The Journal of Web Science, 2019, 7

Partisanship, Propaganda and Post-Truth Politics: Quantifying Impact in Online Debate

Genevieve Gorrell[1], Mehmet E. Bakir[1], Ian Roberts[1], Mark A. Greenwood[1], Benedetta Iavarone[2] and Kalina Bontcheva[1]

[1] *University of Sheffield, UK Scuola*
[2] *Normale Superiore, Italy*

ABSTRACT

The recent past has highlighted the influential role of social networks and online media in shaping public debate on current affairs and political issues. This paper is focused on studying the role of politically-motivated actors and their strategies for influencing and manipulating public opinion online: partisan media, state-backed propaganda, and post-truth politics. In particular, we present quantitative research on the presence and impact of these three "Ps" in online Twitter debates in two contexts: (i) the run up to the UK EU membership referendum ("Brexit"); and (ii) the information operations of Russia-backed online troll accounts. We first compare the impact of highly partisan versus mainstream media during the Brexit referendum, specifically comparing tweets by half a million "leave" and "remain" supporters. Next, online propaganda strategies are examined, specifically left- and right-wing troll accounts. Lastly, we study the impact of misleading claims made by the political leaders of the leave and remain campaigns. This is then compared to the impact of the Russia-backed partisan media and propaganda accounts during the referendum. In particular, just two of the many misleading claims made by politicians during the referendum were found to be cited in 4.6 times more tweets than the 7,103 tweets related to Russia Today and Sputnik and in 10.2 times more tweets than the 3,200 Brexit-related tweets by the Russian troll accounts.

Keywords: Politics, Social Media, Misinformation

ISSN 2332-4031; DOI 10.34962/jws-84

1 Introduction

"Post-truth politics" (Higgins, 2016) and "weaponized relativism"[1] describe strategies by which misleading information can be used to shape debates, redirect attention and sow confusion in order to influence political outcomes. In recent times, concern has been raised about politicians, foreign states, and hyperpartisan media exploiting social media to try to reach out and influence voters and citizens on an unprecedented scale. Where once social media were heralded as the beginning of a new age of interactive democracy, the question in the minds of researchers and many others is now "can democracy survive the internet" (Persily, 2017). A working theory might postulate that the low bar to publishing created by Web 2.0 has resulted in a number of effects that we explore here under three headings:

- **Partisan media**: today's highly competitive online media landscape has resulted in poorer quality journalism and worsening opinion diversity, with misinformation, bias and factual inaccuracies routinely creeping in. Many outlets also resort to highly partisan reporting of key political events, which can have acrimonious and divisive effects.
- Online **propaganda**: State-backed (e.g. Russia Today), ideology-driven (e.g. misogynistic or Islamophobic), or for-profit clickbait websites and social media accounts are engaged in spreading manipulative content and disinformation often with the intent to deepen social division and/or influence key political outcomes.
- **Post-truth politics**, where politicians, parties and governments frame key political issues in propaganda instead of facts. Misleading claims are repeated, even when proven untrue by journalists or independent fact checkers. This has a highly corrosive effect on public trust and informed participation in democratic processes.

While researchers have started studying these recently (Skjeseth, 2017; Ferrara, 2017), most work has focused primarily on misinformation and fake news during elections (Vosoughi *et al.*, 2018; Kaminska *et al.*, 2017) and the role of bots in spreading it (Shao *et al.*, 2018; Howard and Kollanyi, 2016). This paper presents large-scale, quantitative research on the presence and impact of these three "Ps" in online Twitter debates in two contexts: (i) the run up to the UK EU membership referendum ("Brexit"); and (ii) the information operations of Russia-backed online troll accounts. The aggregate data on which this work is based is available online. [2]

[1] https://www.theguardian.com/commentisfree/2015/mar/02/guardian-view-russian-propaganda-truth-out-there

[2] https://gate-socmedia.group.shef.ac.uk/political-polarisation-disinformation-and-bots/ppp-supp-mats/

We first compare the impact of highly partisan versus mainstream media during the Brexit referendum, specifically comparing tweets by half a million "leave" and "remain" supporters.

Next, online propaganda strategies are examined. Late in 2018 Twitter released a set of nine million tweets from accounts they have identified as belonging to the Russian Internet Research Agency (IRA). The IRA dataset covers a time period spanning from the beginning of the Ukraine conflict in 2014 through the Brexit referendum and US presidential election until well into President Donald Trump's term of office. These data provide rich possibilities for investigating propaganda. We present here the first exhaustive analysis of this new dataset, with a focus on what we can learn about how propaganda succeeds and fails under the conditions created by modern social media. We also accurately classify accounts into different activity types (left trolls, right trolls, etc.), enabling a deeper understanding of how different strategies pay off in terms of impact.

Lastly, we study the impact of misleading claims made by the political leaders of the leave and remain campaigns. This is then compared to the impact of the Russia-backed partisan media and propaganda accounts during the referendum. In particular, just two of the many misleading claims made by politicians during the referendum were found to be cited in 4.6 times more tweets than the 7,103 tweets related to Russia Today and Sputnik and in 10.2 times more tweets than the 3,200 Brexit-related tweets by the Russian troll accounts.

1.1 Related Work

The work presented here is set against a backdrop of increasing awareness of the ways in which the internet and social media are changing society. Social media have been widely observed to provide a platform for fringe views. Faris *et al.* (2017) showed that social media seem to amplify more extreme views, with materials linked on Twitter being more outré than the open web, and on Facebook even more so, a finding echoed by Silverman (2015). Barberá and Rivero (2015) and Preoţiuc-Pietro *et al.* (2017) both show that Twitter users with more ideologically extreme positions post more content than those with moderate views.

Researchers also report consistent asymmetries in the way these changed conditions play out. Allcott and Gentzkow (2017), during the run-up to the 2016 US presidential election, found 115 pro-Trump fake news stories, which were shared a total of 30 million times. They found 41 pro-Clinton fake news stories, which were shared a total of 7.6 million times. This disparity is again echoed in Silverman's work (Silverman, 2015). Hare and Poole (2014) find that the increased separation between American left and right wing partisans in recent years is accounted for by a right wing shift to the right; left wing voters have not changed their position.

There is little evidence of a difference in the way information consumers of different political valences respond to materials that might account for this asymmetry (Faris *et al.*, 2017; Allcott and Gentzkow, 2017). Instead, Faris *et al* suggest that in the case of the 2016 presidential election, it was the cooperative behaviour of pro-Trump media themselves that led to an advantage, in a phenomenon they dub "network propaganda". This raises questions about the reach of such a network or the conditions under which it might arise elsewhere, and its relationship to political views if any. The idea of an "alternate reality" created by network propaganda has implications for social polarization given observations by Lewandowsky *et al.* (2017) that where partisans are isolated in echo chambers extremism is rewarded, as a message may reach sympathizers without the cost attached in alienating centric or opposing voters.

A body of work (Lansdall-Welfare *et al.*, 2016; Mangold, 2016) has begun to explore Brexit opinion and sentiment as expressed on Twitter. Matsuo and Benoit [3] focus on differences in the dialogue between leave and remain camps. Mostly manual research by Moore and Ramsay (2017) is focused on analysing the newspaper media during the referendum and highlights differences in the tone of the different campaigns. Our work builds on theirs by exploring how the behaviour they discuss relates to a medium's partisan appeal, as well as focusing on social media, rather than newspapers.

Howard and Kollanyi (2016) share our interest in propaganda. Our novel contributon is in exploiting large-scale, reliable voter classification in order to explore partisan dynamics and polarisation. Their group have also specifically investigated Russian bot involvement in Brexit (Narayanan *et al.*, 2017), but on a significantly smaller scale. Likewise, Bastos and Mercea (2017) study the impact of bot activity during Brexit, and present some observations about the nature of the content they spread. They find that such materials are likely to be user-generated, tabloid-style emotionally orientated materials. The role of Twitter misinformation and bot activity in the context of the 2016 US presidential election has attracted much research attention, as previously discussed. This has primarily focused on the amount of traffic generated by bots or trolls, without providing evidence of impact. In this paper, instead, we focus on quantifying bot impact and exploring the strategies for achieving it.

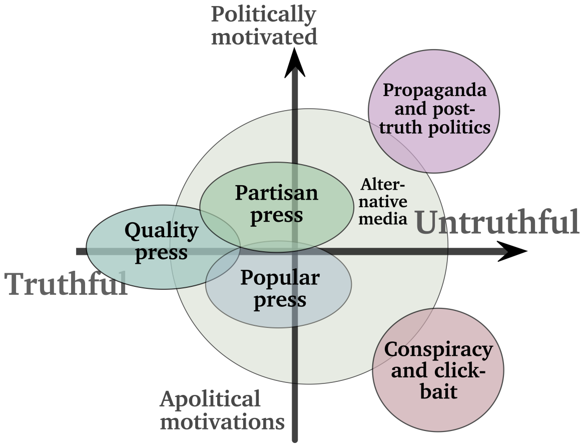
The release of the IRA dataset is so recent as to preclude much in the way of in-depth investigation so far, though descriptive work is available from SMaPP (Yin *et al.*, 2018). The largest prior study available by Linvill and Warren (2018) still had access to only 3 million tweets, which is very significantly less than the 9 million just released by Twitter. This new large corpus constitutes an unprecedented opportunity, since troll accounts are rapidly suspended by the platform, creating a moving target for research.

