, Tories is a subreddit for news concerning the Conservative Party in the UK, also known as the Tories. When there are several subreddits on the same topic (for example, BritishPolitics is also a subreddit for politics in the United Kingdom), we select the subreddit with the largest number of members. We note that Reddit does not ask for or encourages users to share personal data, such as their locations. Statistics on Reddit users are available only through data gathered from independent polls and surveys. For example, we know that the US and UK are the best-represented countries among the Reddit users. In the light of the surveys, we hypothesize that there are many users from the US and UK that engage in political subreddits.

We take all content posted for a period of one year, January 2019 to December 2019, from the PushShift dataset (Baumgartner et al., 2020). On Reddit, a discussion is started by a **submission**, e.g. a news article or a piece of text, and users will engage by writing **comments**. A comment is described by *author*, *body* (the content of the comment), and *score* (computed as upvotes minus downvotes) among others. We remove comments tagged as "[deleted]" or "[removed]", which are comments removed by the moderators or the users themselves. A submission has several properties, including *content* (often linked via a URL), *number of comments*, *score* (upvotes minus downvotes), and *author*. For simplicity, we refer to the submission and the article linked in the submission using the term *submission*. We retrieve the external articles by following the link in the submission. We filter out the submissions whose corresponding articles were not found by the crawler, either because cookie permissions cannot be given automatically or because the link is no longer valid. We also filter out the submissions linking to articles with less than 200 words. We want to focus on journalistic like content, a piece of text large enough to develop well an idea. Overall, we keep around $43 - 71\%$ of the original submissions, depending on the subreddit.

Subreddit	Submissions	Comments
Politics	317K	20M
Democrats	9.8K	54K
Republican	8.2K	41K
UKPolitics	42K	1.8M
LabourUK	7K	58K
Tories	1.1K	12K

Table 1: Reddit dataset

An overview of our dataset is given in Table 1.

To further understand the subreddits' dynamics, we report the overlap between the users commenting or posting a submission in the forums over the period we study. In the US related forums, there are $736K$ unique users, out of which $730K$ unique users in Politics, $8.5K$ in Democrats, and $7.7K$ in Republican. We have that 75% of users in Democrats and 57% of users in Republican also post in Politics, while only 5% of users posting in Republican also post in Democrats. In the UK forums, we have $46K$ unique users, out of which $44K$ post in UKPolitics, $3.3K$ in LabourUK, and $1K$ in Tories. The overlap between the forums shows a more balanced dynamics, with 61% of the LabourUK users and 63% of the Tories users also posting in UKPolitics, and 23% of the Tories users posting in LabourUK.

5 Methodology

The dataset introduced in (Da San Martino et al., 2019) consists of news articles manually annotated with propaganda techniques. The propaganda techniques are in order of frequence of instances: loaded language (2547), name-calling (1294), repetition (767), exaggeration or minimization (571), doubt (562), appeal to fear and prejudice (367), flag-waving (330), causal oversimplification (233), slogans (172), appeal to authority (169), black and white fallacy (134), thought-terminating cliches (95), whataboutism (76), reductio ad hitlerum (66), red herring (48), bandwagon (17), labeling, obfuscation or intentional vagueness (17), straw men (15). The annotations are fined-grained, with each propaganda instance being labeled at the token level. One technique might span more than one sentence.

We define two classification tasks based on the propaganda dataset described in Section 4: *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.

For each task, we test the following classifiers: a random classifier which predicts a class uniformly at random, a suite of transformer classifier BERT (Devlin et al., 2019), ROBERTA (Liu et al., 2019) and XLNet (Yang et al., 2019), and an ensemble classifier that makes a prediction based on the most confident label given by one of the three classifiers (BERT, ROBERTA or XLNet). Finally, we add the multi granularity model proposed in (Da San Martino et al., 2019), MGN ReLU. To fine-tune the transformer models, we add a final linear layer. We use a sequence length of 210, a learning rate of 0.01, a mini-batch of size 16, anneal factor of 0.5, patience of 2, and the maximum number of epochs to 20. To deal with dataset imbalance in both tasks, we weight the loss function samples according to the class weight.

The first task, propaganda identification, is a binary classification task, with classes propaganda and nonpropaganda. We present the results in Table 2. We note that propaganda identification is a difficult task, and all the classifiers obtain moderately good results, however much better than random selection.

The second task allowed us to understand if we have enough instances of each propaganda technique to classify them. We ran an experimental study, and we observed that bandwagon, obfuscation, red herring, straw men, and thought-terminating cliches were never recognized in the test set by our classifiers. Given this, we removed them from the annotations, and we kept the remaining techniques for the first and second tasks. We present the results in Table 3.