1.2 Term Definitions

The politically-motivated actors and strategies that are central to this study (partisan media, propaganda, and post-truth politics) have complex, overlapping characteristics. Figure 1 provides a conceptual diagram of these interrelationships, as examined in the scope of this paper. We distinguish explicitly *political vs. apolitical*, because although there are many other cases where propaganda and partisan media play a significant role, the focus here is on political influences. The sector of the figure that we are interested in in this work is the *top right*;

[3] http://blogs.lse.ac.uk/brexit/2017/03/16/more-positive-assertive-and-forward-looking-how-leave-won-twitter/

Figure 1: Term Definition and Conceptualisation



namely, political and less truthful/unbiased, as we aim to highlight these important new trends in techno-political sociology. Others[4] have explored the "Ps" concept with more coverage of apolitical motivations.

Inevitably there are overlap and grey areas between the media and behaviours we discuss in this work. Motivations for behaviours are unclear; for example, is a popular political message in the press intended to influence political outcomes or sell more newspapers? In this work we confine our interest to media behaviour that is politically engaged *and* misleading. We therefore define:

- **Partisan media** to to be media presenting themselves as news, including:
  - **Partisan press**; mainstream media unambiguously identifying as providers of news reportage, but who may present partisan materials as more factual than they really are;
  - **Alternative media**; a broad and varied ecosystem of new publishers presenting themselves as news, some of whom are politically partisan and therefore of interest;
- **Propaganda** to be politically motivated behaviours and materials with a primary purpose of influencing toward a particular point of view, see e.g. OED.[5] Origin may be veiled;
- **Post-truth politics** to be politically motivated output with little regard for truth and public, political figure or entity as instigator;

We explore our findings below under these headings.

2 Methodology

The first corpus used is a large collection of tweets collected using the GATE Cloud Twitter Collector, [6] a tool that allows tweets to be gathered according to search criteria as they appear, and processed using GATE [7] text processing pipelines to enrich the tweets with relevant background information, including the EU membership stance of the author. The method is described more fully by Maynard *et al.* (2017). In the next section we describe collecting the tweets, then after that the user vote intent classification. The corpus thus enriched was indexed using the Mímir search engine for efficient exploration, which again is described in more detail by Maynard *et al.* (2017).

The second corpus is Twitter's IRA data downloaded from their site.[8] We introduce this corpus at the end of this section.

Partisanship Attention Score

Throughout the work we make use of Partisanship Attention Score (PAS), first introduced by Faris *et al.* (2017). This metric is a simple ratio of the number of times a source is linked by one valence of user, for example "leavers" (Brexit), versus the other valence. In this work we use "leave-PAS" to describe a PAS in which leave linkers outnumber remain linkers, and "remain-PAS" to describe a PAS in which remain linkers dominate. We have grouped sources into five sets; those in which a PAS is greater than 30:1 (one leave set and one remain set), those in which the PAS is greater than 3:1 (leave and remain) and those with a more balanced PAS of less than 3:1. The 30:1 and 3:1 ratios were selected heuristically–throughout the work we are careful to reflect on how that choice might affect the results.

2.1 Brexit Tweet Collection

Around 17.5 million tweets were collected up to and including 23 June 2016 (EU referendum day). The highest volume was 2 million tweets on Jun 23rd (only 3,300 lost due to rate limiting), with just over 1.5 million during poll opening times. Of the 2 million, 57% were retweets and 5% replies. June 22nd was second highest, with 1.3 million tweets. The 17.5 million tweets were authored by just over 2 million distinct Twitter users (2,016,896). The work presented here focuses on a subset of these, covering the month up to and including June 23rd. Within that period, there were just over 13.2 million tweets, from which 4.5 million were original tweets (4,594,948), 7.7 million were retweets (7,767,726) and 850 thousand were replies (858,492). These were sent by just over 1.8 million distinct users. The tweets were collected based on the following keywords and hashtags: *votein, yestoeu, leaveeu, beleave, EU referendum, voteremain, bremain, no2eu, betteroffout, strongerin, euref, betteroffin, eureferendum, yes2eu, voteleave, voteout, notoeu, eureform, ukineu, britainout, brexit, leadnotleave.* These

[4] https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79

[5] http://www.oed.com/view/Entry/152605

[6] https://cloud.gate.ac.uk/shopfront/displayItem/twitter-collector

[7] https://gate.ac.uk/

[8] https://about.twitter.com/en_us/values/elections-integrity.html#data

were chosen for being the main hashtags, and are broadly balanced across remain and leave hashtags, though the ultimate test of the balance of the dataset lies in the number of leavers and remainers found in it, which is discussed below.

Most URLs found in tweets have been shortened, either automatically by Twitter or manually by the user, which has the side-effect of obfuscating the original domain being linked to. For this work we expanded the URLs in tweets using the following approach. From manual analysis of the URLs we accumulated a list of 18 URL shorteners or redirect services: shr.gs, bit.ly, j.mp, ow.ly, trib.al, tinyurl.com, ift.tt, ln.is, dlvr.it, t.co, feeds.feedburner.com, redirect.viglink.com, feedproxy.google.com, news.google.com, www.bing.com, linkis.com, goo.gl, and adf.ly. All URLs from other domains were considered to already be expanded. (A small number of minor URL shorteners have gone unexpanded due to the long tail in this large tweet set and the necessity of manually identifying shortening services.) When we saw a shortened URL it was expanded, either by following HTTP redirects or using the API of the shortener, recursively until the resulting URL no longer pointed to a domain in our list of shorteners.

2.2 User Vote Intent Classification

Classification of users according to vote intent was done on the basis of tweets authored by them and identified as being in favour of leaving or remaining in the EU. Such tweets were identified using 59 hashtags indicating allegiance, given in the online experimental materials.[9] Hashtags in the final position more reliably summarise the tweeter's position, so only these were used. Consider, for example. "is Britain really #strongerin? I don't think so! #voteleave".

This approach was evaluated using a set of users that explicitly declared their vote intent. A company called Brndstr[10] ran a campaign offering a topical profile image modification (a flag overlaid on their profile picture) in response to a formulaic vote intent declaration mentioning their brand. This enabled a ground truth sample to be easily and accurately gathered. On these data, we found our method produced a 94% accuracy even on the basis of a single partisan tweet (where three are required, an accuracy of 99% can be obtained, though only 60,000 such users can be found, as opposed to 290,000 with at least one partisan tweet). The Brndstr data itself, consisting of around 100,000 users of each valence, was also used to supplement the set, raising the accuracy further, and resulting in a list of 208,113 leave voters and 270,246 remain voters. Table 1 gives detailed statistics for three conditions; one matching tweet found for that user, two found or three found. "Total" is the total number of users found with that number of matching tweets. "Brndstr found" is the number of those users found in the Brndstr set, and so able to be evaluated. The remaining figures refer to that set, providing an accuracy for the total list of users found using the given minimum number of partisan tweets.

	Total	Brndstr found	Of found correct	Accuracy	Cohen's kappa
Leavers, 3#	34539	1142	1129	0.987	0.972
Remainers, 3#	26674	603	594		
Leavers, 2#	49080	1368	1350	0.984	0.966
Remainers, 2#	50972	901	882		
Leavers, 1#	114519	1935	1801	0.943	0.885
Remainers, 1#	175042	1744	1667		

Table 1: Brexit Classifier Accuracy

There may be a case for using a threshold of two hashtags in order to produce a more balanced set of leavers and remainers, but this would disproportionately exclude remainers with more moderate feelings (if the number of hashtags can be seen as an indicator of this). The resulting set is somewhat slanted toward remainers, demonstrating the obvious; that Twitter isn't a representative sample of the UK population, who voted to leave the EU to the order of 52%. However, leavers were more vocal and apparent in the data presented below, contrary to what we would expect if the higher number of remainers had affected the result. It is possible that some users changed their mind about how to vote after making their Brndstr declaration, but voters making an online declaration of their vote intent are perhaps those less likely to vacillate, and the work can in either case be seen as an exploration of the behaviour of those who held a particular allegiance during the time period studied.

2.3 IRA Corpus and Account Classification

The Twitter IRA corpus[11] contains 3,836 unique users and 9,041,308 tweets. The tweets are posted in 57 different languages, but most of the tweets are in Russian (53.68%) and English (36.08%), comprising almost 90% of the tweets. The majority of accounts (as opposed to tweets) are self-declared English language (2,384), but note that many of these have Russian display names. Average account age is around four years, and the longest accounts are as much as ten years old. Linvill and Warren (2018) have analyzed the English language accounts and find several key types of account emerging. A large amount of activity in both the English and Russian accounts is given to **news** provision. Secondly, many accounts seem to engage in **hashtag games**, which may be an easy way to establish a history for an account to make it seem more credible. Of particular interest however are the political trolls. **Left trolls** pose as individuals interested in the Black Lives Matter campaign. **Right trolls** are patriotic, anti-immigration Trump supporters. Among left and right trolls, several have achieved large follower numbers and even a degree of fame.[12] Finally there are **fearmonger** trolls, that propagate scares, and a small number of **commercial** trolls. The Russian language accounts may also provide news, or may pose as indi-

[9] https://gate-socmedia.group.shef.ac.uk/political-polarisation-disinformation-and-bots/ppp-supp-mats/

[10] http://www.brndstr.com/

[11] https://about.twitter.com/en_us/values/elections-integrity.html#data

[12] https://www.theguardian.com/technology/shortcuts/2017/nov/03/jenna-abrams-the-trump-loving-twitter-star-who-never-really-existed

Actual\Predicted	Hash.	Left	Right	Fear	News	Comm.
Hashtag Gamer	23	0	4	0	0	0
Left Troll	0	42	14	0	0	0
Right Troll	1	8	141	0	1	0
Fearmonger	0	0	2	26	0	0
Newsfeed	0	0	1	0	12	0
Commercial	0	0	0	0	0	1

Table 2: Troll Classification Confusion Matrix

viduals with opinions about for example Ukraine or western politics. These troll types provide insight into how IRA effort was targeted and to what extent these different behaviour types translate into impact, such as followers attracted to the accounts and retweets achieved. For this reason we took their dataset and built a classifier enabling us to classify all the accounts.

Linvill and Warren manually categorized 1,102 IRA-associated handles into the six categories described above, providing us with an adequate training set to build a classifier. 55% of their labelled accounts are right trolls, 20% are left trolls, 10% are fearmonger and hashtag gamer accounts, 5% are newsfeeds and less than 1% are commercial accounts. We used a support vector machine (SVM) to predict the categories of the remaining accounts. Features were term frequency-inverse document frequency (tf-idf) of English tweet texts, the domain of shared links including the domains of the shortened and expanded versions of the links, and the topic distribution of the tweet text.

We used 75% of the dataset for training and 25% for testing. The accuracy was 0.89. Table 2 gives the confusion matrix of the test data. The only significant area of confusion is between left and right trolls, which may be partially explained by accounts being repurposed; in this work we did not investigate account repurposing. Alternatively it may be that these account types are confusable for other reasons. The final model was trained on all data and was used to classify the remaining 2157 accounts which had English tweets. No attempt was made to classify an account that had no English language tweets. The resulting fully classified dataset contains 60% right trolls, 12% fearmongers, 11% having no English language tweets, 10% left trolls, 5% hashtag gamers, 3% newsfeed accounts and negligible commercial accounts (n=6). The reason for the change in class proportions is likely to be the criteria that Linvill and Warren used for selecting accounts to manually classify. They classified accounts represented in their tweet set, which was collected via retrospective search on IRA account names in late 2017, and collected therefore only tweets still available at that point going back to mid-2015. We find generally speaking more left and right trolls than in their sample, and fewer newsfeeds and hashtag gamers.

3 Findings

We now present findings under the headings of the three "Ps", beginning with partisan media, then moving on to propaganda, then post-truth politics.

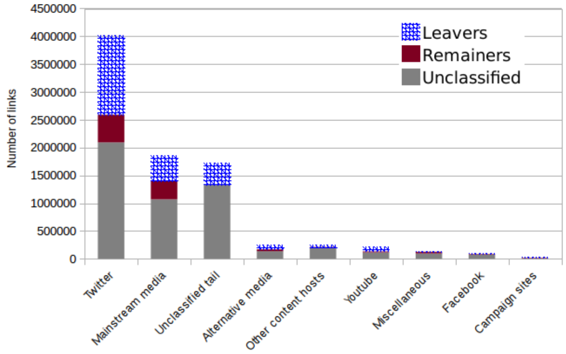


Figure 2: Types of links posted

3.1 Partisan Media

We begin our investigation with the Brexit tweet collection described above. As a starting point for quantifying the various influences and evidence of partisanship, the top 100 most posted domains were manually grouped into high level categories, as shown in figure 2. The dominant domain to appear was Twitter itself, appearing whenever anyone posts an image, as well as when they link to another tweet. After that, the greater proportion of the links are to items in a wide variety of mainstream news media. "Other content hosts" refers to smaller content platforms such as Instagram. YouTube and Facebook are listed separately. Finally, smaller amounts of material are linked from referendum campaign sites and alternative media. (Alternative media range from publications that are nearly mainstream through to conspiracy sites and fake news.) The "long tail" of a further 17,000 less linked domains that haven't been manually classified are included in the chart to give a quantification of the unknown; note that this unknown section is likely to contain many more small alternative media, blogs etc. than mainstream media. Also only domains that were tweeted at least once by a user that has been classified for vote intent were included. The actual number of domains mentioned in the set is much greater. The graph broadly agrees with table 1 of Narayanan *et al.* (2017). We are also able divide each count into three parts, indicating the proportion of tweets in that section by unclassified users, remainers and leavers. It is evident at a glance that remainers were tweeting less linked material, since their representation is smaller. Also there were fewer remainers in the unclassified tail (that is, the column of unclassified sites, not the unclassified users), suggesting perhaps a preference for more popular sites on the part of remainers. It is unknown how many leavers, remainers and undecideds constitute the unclassified users (the grey bottom section of the columns), though the domains tweeted by them suggest greater neutrality than the classified users (Guardian, BBC, Telegraph).

PAS of High Impact Media

Figure 3 shows the sites that had the most impact, in terms of total number of times they appeared in tweets in the Brexit dataset. These were almost entirely mainstream media, mostly UK media, with the exception of the remain campaign site "uk-

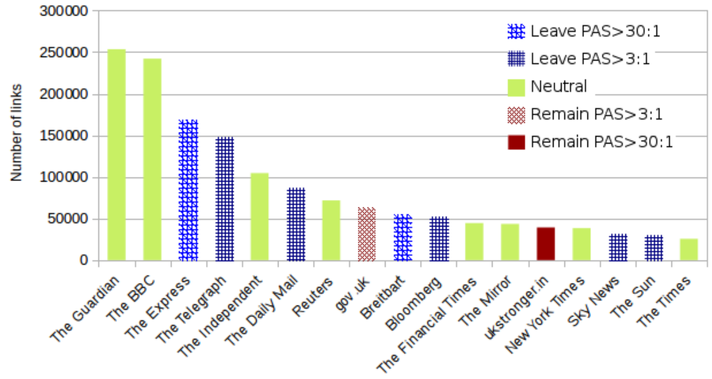


Figure 3: Number of appearances of high impact sites

stronger.in" and the UK government domain. The graph gives total counts of appearances of the most influential domains, colour coded by partisanship attention score (PAS); the ratio of links from leave voters to remain voters or vice versa. Platforms such as Facebook, where the site doesn't author the content, are excluded. Only link appearances in original tweets are used in this graph (not appearances in retweets or replies). Tables 1 and 2 in the supplementary materials give a longer list of sites (the full set is also available for download). [13]

On page 13 of Moore and Ramsay (2017) a similar graph shows the number of referendum-related articles published by UK media. The number of Brexit articles published by a medium shows a strong correlation to its link presence on Twitter (0.71). In fact, the Express has been somewhat less taken up on Twitter than its engagement with the subject might predict; figure 7 and its discussion later in the paper may offer further insights on this point.

It is evident that mainstream media were the dominant source of linked materials in the Brexit discussion on Twitter, with the six most influential domains all being British mainstream media as shown in figure 3. Smaller in influence but nonetheless significant were alternative media, with Breitbart appearing in ninth place in figure 3, user-shared content on other content platforms such as Facebook, and campaign sites. This suggests a continuing important role for traditional media, though leaves questions about how social media, and indeed alternative media, may interact to popularize certain materials and influence the focus. It is also apparent that the most popular domains were either neutral in their appeal or appealed to leavers, with only two smaller sources, the government and the "Stronger In" campaign, appealing to remainers. This subject is taken up more fully in the next section.

3.1.1 Ground-Truthing Mainstream Media

Figure 4a shows British mainstream newspapers ranked from left to right in order of their PAS ratio. For those media with negative leave PAS ratios, the remain PAS ratio has been plotted (ratio of appearances in remain tweets against those in leave