Classifier	Precision	Recall	F1
Random	24.14	25.65	28.87
BERT	58.52	52.02	55.08
ROBERTA	**63.96**	41.41	50.28
XLNet	53.27	59.29	56.12
Ensemble	62.72	48.57	54.74
MGN ReLU	60.41	**61.58**	**60.98**

Table 2: Precision, recall and $F1$ score for propaganda identification.

Classifier	F1-micro	F-macro
Random	17.07	15.76
BERT	29.75	22.17
ROBERTA	**26.96**	22.00
XLNet	29.07	**23.95**
Ensemble	28.17	22.71

Table 3: $F1$ score for technique identification.

Topical confounds. Finally, we study the effect of topical confounds in propaganda and technique classification. This analysis aims to understand if there are topical biases in the annotated dataset, which might bias our analysis. For example, if the data contains many articles on Trump, we might tend to label as propaganda any article referring to him. To identify topical biases, we use the approach presented in (Kumar et al., 2019). We first identify statistically overrepresented words in each propaganda technique in the training set and then replace them with a special token in the test set. The overrepresented words are computed using log-odds ratio with Dirichlet prior (Monroe et al., 2008), and we present the results in Table 4. We recall that we removed the techniques bandwagon, obfuscation, red herring, straw men, and thought-terminating cliches from our labeled dataset. As we can observe, for certain categories, the words are very intuitive. For example in reductio ad hitlerum we have many words related to totalitarian regimes, or in flag-waving we have many words around the notion of country. However, for most techniques, the words do not form cohesive topics, which is

Appeal to authority	thousands, regard, voter, altering, bea, notes, schema, muhammad, homosexual, viganò
Appeal to fear or prejudice	student, eucharist, jewish, easiest, bds, cliff, lew, eucharistic, campus, mcgill
Black-and-White Fallacy	easiest, uk, die, focus, burn, throw, dear, blessing, bless, chop
Causal oversimplification	anderson, backlash, cia, hillary, continued, god, alleging, knows, ruined, shahada
Doubt	guardian, story, wills, harding, assange, fake, claims, luke, failed, evidence
Exaggeration,Minimization	absolutely, history, worse, impossible, biggest, world, extraordinary, greatest, ukrainians, ingenious
Flag-Waving	american, america, europe, country, people, hungary, orban, nation, americans, citizen
Loaded Language	advantage, neo, devastating, democrats, shocking, lies, church, voice, grave, jews
Name calling, labeling	partisan, bergoglio, witch, hunt, dems, spy, assange, google, guardian, righteous
Reductio ad hitlerum	hitler, german, nazis, vichy, labour, communists, farrakhan, nazi, soviet, occupation
Repetition	muslim, san, entry, invaders, port, hungary, hat, inconvenient, orban, tijuana
Slogans	hat, character, jimenez, duke, school, america, sadikov, whataburger, foot, home
Whataboutism	admitting, guys, prosecute, focus, jihad, interpreted, ship, prosecuting, west, happened

Table 4: Top 10 words statistically overrepresented in each propaganda technique in the training set.

expected as propaganda represents a communication technique, and it is not restricted to a topic. To further verify that our classifiers learn style and not the topic, in the test set, we replace with a special token the top k words strongly associated with each technique, computed from the training set. For both $k = 10$ and $k = 20$ we report a very small decrease in $F1$ score for the BERT classifier in the propaganda classification task, from 55.08 to 52.47 and from 55.08 to 53.08. For the technique classification task, for $k = 10$ we do not observe a drop in performance, while for $k = 20$ we pass from 29.75 $F1$-micro score to 27.26, and from 22.17 $F1$-macro to 19.85. Besides, we note that the decrease in performance for this task is distributed among techniques. For flag waving and reductio ad hitlerum, for which certain words were important with respect to their definition, we do not observe a large decrease in $F1$ score. For example, the F1 scores for flag waving for $k = 10$ and $k = 20$ decrease from 43.98 to 39.57 and to 39.36, respectively. Given the small decrease in performance, we can conclude that our classifier does not learn topical confounds but the language patterns of propaganda techniques.

We leverage the propaganda identification classifier to define a **propaganda score**. The propaganda score of a document is the percentage of sentences that were labelled as containing propaganda. We compute the propaganda score of each submission, and based on the distribution of the score values in a subreddit, we define two groups: the **least propaganda**, which represents the 25% submissions with the lowest propaganda score, and the **most propaganda**, which represents the 25% submissions with the highest propaganda score. Our aim in defining the two groups is to mitigate part of the classifier's imprecision and make our analysis more robust.