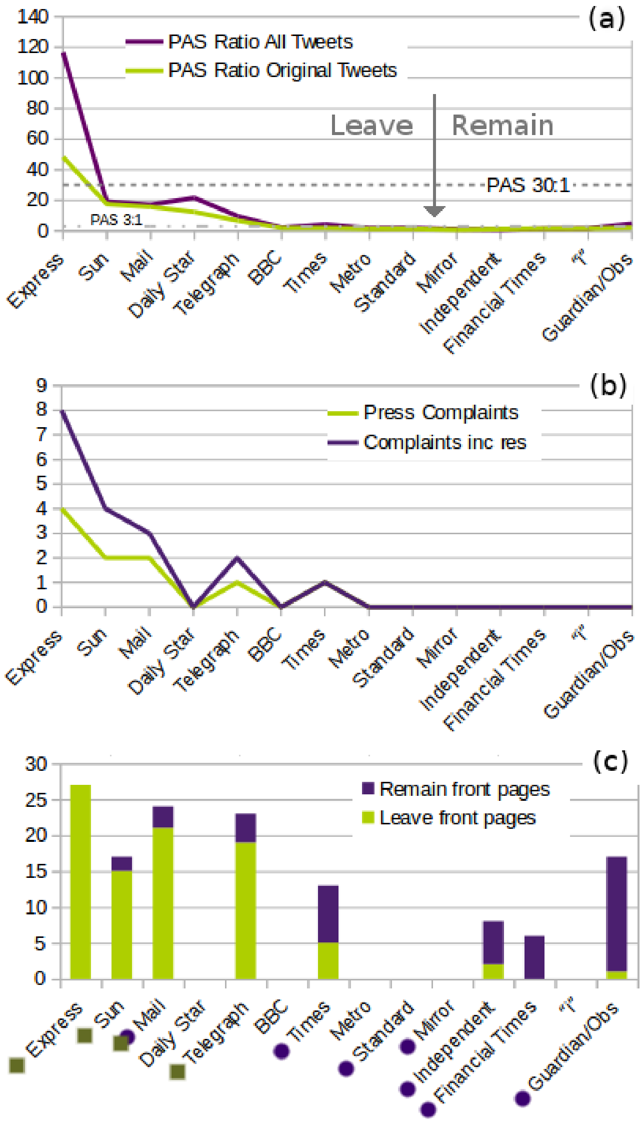


Figure 4: PAS (a), Press Complaints (b) and Partisan Front Page Counts (c) for UK Mainstream News Media

[13] https://gate-socmedia.group.shef.ac.uk/political-polarisation-disinformation-and-bots/ppp-supp-mats/

tweets). In this way, both leave and remain media can be shown commensurately on the same graph. The point at which the PAS ratios switch direction is indicated with a vertical arrow. The extreme right of the graph, therefore, shows the newspaper with the highest remain PAS ratio (The Guardian/Observer). Two horizontal lines indicate PAS ratios of 3:1 and 30:1. PAS ratios for link appearances in all tweets and just original tweets are shown.

In figure 4b, the green line indicates the number of upheld press complaints for that medium. The purple line also includes the number of complaints for which a resolution was found. The majority of press complaints regarded articles that were anti-immigration in their focus.

In figure 4c, newspaper front pages provided by Moore and Ramsay (2017) for the two month period preceding the referendum have been manually classified as leave, remain or neutral in their orientation. An example of a leave front page might be "EU 'very bad' for pensions" (The Express, June 21st 2016). An example of a remain front page might be "Vote remain today" (The Mirror, June 23rd 2016). Bars show leave front pages in green and remain in purple. Where possible, the original article was consulted before classifying a front page. However, in many cases this information wasn't accessible. In these cases, a conservative judgment was reached, but this means that counts for the Sun and the Independent may be a little depressed, since the full article usually wasn't available for them. Note also that the work was completed by a single annotator, and that in many cases, classifying the headlines was quite a subtle judgment call.

Several British newspapers declared their allegiances regarding Brexit, reportedly giving media supporting the UK leaving the EU an audience of around 4.8 million, while those in favour of remaining in the EU reach just over 3 million [14]. Stance information is included in figure 4c in the form of coloured marks–a green square for leave and a purple circle for remain. Both marks appear for the Mail because the Daily Mail shares its domain with the Mail on Sunday. The Daily Mail were in favour of leaving the EU, and the Mail on Sunday, with a slightly lower circulation, were in favour of remaining.

PAS was found to correlate with press complaints (0.922, $p<0.001$) as well as bias as quantified by the magnitude of the difference between pro- and anti-Europe front page counts (0.842, $p<0.001$).

Figure 4a shows that all of the media that declared their support for the remain cause were broadly neutral in their appeal, with the exception of the Guardian/Observer, who, when retweets and replies are counted, has a leave PAS greater than 3:1. The media that declared their official support for leave all to varying extent appealed more to leavers. This brings to mind the conclusion of Faris *et al.* (2017) from their study of the 2016 US presidential election that mainstream media ranging from left to centre right show more investment in principles of neutrality. The Brexit question cut across the political spectrum, although in terms of media stance, the left-leaning papers favoured remain and the right, leave. However, it is also possible that leavers engaged with remain materials for other reasons. Press complaints and front page partisanship data provide further insights. It is interesting to note that PAS seems to echo upheld press complaints better than it does partisanship as indicated by front pages. There are prominent cases where media published many stories in keeping with their Brexit stance, but without attracting press complaints; most notably the Telegraph and the Guardian. Materials supportive of a particular stance don't *per se* seem to draw partisan attention—the PAS of both these media is low.

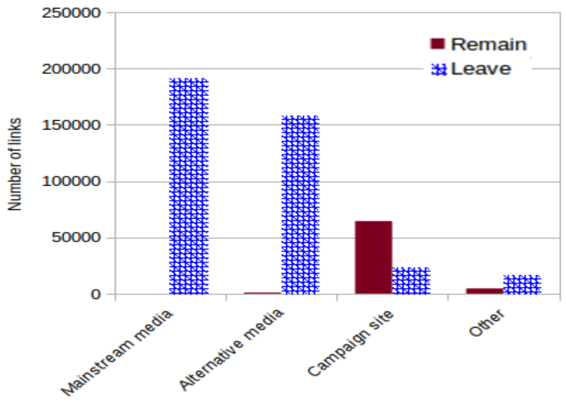


Figure 5: Who are the PAS>30:1 influencers?

This is important in correctly interpreting figure 3. The medium with the biggest impact is the Guardian, which published many pro-remain articles. So in this sense, there wasn't a lack of attention to pro-remain materials, and if the colour coding of the graph were based on the “front page diff” used above, the impression created would be quite different. PAS captures something different. Manual review of the tweets suggests that Guardian articles tend to be factual in tone, and attract critical engagement from leavers. Express articles tend to use emotive and suggestive language, and seem to attract less discussion. Moore and Ramsay's analysis (Moore and Ramsay, 2017) gives much information about the rhetorical styles employed by the press in the run-up to the referendum. Circulation size does not explain the number of complaints received, with the Express having less than half the readership of any of the four largest media.[15]

Extreme/Affective Materials

We saw in section 3.1 that high PAS scores show a potential relationship with upheld press complaints, and that polarity of PAS is a good indicator of the stance of the source, as determined from press front pages. We now use PAS scores of greater than 30:1 to select sources that may be misleading for further examination. Sites of either camp with at least 1000 total mentions in tweets in the dataset and at least 50 tweets, retweets or replies by leavers or remainers were manually analysed. We present the sites divided into 4 categories; mainstream media, alternative media, campaign sites and other

[14] https://www.huffingtonpost.co.uk/entry/which-newspapers-support-brexit_uk_5768fad2e4b0a4f99adc6525

[15] http://www.pressgazette.co.uk/nrs-national-press-readership-data-telegraph-overtakes-guardian-as-most-read-quality-title-in-printonline/

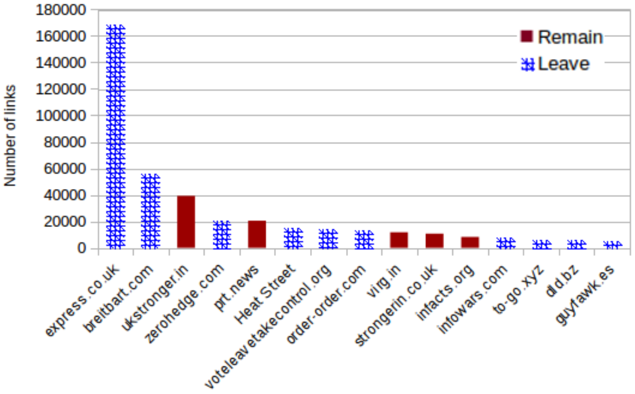


Figure 6: Who are the PAS>30:1 sites?

sites. "Others" includes for example personal blogs or special interest websites not primarily focused on Brexit.

Figure 5 shows that remain PAS>30:1 sites are dominated by explicit campaign sites. As we would expect given the data above, among leave influencers we see more mainstream media—note that the only high PAS mainstream media were leave media; namely the Express. We also see a much greater role for alternative media in the leave campaign. The total impact of leave PAS>30:1 media was 389,000 mentions. For remain it was 70146 mentions, or 18% of the PAS>30:1 impact. All sites with a PAS higher than 30:1 and more than 5000 mentions are shown in figure 6. The Express dominates, with the US alternative medium Breitbart in second place. As indicated above, remain sites are mainly campaign sites. Other leave sites are media ranging from alternative to conspiracy, plus the campaign site "voteleavetakecontrol.org". A longer list can be found in table 2 in the supplementary materials.[16]

Key observations from figure 5 include that in terms of mentions in tweets, the influence of leave sites dwarfs that of remain sites. It is also notable in that figure that high remain-PAS sites were mostly explicit campaign sites; in other words, openly partisan, with no suggestion of providing reportage. The range of media providing high leave-PAS materials, plus the presence of Breitbart raises the question of whether these findings demonstrate a similar phenomenon happening in the UK as described by Faris *et al*, or whether indeed it is simply the same phenomenon - an extension of the same network of propaganda.