6 Propaganda on Reddit

In this section, we focus on several research questions around propaganda on online forums.

RQ1. Who is posting propaganda? In the context of political forum discussions, this question targets two different groups: *media outlets and social media users*. The initial publishers are the media outlets, but users handpick what news to share on political forums. To study what media outlets are present and in which measure they are responsible for the propaganda content, we look at the groups defined in Section 5, least propaganda submissions, and most propaganda submissions. We compute the top-level domain for each submission in the two groups, which corresponds to the media outlet. We give each media outlet a label measuring its political leaning, according to MediaBiasFactCheck: center, left-center, right-center, left, right, questionable, and others. The center label is interpreted as having no or little political bias, left-center and right-center have a slight bias, left and right have a moderate bias, while the questionable label has a strong bias. The others label is given to sites that are not found in our dataset. MediaBiasFactCheck computes a media source's political bias taking into account bias by story selection, bias by omission, or bias by labeling, among others. In Table 5, we observe a strong relationship between the political bias of the media sources and the groups we computed using our propaganda score. Hence, we can infer that political bias often translates into the use of propaganda techniques.

Concerning users posting propaganda content on Reddit, we cannot link them to real entities; however, we can observe them as a community. We find

	Politics
Least Propaganda	(LeftCenter, 34.49%), (Center, 24.81%), (Left, 22.4%), (RightCenter, 3.08%), (Right, 2.91%), (Questionable, 1.44%), (Others, 10.88%)
Most Propaganda	(Left, 39.1%), (LeftCenter, 25.41%), (Center, 17.09%), (Right, 6.92%), (RightCenter, 3.3%), (Questionable, 3.07%), (Others, 5.11%)
	Democrats
Least Propaganda	(LeftCenter, 33.82%), (Left, 27.98%), (Center, 21.74%), (RightCenter, 2.76%), (Right, 0.53%), (Questionable, 0.41%), (Others, 12.78%)
Most Propaganda	(Left, 41.74%), (LeftCenter, 24.44%), (Center, 15.72%), (Right, 1.13%), (RightCenter, 1.05%), (Questionable, 0.4%), (Others, 15.51%)
	Republican
Least Propaganda	(Right, 35.94%), (Questionable, 23.69%), (LeftCenter, 7.67%), (RightCenter, 7.28%), (Center, 6.27%), (Left, 1.98%), (Others, 17.14%)
Most Propaganda	(Right, 41.58%), (Questionable, 29.28%), (RightCenter, 6.58%), (Left, 2.81%), (LeftCenter, 2.58%), (Center, 2.24%), (Others, 14.93%)
	UKPolitics
Least Propaganda	(LeftCenter, 47.66%), (Center, 10.42%), (Right, 3.84%), (Questionable, 2.22%), (RightCenter, 2.18%), (Left, 1.04%), (Others, 32.64%)
Most Propaganda	(LeftCenter, 40.65%), (Right, 11.31%), (Questionable, 6.11%), (Left, 5.33%), (RightCenter, 4.48%), (Center, 3.76%), (Others, 28.37%)
	LabourUK
Least Propaganda	(LeftCenter, 48.87%), (Left, 3.94%), (Center, 3.1%), (RightCenter, 1.69%), (Right, 0.96%), (Questionable, 0.45%), (Others, 40.99%)
Most Propaganda	(LeftCenter, 49.63%), (Left, 10.46%), (RightCenter, 2.7%), (Right, 1.74%), (Center, 1.24%), (Questionable, 0.73%), (Others, 33.5%)
	Tories
Least Propaganda	(LeftCenter, 47.18%), (Right, 9.86%), (Center, 4.93%), (Questionable, 2.11%), (RightCenter, 1.76%), (Left, 1.06%), (Others, 33.1%)
Most Propaganda	(LeftCenter, 28.87%), (Right, 27.11%), (Questionable, 5.63%), (RightCenter, 3.87%), (Center, 2.82%), (Left, 1.06%), (Others, 30.63%)

Table 5: There is a strong relation between the political bias of the media sources and the groups we computed using our propaganda score. For example, on Politics, the majority of media sources are *left center* in the least propaganda group, while the majority of sources are *left* in the most propaganda group.

that the most propaganda group's submissions are created by a smaller number of unique users than the submissions in the least propaganda group, on all subreddits except LabourUK, where we observe the opposite trend. On Politics, the least propaganda group has 9% more unique submitters, on Democrats 5%, on Republican 32%, on UKPolitics 9%, on Tories 10% while on LabourUK the most propaganda group has 28% more unique users that created a submission. This trend might indicate that certain users are more active in publishing propaganda content.

One follow-up question that we ask is how many of these users are bots. The presence of bots could explain why in one group there are fewer users posting articles. While there are several lists of Reddit bots, none of them is complete. Given this, we employ Rest-Sleep-and-Comment (RSC) (Ferraz Costa et al., 2015), a generative method that can distinguish human from bot posting activity. The method receives in input the intervals between two consecutive posts of a user, and these intervals are then compared with the aggregated distributions of intervals of all the users. The authors provide an initial training set of normal users and bots consisting of 37 bots and 999 users, to which we add 94 extra bots to make the model more robust. RSC has an average $F1$-score of 77.3 in cross-validation. The model requires at least 800 consecutive timestamps at which a user has written a comment. We retrieve from our subreddits all the users that posted a submission, and we keep the users for which we could retrieve the required number of timestamps. We note that the timestamps for a user are retrieved such that they represent consecutive chronological posts. Hence we do not restrict the subreddits in which the user might have posted. We find 748 possible bots on Politics, 91 on Democrats, 21 on Republican, 135 on UKPolitics, 23 on LabourUK and 9 on Tories. We investigate if these suspicious users posted a larger percentage of most propaganda articles in comparison with least propaganda articles. We find that this is the case on all subreddits, except Republican. However, the results are not statistically significant ($p > 0.05$) and the differences are close, as seen in Table 6. Hence, we can conclude that the bots' automatic activity in our dataset is not necessarily linked to posting propaganda content. Also, the small percentage of content in most propaganda group published by the bots shows that the majority of the propaganda content is published by real users.

Subreddit	% articles in LP	% articles in MP
Politics	2.86	3.17
Democrats	7.70	9.13
Republican	2.36	1.95
UKPolitics	3.46	4.19
LabourUK	0.78	1.23
Tories	2.81	3.16

Table 6: The percentage of submissions posted by suspicious users in the least propaganda group (LP) and the most propaganda group (MP)

RQ2. Does propaganda differ across the political spectrum? For this analysis, we will distinguish between US-based subreddits and UK sub-

reddits. We compare these subreddits using our propaganda score. We find that there is a statistically significant difference between the median propaganda score of articles on all subreddits in US ($p < 0.001$), with the most propagandistic content being shared on the subreddit Republican ($median = 0.307$), followed by Democrats ($median = 0.250$), and finally Politics ($median = 0.222$). In the UK subreddits, UKPolitics ($median = 0.214$) and Tories ($median = 0.217$) contains less propaganda than LabourUK ($median = 0.257$). There is no statistical difference between UKPolitics and Tories, tested using Kruskal–Wallis one-way analysis of variance, followed by Conover posthoc tests. These results indicate that *right leaning forums are not more likely to post propaganda than left leaning*. However, the tendance of using propaganda could result from the popularity of the respective party in the country. The Conservative Party in the UK has been in government since 2010, and a 2019 survey showed the party 15 points ahead of the Labour party. Even if the Republican party in the US won the White House in 2016, it didn't win the popular vote, and according to surveys more Americans identify as democrats. We also note that the subreddits that don't claim any political affiliation, Politics and UKPolitics, have less propagandistic content, which is consistent with the results in Table 5.

A second question is if the propaganda techniques employed differ according to the subreddits' political leaning or according to the country. We annotate using our propaganda technique identification classifier the sentences we previously labeled as propaganda to test this. We restrict ourselves to articles in the group most propaganda, using the intuition that if many sentences in the same article raise flags in the classifier, it is more likely that the article contains propaganda. For each subreddit, we rank the propaganda techniques by their frequency. We find that the relative ranking of techniques does not differ much between subreddits from the same country. The top 5 most frequent techniques are in the US *loaded language, name-calling, exaggeration or minimization, flag waving, doubt*, while in the UK based subreddits we have *loaded language, name calling, doubt, appeal to fear or prejudice, exaggeration or minimization*. Given the low accuracy of our technique classifier, we cannot make any definitive claims. However, such differences between the subreddits discussing politics in the two countries are plausible when considering the cultural differences. For example, Americans might be more susceptible to flag-waving, the technique of using patriotic feelings to justify an action. In 2017, 67% of Americans believed that the US is the leader of the free world according to a survey by the Public Broadcasting Service.