Figure 7 presents counts of sites according to their PAS status. A threshold of 20 total original tweets by leavers and remainers was applied, in order to exclude sites for which too little evidence was available to classify them. The graph shows peaks to either extreme, despite the stringent 30:1 criterion, reinforcing previous researchers' findings that extreme content tends to proliferate on social media (Faris *et al.*, 2017; Silverman, 2015; Barberá and Rivero, 2015; Preoţiuc-Pietro *et al.*, 2017). The neutral peak most likely arises because content-neutral platforms such as Facebook are counted here, rather than because there is a peak in neutral materials such as unbiased news providers. On the right we see the actual link counts to the sites. Links to Twitter have not been included, since they give a large, uninformative boost to the neutral count. Were other content-neutral platforms to be excluded, this count would be lower still. Nonetheless, we see that the extremes no longer outnumber the moderate sites. Evidently most Twitter users prefer less extreme materials of those on offer. However, this provides evidence of the diet Twitter is offering.

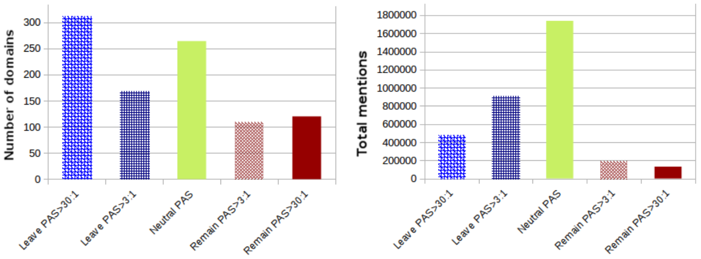


Figure 7: All domains vs total mentions by PAS of domain

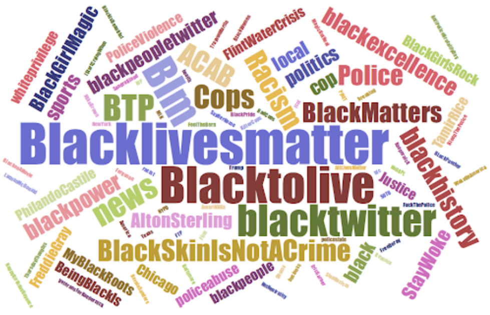


Figure 8: Left Troll Hashtags

3.2 *Online Propaganda*

Recall that political propaganda is non-objective information, which is aimed at influencing citizens and/or furthering a political agenda. In this section we use the Twitter IRA tweet collection, introduced in Section 2.3, to explore evidence for the impact of different propaganda strategies.

Initially, in the autumn of 2017, Twitter released a list of around 3,000 Twitter accounts to US Congress that they had identified as being Russian state-controlled troll accounts, and

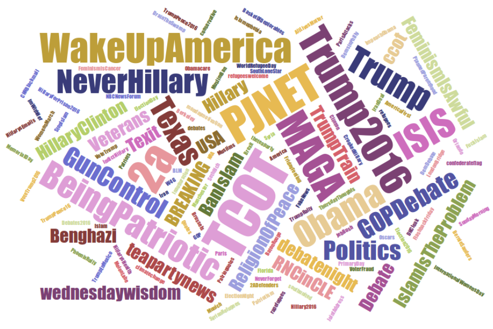


Figure 9: Right Troll Hashtags

[16] https://gate-socmedia.group.shef.ac.uk/political-polarisation-disinformation-and-bots/ppp-supp-mats/

Type	Num	Av Tw	Av Orig	Retw Rec	Av Foll	Retw Rat
Right	2194	2560	1436	8600	1609	5.989
Left	339	2755	1025	30047	1815	29.305
Fearmonger	432	487	481	10	62	0.022
Hashtag	189	3041	1582	922	2225	0.583
News	99	9981	9859	13921	9552	1.412
All trolls	3667	2466	1537	8522	1741	5.546

Table 3: Troll Impact

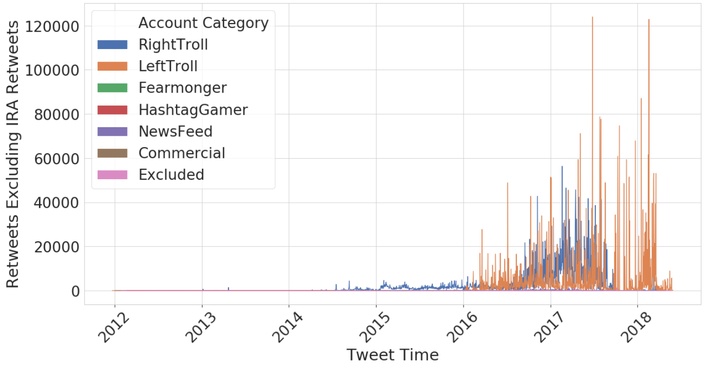


Figure 10: Timeline of Retweets Achieved by Troll Type

had suspended. In the autumn of 2018, the full set of 9 million tweets by these IRA propaganda accounts were released. The majority of tweets are in Russian as noted above, primarily with Ukraine-related focus. In contrast, the English language tweets focus predominantly on US politics.

Prior to the release of the full 9 million tweet set, Linvill and Warren (2018) researched a partial set of 3 million tweets by most of the IRA accounts, which they gathered and released independently. They found differing patterns of troll activity, with news accounts keeping up a relatively steady output of genuine news and achieving a fair reach, hashtag trolls showing bursty activity around playing "hashtag games"[17] (i.e., seeking to get many retweets and favourites through exploiting hashtags), and left and right trolls being more event-triggered. Political trolls in some cases achieve a significant following. Examples are given in table 4, and include both left and right trolls and news feeds.

Figures 8 and 9 give word clouds we generated for the subset of left and right troll accounts that were manually identified by Linvill and Warren (2018). Left troll material has a strong Black issues focus, and often talks about conflict with the police. Right troll material is political, supportive of Trump, against the Democrats and anti-Muslim.[18] We also find differences in the web domains left and right trolls tend to link. The most-linked domains of we found for Linvill and Warren's left and right trolls are included in table 3 in the supplementary materials.[19] Domains intersect with domains linked by leavers and remainers, as described above and also included in the supplementary materials. Three sites frequently linked by left trolls appear on the Brexit list; the Independent, the Huffington Post and the New York Times. All had a neutral PAS. Three highly partisan sites frequently linked by right trolls also appear on the Brexit list; Breitbart, Infowars and the Express. All had a leave PAS of greater than 30:1. This suggests an overlap in outlook between Brexit leave voters and the right troll persona. Left trolls link neutral sites as well as Black-focused sites that aren't relevant to Brexit.

Table 3 gives impact statistics for the different troll types, according to our classifier. First we give number of accounts, then average number of original tweets (excluding retweets). Then we report average number of retweets received, average number of followers and rate of retweets per original tweet. It is clear that political trolls achieve by far the best ratio of retweets to original tweets. Left trolls achieve more retweets per original tweet than right trolls. However, other account types are more highly followed, and news and hashtag accounts may influence their followers even though their tweets do not inspire retweets to the same extent. Where an agent retweets someone else's tweet rather than authoring an original tweet, we don't have data about how widely retweeted that tweet was, as it counts for the original author; it is possible that agents retweeting the tweets of others are having significant impact in amplifying a message. Of the account types shown, all have average longevities of active life approaching a couple of years with the exception of fearmonger trolls, where the average duration of active life (first activity to last activity) is less than six months. Follower count correlates with retweet rate per original tweet to the tune of 0.35, which is highly significant (p<0.001), but as we see, different types of tweeting behaviour produce different profiles in terms of being followed and being retweeted.

In figure 10 we see a timeline of retweets achieved for the different types of trolling behaviour. This gives an indicator of the effectiveness of the different troll types. It is notable that political trolls are achieving many more retweets than any other type, with the others barely appearing in the graph. Retweets by other IRA trolls have been removed from these counts. As a whole, IRA trolls have not tended to retweet each other a great deal; 27% of retweets in the corpus are of other trolls, but this was extremely variable; right trolls retweeted each other significantly until the end of 2016 then stopped. Hashtag gamers do retweet each other to a minor extent.

Figure 11 gives a network diagram of only trolls with more than 5,000 followers. Connections are based on the trolls mentioning, retweeting, replying to or quoting each other, not whether they follow each other, as we do not have access to that information in the dataset released by Twitter. "Not English" accounts are mostly Russian, and consist of a large number of newsfeed accounts ("novosti") as well as others.

In the following subsections we discuss a selection of cases illustrating different aspects of the dataset that shed light on some aspect of online propaganda. We discuss prominent "spikes"; brief periods of much escalated tweeting. We also briefly cover an attempt at a "scare" from 2014, before concluding with an analysis of the relevance of Russian Twitter propaganda to

[17] https://www.huffingtonpost.com/jeffrey-dwoskin/you-should-be-playing-has_b_7910728.html

[18] "TCOT" means "Top Conservatives on Twitter"; "PJNET" means "Patriot Journalist Network".