RQ3. How is propaganda received on political forums? To answer this last question, we aim to understand if more propaganda content will create more engagement. On Reddit, engagement is measured in the number of comments or the number of votes.

Firstly, we investigate if users comment more on submissions with higher propaganda score. We compare the median number of comments between the least propaganda group and the most propaganda group for each subreddit using the one sided Mann–Whitney U test. We find that on Politics, Democrats, Republican, UKPolitics and LabourUK, submissions in the most propaganda group receive more comments, while on Tories we observe the opposite effect.

We usually associate propaganda with media outlets. However, people can employ the same techniques to persuade the audience. We investigate how comments with propagandistic undertones are received on Reddit. For this we look at the comment's score, which is the difference between the upvotes and downvotes that a comment received. We construct two groups of comments: *positively received* comments that have the $score \geq t_{pos} \geq 10$, and *negatively received* comments with the score $score \leq t_{neg} \leq -5$. We compute the average propaganda score in the positively and negatively received groups while increasing the absolute value of the thresholds (t_{pos} from 10 to 50 and t_{neg} from -5 to -50), as shown in Figure 1. We observe that the average propaganda score of a comment increases with the engagement it generates, measured as the number of upvotes or downvotes it received. However, the trend is not observed on Republican and on Tories, one of the smaller subreddits for which we have very few data points in the plot.

7 Conclusion

In this work, we perform an extensive analysis of propaganda on online forums. We study for one year six subreddits from two English-speaking

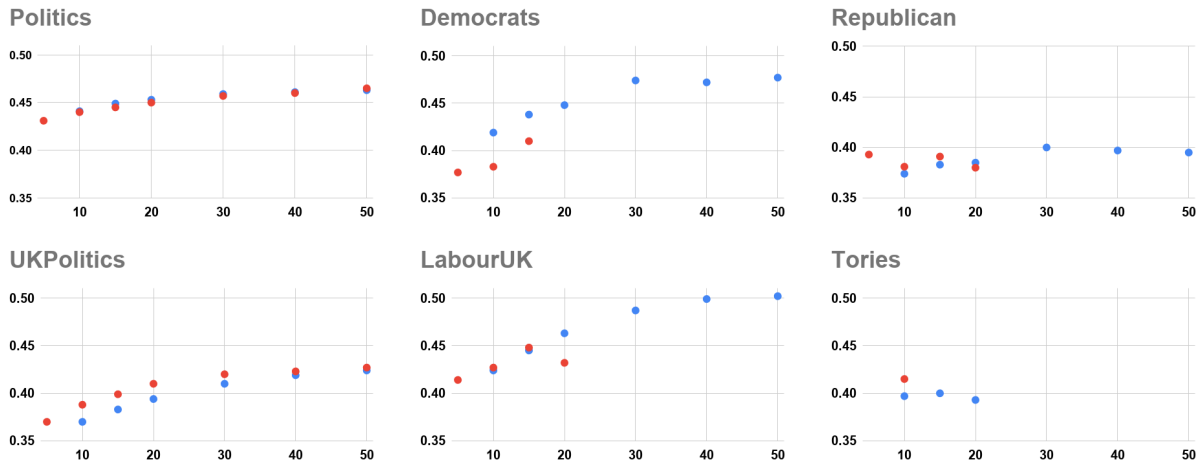

Figure 1: The average propaganda score in the positively received (blue) and negatively received (red) groups, while increasing the absolute value of the threshold. A data point represents a group with more than 100 comments.

countries, the US and the UK. We find several interesting patterns that can be leveraged by Reddit users and moderators to create better online discussions. We have found trends that we believe were not observed before in the literature. For example: *i*) the parties which represent a minority in a country might tend to use more propaganda; *ii*) political bias (either towards the right or the left) might be an indication of propaganda; *iii*) users that post more biased content form smaller communities; *iv*) differences in the use of propaganda techniques across countries might be rooted in cultural differences; *v*) submissions and comments having more propaganda content tend to receive more engagement in the form of number of comments, upvotes or downvotes. We note that while we have thoroughly tested all our hypotheses, our work is based on the automatic labelling of submissions and comments, with all the imprecision of such a method. We believe that understanding how propaganda affects us is of utmost importance for ensuring we live in democratic societies.

Acknowledgements

We thank Ioana Manolescu for our initial discussions on propaganda identification. We also thank the anonymous reviewers for their time and insightful comments.