[19] https://gate-socmedia.group.shef.ac.uk/political-polarisation-disinformation-and-bots/ppp-supp-mats/

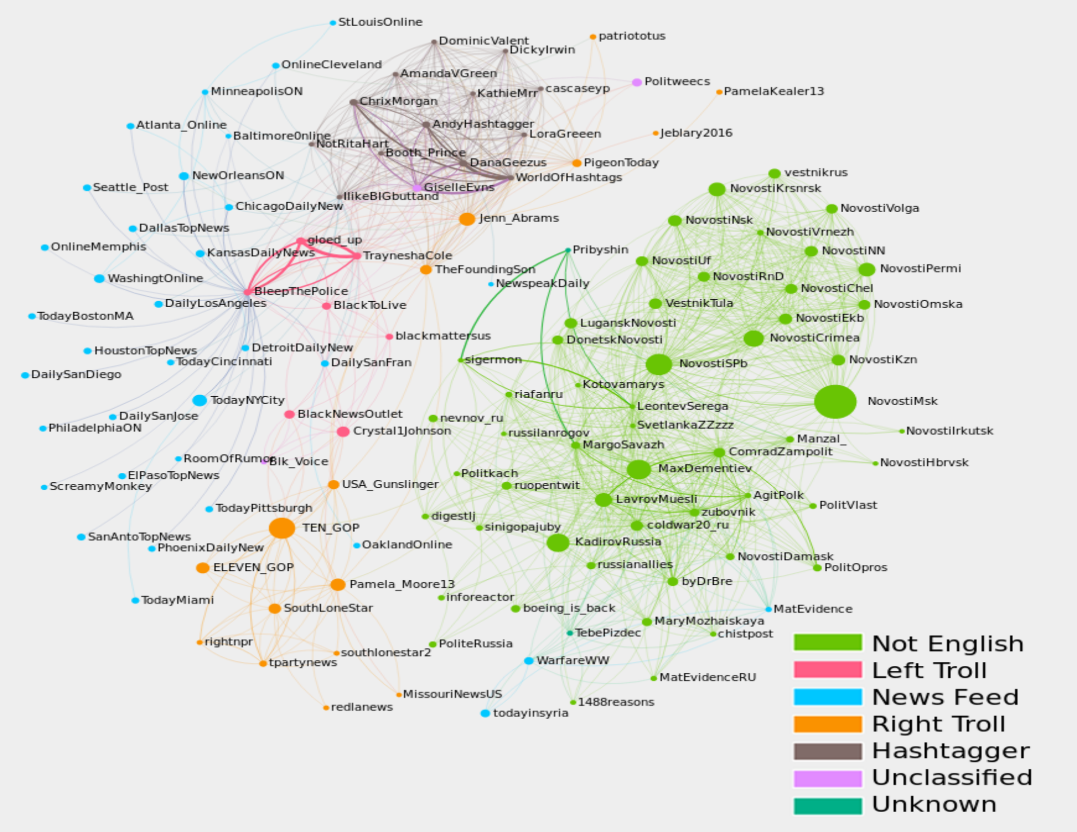


Figure 11: Network of IRA Troll Accounts

Name	Bio	Followers	Tweets
TEN_GOP	Unofficial Twitter of Tennessee Republicans. Covering breaking news, national politics, foreign policy and more. #MAGA #2A	147,767	10,794
Jenn_Abrams	Calm down, I'm not pro-Trump. I am pro-common sense. Any offers/ideas/questions? DM or email me jennabrams@gmail.com (Yes, there are 3 Ns)	79,152	25,378
Pamela_Moore13	Southern. Conservative. Pro God. Anti Racism	72,121	6,203
TodayNYCity	New York City's local news on Twitter. Breaking news, sports, events and international news. Tweet us or DM	66,980	59,420
ELEVEN_GOP	This is our back-up account in case anything happens to @TEN_GOP	59,279	115
wokeluisa	APSA. #Blackexcellence. Political science major	57,295	2,288
Crystal1Johnson	It is our responsibility to promote the positive things that happen in our communities.	56,581	7,915
SouthLoneStar	Proud TEXAN and AMERICAN patriot #2a #prolife #Trump2016 #TrumpPence16 Fuck Islam and PC. Don't mess with Texas!	53,999	3,600

Table 4: High Impact IRA Trolls

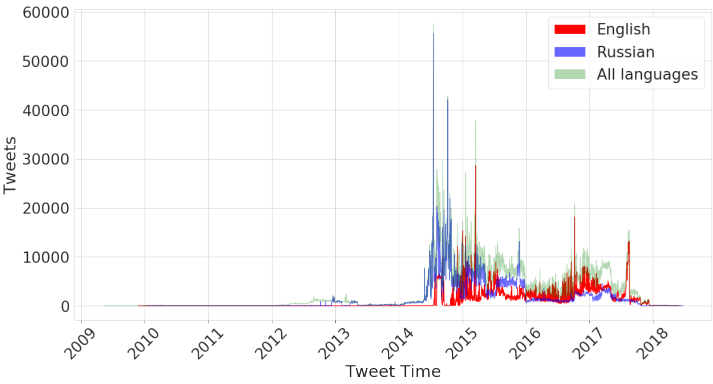


Figure 12: Timeline of Tweet Activity

Brexit.

Cases

There are three prominent spikes in activity among English language tweets, and three among the Russian ones, as can be seen in figure 12. The first and greatest of the English spikes shows little in the way of meaningful content. Impact (retweets) in this period was negligible despite a high number of original tweets. The second was timed well, in October 2016, as an attempt to influence Americans who would go to the polls to elect a new president the following month. The final

When	Lang	Tweet Total	% Retw	Retw Rec	Retw Rat
17-20 Jun 2014	Rus	118,219	17%	30,287	0.31
8-10 Oct 2014	Rus	70,233	44%	22,569	0.58
17-19 Mar 2015	Eng	57,710	1%	637	0.01
23-25 Nov 2015	Rus	28,252	72%	38,760	4.94
5-7 Oct 2016	Eng	31,111	90%	119,635	38.54
11-18 Aug 2017	Eng	95,112	36%	272,575	4.51

Table 5: Statistics of Tweet Spikes

of three spikes in English language tweets occurred in August 2017 and focuses on the incidents in Charlottesville (Phillips and Yi, 2018). Table 5 gives an overview of the spikes. "% Retw" gives the percentage of the tweets that were retweets of others, whereas "Retw Rec" gives number of times the tweets were retweeted, and "Retw Rat" gives the ratio of retweets to original tweets.

31,111 tweets were found in the set between October 5th and 7th 2016, which constituted the second largest English language "spike" in the dataset. It is evident in the table that the number of these tweets that were retweets was high, and at 90% much higher than the corpus-wide average rate of 38%. In the two day window from October 5th to 7th, almost half the tweets originated in the most active twenty accounts and consisted almost entirely of retweets. These accounts had on average 1,300 followers each. Prominent trolls continued their activity as usual during this period, and the top 15, which each had more than 500 retweets and are familiar, established accounts such as "TEN_GOP" and "Crystal1Johnson", achieved 98% of retweets (of original tweets) in this period. The retweet rate of original tweets in this period was 39 retweets per original tweet, which is much higher than the corpus-wide retweet rate of 3.46 retweets per original tweet. It is possible that the retweeting activity boosted the impact of the original tweets during this time; however the retweet quality is generally low and the retweets were not generally of other troll accounts. It is perhaps more likely that the political climate in this period enabled skilled political trolls to be particularly effective.

In the Charlottesville spike we again see the overwhelming majority of retweets achieved by a handful of prominent trolls. 97% of retweets were achieved by the 19 trolls with retweet counts over 500. Among those 19 we see familiar faces, who continued to operate as usual and with their usual high impact, most notably "TEN_GOP" who achieved 130,000 retweets in that period. However there is also a presence of a cluster of accounts that became active at the end of July 2017 and remained active for short durations only, often posing as patriotic, Trump-supporting individuals and notably giving as their profile URL a link to "ReportSecret.com", a now-defunct alternative news site also run by the Internet Research Agency. 65% of tweets in this period originated in accounts with "ReportSecret.com" profile URLs.

These accounts used IFTTT, a web scripting service, suggesting some degree of automation. Retweets created this way don't appear to Twitter as retweets, which means unlike normal retweets they receive a retweet count rather than passing it back to the actual author–we refer to them here as "manual retweets". Manual retweets are rare in the corpus, but became prevalent during this period, giving us an opportunity to calculate success rate with retweets, which normally isn't possible. 0.78 retweets per manual retweet were accrued during this period. (In the table above, retweets of manual retweets are excluded from the counts of retweets received, in order to make all periods comparable.) During the Charlottesville period, one of these accounts achieved 21,000 retweets, a return of four retweets per tweet, notable given that the account was active for only eighteen days but most likely arising from luck rather than skill given that the most successful tweets were retweets of other tweets. The tone of the material is pro-Trump, consisting of a fair percentage of original tweets, and retweets that are consistent with the message. A total of 72,847 manual retweets were found in the full dataset, of which 60,618 were created using IFTTT; more than half of those were found in the Charlottesville spike. IFTTT began being used to create tweets in late 2016; by the end of the dataset in the latter half of 2017 it was used to create half the tweets.

In contrast, a tweet set from a single day in September 2014 illustrates a further early unsuccessful attempt at influence. 8,520 tweets in total contained the hashtag "#ColumbianChemicals", spreading false rumours of an accident at a US chemical plant, and consisting of 275 tweets in Russian, most of which came earliest in the day, 3,119 tweets targeted at prominent individuals that achieved just eight retweets, 3,821 original tweets that achieved 1360 retweets, and 1305 retweets by the IRA trolls themselves, accounting for most of the retweets of original tweets. This attempt at a scare clearly fell flat. Here is an example tweet from the set:

> @BarackObama Barack , Are you kidding?? I saw the video #ColumbianChemicals and it looks like hell!!! What a nightmare!