References

Ahmed Al-Rawi and Jacob Groshek. 2020. Jihadist propaganda on social media: An examination of isis related content on twitter. In *Cyber Warfare and Terrorism: Concepts, Methodologies, Tools, and Applications*, pages 1442–1457. IGI Global.

Jisun An, Haewoon Kwak, Oliver Posegga, and Andreas Jungherr. 2019. Political discussions in homogeneous and cross-cutting communication spaces. In *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*, pages 68–79. AAAI Press.

Vít Baisa, Ondrej Herman, and Ales Horák. 2019. Benchmark dataset for propaganda detection in czech newspaper texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, September 2-4, 2019*, pages 77–83. INCOMA Ltd.

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

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset.

Halil Bisgin, Hasan Arslan, and Yusuf Korkmaz. 2019. Analyzing the dabiq magazine: The language and the propaganda structure of isis. In *Social, Cultural, and Behavioral Modeling*, pages 1–11, Cham. Springer International Publishing.

Gillian Bolsover and Philip Howard. 2017. Computational propaganda and political big data: Moving toward a more critical research agenda. *Big Data*, 5:273–276.

Mark J. Carman, Mark Koerber, Jiuyong Li, Kim-Kwang Raymond Choo, and Helen Ashman. 2018. Manipulating visibility of political and apolitical threads on reddit via score boosting. In *17th IEEE International Conference On Trust, Security And*

Privacy In Computing And Communications / 12th IEEE International Conference On Big Data Science And Engineering, TrustCom/BigDataSE 2018, New York, NY, USA, August 1-3, 2018, pages 184–190. IEEE.

Alceu Ferraz Costa, Yuto Yamaguchi, Agma Juci Machado Traina, Caetano Traina Jr., and Christos Faloutsos. 2015. RSC: mining and modeling temporal activity in social media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 269–278. ACM.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, Online. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pages 273–274. ACM.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alceu Ferraz Costa, Yuto Yamaguchi, Agma Juci Machado Traina, Caetano Traina, and Christos Faloutsos. 2015. Rsc: Mining and modeling temporal activity in social media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 269–278, New York, NY, USA. Association for Computing Machinery.

Anna Guimaraes, Oana Balalau, Erisa Terolli, and Gerhard Weikum. 2019. Analyzing the traits and anomalies of political discussions on reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):205–213.

Philip Howard and Bence Kollanyi. 2016. Bots, #strongerin, and #brexit: Computational propaganda during the uk-eu referendum. *SSRN Electronic Journal*.

Sofia Hurtado, Poushali Ray, and Radu Marculescu. 2019. Bot detection in reddit political discussion. In *Proceedings of the Fourth International Workshop on Social Sensing*, SocialSense'19, page 30–35, New York, NY, USA. Association for Computing Machinery.

Ansgar Kellner, Christian Wressnegger, and Konrad Rieck. 2020. What's all that noise: Analysis and detection of propaganda on twitter. In *Proceedings of the 13th European Workshop on Systems Security*, EuroSec '20, page 25–30, New York, NY, USA. Association for Computing Machinery.

Zoltán Kmetty, Veronika Vincze, Dorottya Demszky, Orsolya Ring, Balázs Nagy, and Martina Katalin Szabó. 2020. Pártélet: A Hungarian corpus of propaganda texts from the Hungarian socialist era. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2381–2388, Marseille, France. European Language Resources Association.

Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. 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)*, pages 4153–4163, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4826–4832. ijcai.org.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Robert Mueller. 2018. *Indictment of Internet Research Agency*.

Institute for Propaganda Analysis. 1938. *Propaganda Analysis: Volume I of the Publications of the Institute for Propaganda Analysis*. Institute for Propaganda Analysis, Inc., 130 Morningside Drive, New York, N.Y.

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 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2931–2937. Association for Computational Linguistics.

Marian-Andrei Rizoiu, Timothy Graham, Rui Zhang, Yifei Zhang, Robert Ackland, and Lexing Xie. 2018. #debatenight: The role and influence of socialbots on twitter during the 1st U.S. presidential debate. *CoRR*, abs/1802.09808.

Jon Roozenbeek and Adrià Salvador Palau. 2017. I read it on reddit: Exploring the role of online communities in the 2016 us elections news cycle. In *Social Informatics*, pages 192–220, Cham. Springer International Publishing.

Ahmed Soliman, Jan Hafer, and Florian Lemmerich. 2019. A characterization of political communities on reddit. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, page 259–263, New York, NY, USA. Association for Computing Machinery.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.