IRA and Brexit

With regards to Brexit, we looked at tweets posted by the IRA accounts in our own Brexit tweet dataset in a one month period before the referendum. Furthermore, using our data, a further forty-five troll accounts were able to be identified and subsequently suspended by Twitter, in work described by Buzzfeed News.[20] Influence by those accounts was modest. Amongst the 3,200 total tweets, 830 came from the 45 newly identified accounts (26%). Brexit interest in the new corpus echoed previous findings provided in the Buzzfeed article showing little interest in advance of the referendum and a peak on the day of the referendum almost entirely in languages other than English, most notably German.

Table 6 shows all tweets posted one month before 23 June 2016, which were either authored by Russia Today or Sputnik, or are retweets of these. This gives an indication of how much activity and engagement there was around these accounts. To put these numbers in context, the table also includes the equivalent statistics for the two main pro-leave and pro-remain Twitter accounts. It is likely therefore that influence was modest (although real world influence is difficult to quantify, depending on factors such as who was reached).

[20] https://www.buzzfeed.com/tomphillips/we-found-45-suspected-bot-accounts-sharing-pro-trump-pro

Account	**Orig. tweets**	**Retweeted**	**Retweets**	**Replies**	**Total**
@RT_com	39	2,080	62	0	2,181
@RTUKnews	78	2,547	28	1	2,654
@SputnikInt	148	1,810	3	2	1,963
@SputnikNewsUK	87	206	8	4	305
TOTAL	352	6,643	101	7	7,103
@Vote_leave	2,313	231,243	1,399	11	234,966
@StrongerIn	2,462	132,201	910	7	135,580

Table 6: Russian Account Activity vs Campaign Sites

Automation in the Brexit Tweets

Automation is another area of concern with regards to propaganda, as it may be used to increase reach at low cost. We saw evidence above suggesting that it is difficult to achieve a high impact with automated accounts. However, other research finds a role for automated accounts in information spread (Shao *et al.*, 2018). With regards to Brexit, whilst it is hard to quantify automation among the accounts, Bastos and Mercea (2017) identified 13,493 suspected bot accounts, among which Twitter found only 1% to be linked to Russia. In our Brexit dataset there are tweets by 1,808,031 users in total, which makes these bot accounts only 0.74% of the total. If we consider Twitter accounts that have posted more than 50 times a day (widely considered to indicate a high degree of automation), then there are only 457 such users in the month leading up to the referendum on 3 June 2016. The most prolific accounts were "ivoteleave" and "ivotestay", both suspended, which were similar in usage pattern. There were also a lot of accounts that did not really seem to post much about Brexit but were perhaps using the hashtags in order to gain attention for commercial reasons. We also analysed the leaning of these 457 high automation accounts and identified 361 as pro-leave (with 1,048,919 tweets), 39 pro-remain (156,331 tweets), and the remaining 57 as undecided. This leaning towards leave echoes our above findings that the leave campaign was much more vocal on Twitter.

3.3 Post-Truth Politics–A Tale of Two Claims

The rise of post-truth politics has been linked to the lowered bar to publication offered by Web 2.0 and the consequent momentum that can be gained for organized disinformation campaigns (Faris *et al.*, 2017). A House of Commons Treasury Committee Report published on May 2016, states that: “The public debate is being poorly served by inconsistent, unqualified and, in some cases, misleading claims and counter-claims. Members of both the ‘leave’ and ‘remain’ camps are making such claims.” We analysed the number of Twitter posts around some of the these disputed claims. A study of the news coverage of the EU Referendum campaign established that the economy was the most covered issue, and in particular, the remain claim that Brexit would cost households £4,300 per year by 2030 and the leave campaign’s claim that the EU cost the UK £350 million each week. Therefore, we focused on these two key claims and analysed tweets about them.

With respect to the disputed £4,300 claim[21] (made by the Chancellor of the Exchequer), we identified 2,404 posts in our dataset (tweets, retweets, replies), referring to this claim. For the £350 million a week disputed claim (same reference) there are 32,755 pre-referendum posts (tweets, retweets, replies) in our dataset. This is 4.6 times the 7,103 posts related to Russia Today and Sputnik and 10.2 times more than the 3,200 tweets by the Russia-linked accounts suspended by Twitter.

In particular, there are more than 1,500 tweets from different voters *within our sample*, with one of these wordings:

> I am with @Vote_leave because we should stop sending £350 million per week to Brussels, and spend our money on our NHS instead.
>
> I just voted to leave the EU by postal vote! Stop sending our tax money to Europe, spend it on the NHS instead! #VoteLeave #EUreferendum

Many of those tweets have themselves received over a hundred likes and retweets each. This false claim is popularly regarded as one of the key ones behind the success of the leave campaign. Regarding the impact of these claims, a potentially useful indicator comes from an Ipsos Mori poll published on 22nd June 2016, which showed that for 9% of respondents the NHS was the most important issue in the campaign.

The leave claim notably appeared as a bus advert, so spreading its message to the voting public via a different channel. To assess the impact of this, the number of appearances of pictures of the red bus in our sample was counted; a high recall OCR step was followed by a manual classification to find these images. 913 images of the bus were found. Furthermore, 21,240 appearances of the leave claim in some form of image were found, using a fully automated OCR method with an F1 of 0.87, substantially increasing the textual count for that claim. Moore and Ramsay (2017) state that the remain claim was discussed in 365 newspaper articles, whereas the leave claim was discussed in only 147. The greater media interest in the Osborne claim is unsurprising given his position of authority, but this didn’t translate into interest on Twitter.

Note that not all Twitter discussion of the misleading headlines is uncritical propagation. The tweets often talk about the credibility of the headline. The 21,240 leave claim images were tweeted by 16,490 unique users. Of those, a higher number were remainers (5,369 vs. 4,950, with the remainder unclassified), suggesting a high proportion of Twitter interest in the claim was at least somewhat critical. Note also that although pictorial versions of the claim were tweeted by more remainers, the leavers that did tweet it tweeted it more; in terms of actual tweets containing pictures making the claim (buses as well as other imagery containing the claim) leavers accounted for 7531, compared with 6585 remainers, with the remainder unclassified, suggesting a greater enthusiasm for sharing the imagery among leavers, as one might expect. Recall that as we found above, our sample contains more remainers, but the leavers were more vocal. These findings recall Venturini (2019), who notes that the spreading of information is largely independent of whether the spreader actually believes it, and that this viral tendency and the resulting deluge of valueless information may be the more significant aspect of the problem. A similar

[21] https://www.kcl.ac.uk/sspp/policy-institute/CMCP/UK-media-coverage-of-the-2016-EU-Referendum-campaign.pdf

result is found when considering another prominent pictorial campaign; the UK Independence Party's poster showing a large queue of people alongside the slogan "Breaking Point" and the suggestion that "we must take back control of our borders". The poster has been criticised for implying that the people in the poster are entering the UK as immigrants, whereas in fact the picture was taken in Slovenia [22]. This claim was found in 3,388 tweets in pictorial form, of which leavers account for 948 and remainers, 1,007, the greater number, and the rest unclassified. In terms of unique users, 843 leavers posted the claim in image form and 890 remainers did so (1,331 unclassified). It is evident from the above that in this case, remainers repeated the leave claim more than leavers.

4 Discussion

We have presented evidence addressing the presence of partisan media, propaganda and post-truth politics in the run-up to the UK EU membership referendum on Twitter and in the media, as well as more broadly. With regards to partisanship in Brexit, we saw that websites linked in topically related tweets were most often neutral or bipartisan in their appeal. However, sources with **partisan** appeal also captured a sizeable portion of the debate, and of those, the leave-partisan materials were much more heavily propagated. Mainstream media with a stated remain stance produced materials appealing to both sides of the debate. Some mainstream media with a stated leave stance produced materials predominantly appealing to leavers.

A high degree of imbalance between leavers and remainers in those linking to a medium's website was found to suggest partisanship or even propaganda; materials with a strong appeal to leavers rather than remainers were plentiful and diverse, and included mainstream media and alternative media including US and other foreign sources. Materials with a strong appeal to remainers were fewer and less influential, and mainly comprised explicit campaign sites. Number of upheld press complaints correlates more strongly with a site's partisan appeal than the bias of the source as determined by the difference between its pro- and anti-Europe front pages (though both correlations are highly significant), suggesting that partisan appeal is capturing something other than the extent to which a source provides a voice for a particular opinion, and that misinformation may be a part of it. More datapoints would be desirable, however, to explore this more convincingly. Evidence of Russian state involvement was modest. Automated accounts were in evidence.

The main evidence presented regarding **propaganda** was taken from a dataset identified by Twitter as originating in the Russian Internet Research Agency, an organization known to seek global influence through the dissemination of propaganda materials. Observation of this data suggests a learning process on their part regarding how impact can effectively be achieved. Tapping into deeply felt issues such as Black equality and patriotism has allowed a few skilled agents to build a large following, accounting for by far the greater part of the IRA's reach. The appetite of the audience for a particular message might therefore be seen as the "Trojan Horse", via which the desired message may then be insinuated. Indeed some difficulty may arise in distinguishing the vehicle message from the propagandistic message that motivates the efforts. A good vehicle may bide its time, or indeed be an end in itself (for example leading to financial benefit through advertising revenue). Low effort approaches, such as possibly automated retweeting and large scale tweeting of pleasing but vague content, didn't appear to result in a high reach. One observed case of a fabricated scare fell entirely flat.

Future work exploiting this corpus should involve a deeper review of the Russian language IRA tweets. This would provide a greater understanding of the early history of an internet propaganda operation. Linked materials also provide more detailed material. The website "ReportSecret.com" has been highlighted above, along with other partisan press and alternative media in reference to the Brexit case. Furthermore the Russian accounts linked thousands of times to pages on the website LiveJournal, where extensive material more in the nature of personal opinion achieved a high reach; most-linked pages discuss the shooting down of Malaysian Airlines Flight 17, and are pro-Russian, anti-Ukraine. The material has provided an opportunity to benefit from the IRA's learning process in understanding how messages spread or fail to spread. However, the observations made here are preliminary only, and must form part of a more rigorous and complete picture formed of all available data, not just part, and backed up by controlled studies.

Claims made by leave and remain campaigns were reviewed in the context of **post-truth politics**. Echoing findings above, uptake of misleading leave claims was found to be high, dwarfing, for example, any evidence of Russian influence on Brexit. The greater hazard for public information may be the increasing tendency for public figures to take liberties with the truth.

A background issue through the findings is the issue of polarization. It has been observed that "social media prompt people to sort themselves into relatively closed communities of the like-minded, and encourage them to see things in a peculiarly urgent and intense way", furthermore noting that in a polarized climate, neutral media can struggle to retain an audience (Lynch, 2015). Nagle (2017) notes that the language of "transgression" can be turned to different, even opposing causes. This and other partisan mentalities find rich soil where some form of conflict or desire for change is already present. In the section on partisan media we found that the pro-remain Guardian newspaper attracted critical comment, which the Express did not do to the same extent, instead attracting upheld press complaints. This raises questions about the factors that encourage, or discourage, bipartisan discussion.

Highly partisan materials were found to be evident in great quantities in the form of linked materials in the Brexit tweet sample. Whilst these materials are of concern in that they are prolific and more often misleading, and are attracting significant attention, information consumers show a preference for linking more moderate materials, supporting previous research suggesting that there is a polarizing pull from those putting out their message on the internet. In the IRA materials we

[22] https://www.theguardian.com/politics/2016/jun/16/nigel-farage-defends-ukip-breaking-point-poster-queue-of-migrants

found that political trolls attracted the greatest following and achieved the greatest impact pushing at a small number of what might be seen as "open doors"; topics where feelings are already running high. These existing cracks in society may offer opportunities for those that wish to create further division.

The release of the IRA dataset by Twitter is an important step forward in platforms working together with scientists to enable a better understanding of the new social dynamic they have created. Controversial posts and accounts are suspended at a very high rate, creating an issue for open and repeatable science on social media data. However the dataset was limited in that follower/followee networks weren't included. Gaining a full picture requires access to all related data, not only tweets from a particular set of accounts. Similarly the impact of retweets cannot be understood without information about the retweet rate of retweets. Fully understanding impact requires information about how often a tweet appeared on someone's screen. Moving forward requires a careful debate about privacy. Failing to have that debate may result in information being richly available to those with commercial objectives, namely the platforms themselves, but denied to a society reeling from the effects.

As already discussed above, disinformation and biased content reporting are not just the preserve of fake news and state-driven propaganda sites and social accounts. A significant amount also comes from media and factually incorrect statements by prominent politicians. The impact of widely known and influential claims made by politicians from both sides of the referendum campaign was already discussed above. Therefore, effectively combating deliberate online falsehoods must address such cases. Furthermore transparency in political advertising on social platforms and a review process for political advertising are likely to help with reducing the impact of all other kinds of disinformation already discussed above (i.e. fake news sites, Russian propaganda, etc).

Acknowledgments

This work was partially supported by the European Union under grant agreements No. 825297 "WeVerify" and No. 654024 "SoBigData", the UK Engineering and Physical Sciences Research Council grant EP/I004327/1 and the British Academy under call "The Humanities and Social Sciences Tackling the UK's International Challenges".

References

Allcott, H. and M. Gentzkow (2017). "Social media and fake news in the 2016 election". *Journal of Economic Perspectives*. 31(2): 211–36.

Barberá, P. and G. Rivero (2015). "Understanding the political representativeness of Twitter users". *Social Science Computer Review*. 33(6): 712–729.

Bastos, M. T. and D. Mercea (2017). "The Brexit Botnet and User-Generated Hyperpartisan News". *Social Science Computer Review*.

Faris, R., H. Roberts, B. Etling, N. Bourassa, E. Zuckerman, and Y. Benkler (2017). "Partisanship, Propaganda, and Disinformation: Online Media and the 2016 US Presidential Election". *Berkman Klein Center for Internet & Society Research Paper*.

Ferrara, E. (2017). "Disinformation and social bot operations in the run up to the 2017 French presidential election". *First Monday*. 22(8).

Hare, C. and K. T. Poole (2014). "The polarization of contemporary American politics". *Polity*. 46(3): 411–429.

Higgins, K. (2016). "Post-truth: a guide for the perplexed". *Nature News*. 540(7631): 9.

Howard, P. N. and B. Kollanyi (2016). "Bots, #strongerin, and #brexit: Computational propaganda during the UK-EU referendum". *Tech. rep.* Working Paper 2016.1. Oxford, UK: Project on Computational Propaganda.

Kaminska, M., B. Kollanyi, and P. N. Howard (2017). "Junk News and Bots during the 2017 UK General Election: What Are UK Voters Sharing Over Twitter?" *Tech. rep.* Data Memo 2017.5. Oxford, UK: Project on Computational Propaganda.

Lansdall-Welfare, T., F. Dzogang, and N. Cristianini (2016). "Change-point analysis of the public mood in UK Twitter during the Brexit referendum". In: *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*. IEEE. 434–439.

Lewandowsky, S., U. K. Ecker, and J. Cook (2017). "Beyond misinformation: Understanding and coping with the "post-truth" era". *Journal of Applied Research in Memory and Cognition*. 6(4): 353–369.

Linvill, D. and P. Warren (2018). "Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building". *pwarren.people.clemson.edu/Linvill_ Warren_ TrollFactory.pdf*.

Lynch, M. (2015). "After the Arab Spring: How the Media Trashed the Transitions". *Journal of Democracy*. 26(4): 90–99.

Mangold, L. (2016). "Should I stay or should I go: Clash of opinions in the Brexit Twitter debate". *Computing*. 1(4.1).

Maynard, D., I. Roberts, M. A. Greenwood, D. Rout, and K. Bontcheva (2017). "A framework for real-time semantic social media analysis". *Web Semantics: Science, Services and Agents on the World Wide Web*. 44: 75–88.

Moore, M. and G. Ramsay (2017). *UK media coverage of the 2016 EU referendum campaign*. King's College London.

Nagle, A. (2017). *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. John Hunt Publishing.

Narayanan, V., P. N. Howard, B. Kollanyi, and M. Elswah (2017). "Russian Involvement and Junk News during Brexit". *Tech. rep.* Data Memo 2017.10. Oxford, UK: Project on Computational Propaganda.

Persily, N. (2017). "The 2016 US Election: Can democracy survive the internet?" *Journal of democracy*. 28(2): 63–76.

Phillips, J. and J. Yi (June 2018). "Charlottesville Paradox: The 'Liberalizing' Alt-Right, 'Authoritarian' Left, and Politics of Dialogue". *Society*. 55(3): 221–228. ISSN: 1936-4725. DOI: 10.1007/s12115-018-0243-0. URL: https://doi.org/10.1007/s12115-018-0243-0.

Preoţiuc-Pietro, D., Y. Liu, D. Hopkins, and L. Ungar (2017). "Beyond binary labels: political ideology prediction of Twitter users". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 729–740.

Shao, C., G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer (2018). "The spread of low-credibility content by social bots". *Nature communications*. 9(1): 4787.

Silverman, C. (2015). "Lies, damn lies and viral content". *Tech. rep.* Tow Center for Digital Journalism.

Skjeseth, H. T. (2017). "All the president's lies: Media coverage of lies in the US and France". *Tech. rep.* Reuters Institute for the Study of Journalism, University of Oxford.

Venturini, T. (2019). "From Fake to Junk News, the Data Politics of Online Virality". In: *Data Politics: Worlds, Subjects, Rights*. Ed. by D. Bigo, E. Isin, and E. Ruppert. London: Routledge.

Vosoughi, S., D. Roy, and S. Aral (2018). "The spread of true and false news online". *Science*. 359(6380): 1146–1151.

Yin, L., F. Roscher, R. Bonneau, J. Nagler, and J. A. Tucker (2018). "Your Friendly Neighborhood Troll: The Internet Research Agency's Use of Local and Fake News in the 2016 US Presidential Campaign". *SMaPP Data Report. New York: Social Media and Political Participation Lab, New York University.*