Twits, Twats and Twaddle: Trends in Online Abuse towards UK Politicians

Genevieve Gorrell, Mark A. Greenwood, Ian Roberts,
Diana Maynard and Kalina Bontcheva
University of Sheffield, UK
g.gorrell, m.a.greenwood, i.roberts,
d.maynard, k.bontcheva@sheffield.ac.uk

Abstract

Concerns have reached the mainstream about how social media are affecting political outcomes. One trajectory for this is the exposure of politicians to online abuse. In this paper we use 1.4 million tweets from the months before the 2015 and 2017 UK general elections to explore the abuse directed at politicians. Results show that abuse increased substantially in 2017 compared with 2015. Abusive tweets show a strong relationship with total tweets received, indicating for the most part impersonality, but a second pathway targets less prominent individuals, suggesting different kinds of abuse. Accounts that send abuse are more likely to be throwaway. Economy and immigration were major foci of abusive tweets in 2015, whereas terrorism came to the fore in 2017.

Introduction

The UK EU membership referendum and the US presidential election, among other recent political events, have drawn attention to the power of social media usage to influence important international outcomes. Such media profoundly affect our society, in ways which are yet to be fully understood. One particularly unsavoury way in which people attempt to influence each other is through verbal abuse and intimidation.

There is a broad perception that intolerance, for example religious or racial, is on the increase in recent years.[1] In the UK, the outcome of the EU membership referendum, in which the British public chose to leave the EU, was also associated with the legitimisation of racist attitudes and an ensuing increased expression of those attitudes.[2] Twitter provides a window on these mindsets, providing a forum where users can communicate their message to public figures with relatively little personal consequence.

In this work we explore a collection of abusive replies to tweets by UK politicians in the run-up to the 2015 and 2017 UK general elections. These data allowed us to investigate what influences the amount of abuse a politician receives, what can we learn about those who send abuse, what are the topics of concern to those who send abuse and what difference we see in abuse over time. Previous work has examined abusive behaviour online towards different groups, but the reasons why a politician might inspire an uncivil response are very different to an ordinary member of the public, with resulting different implications for democracy. To the best of our knowledge, this is the first study to contrast quantitative changes in this across two comparable but temporally distinct samples (the two general election periods).

[1] https://www.opendemocracy.net/transformation/ae-elliott/assemble-ye-trolls-rise-of-online-hate-speech

[2] https://www.theguardian.com/commentisfree/2016/jun/27/brexit-racism-eu-referendum-racist-incidents-politicians-media

Related Work

Whilst online fora have attracted much attention as a way of exploring political dynamics (Nulty et al. 2016; Colleoni, Rozza, and Arvidsson 2014), and the effect of abuse and incivility in these contexts has been explored (Vargo and Hopp 2017; Rüsel 2017), little work exists regarding the abusive and intimidating ways people address politicians online - a trend that has worrying implications for democracy. Theocharis *et al* (2016) collected tweets centred around candidates for the European Parliament election in 2014 from Spain, Germany, the United Kingdom and France posted in the month surrounding the election. They find that the extent of the abuse and harrassment a politician is subject to correlates with their engagement with the medium. Their analysis focuses on the way in which uncivil behaviour negatively impacts on the potential of the medium to increase interactivity and positively stimulate democracy. Stambolieva (2017) studies online abuse against female Members of Parliament (MPs) only; in studying male MPs as well, we are able to contrast the level of abuse they each receive. Furthermore, we contrast proportional with absolute figures, creating quite a different impression from the one she gives. A larger body of work has looked at hatred on social media more generally (Bartlett et al. 2017; Coe, Kenski, and Rains 2014; Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015).

Data Collection

The corpus was created by downloading tweets in real-time using Twitter's streaming API. Tweets posted from the end

	#collected	#hadabuse	%abusive
2015	597 411	16 628	2.8%
2017	821 662	32 791	4%

Table 1: Corpus statistics

of May 6th to the end of June 6th 2015, and April 7th to May 7th 2017 (the day before each election) were collected. We obtained a list of all current MPs[3] and all currently known election candidates[4] (at that time) who had Twitter accounts. For 2015, that was 506 currently elected MPs and 1811 candidates, of whom 444 MPs were also standing for re-election, and for 2017, 1952 candidates and 480 sitting MPs, of whom 417 were also candidates. We collected every tweet by each of these users, and every retweet and reply (by anyone). Data were of a low enough volume not to be constrained by Twitter rate limits. Numbers of tweets thus collected are given in table 1, along with the percentage of abusive tweets.

In order to identify abusive language, its targets and topics, we use a set of NLP tools, combined into a semantic analysis pipeline. Topic detection finds mentions in the text of political topics (e.g. environment, immigration). The list of topics was derived from the set used to categorise documents on the gov.uk website,[5] first seeded manually and then extended semi-automatically as described by Maynard *et al* (2017). We also perform hashtag tokenization, to find abuse and threat terms that otherwise would be missed; for example in the hashtag "#killthewitch".

A dictionary-based approach was used to detect abusive language in tweets. An abusive tweet is considered to be one containing one or more abusive terms from the vocabulary list.[6] This contains 404 abuse terms in British and American English, comprising mostly an extensive collection of insults, with a few threat terms such as "kill" and "die" also included. Racist and homophobic terms are included as well as terms that denigrate a person's appearance or intelligence. In this way, abuse is broadly defined here.

Data from Kaggle's 2012 challenge, "Detecting Insults in Social Commentary"[7], were used to evaluate the success of the approach, demonstrating an accuracy of 0.78 (Cohen's Kappa: 0.37), with a precision of 0.61, a recall of 0.44 and an F1 0.51. This performance is comparable to that obtained by Stambolieva (2017). Manual review of the errors shows some false positives particularly on threat terms, but no evidence of any particular bias that might affect the results reported here.

[3]From a list made publicly available by BBC News Labs, which we cleaned and verified

[4]List of candidates obtained from `https://yournextmp.com`

[5]e.g. `https://www.gov.uk/government/policies`

[6]Warning; strong language and offensive slurs: `http://www.dcs.shef.ac.uk/~genevieve/publications-materials/abuse-terms.txt`

[7]`https://www.kaggle.com/c/detecting-insults-in-social-commentary/data`

Who is receiving the abuse?

Table 1 reveals that both in terms of quantity and proportion, abuse increased between the two elections. In this section, we investigate the factors at play in individuals receiving the quantity of abuse they do. Multiple factors may be involved in this. Structural equation modeling (SEM, see Hox and Bechger (2007) for an introduction) offers the possibility of producing an overall theory relating multiple factors. For example, gender may confound an attempt to relate party membership to abuse received, since there are more women in the Labour party, and this may account for any relationship seen, but SEM can accommodate this. SEM was used here to broadly relate three main factors with the amount of abuse received: prominence, Twitter prominence (which we hypothesise differs from prominence generally) and Twitter engagement. We obtained Google Trends data for the 50 most abused MPs in each of the time periods, and used this variable as a measure of how high-profile that individual is in the minds of the public at the time in question. Search counts for the month running up to each election were totalled to provide a figure. We used number of tweets sent by that politician as a measure of their Twitter engagement, and tweets received as a measure of how high-profile that person is on Twitter. The model in figure 1, in addition to proposing that the amount of abuse received follows from these three main factors, also hypothesises that the amount of attention a person receives on Twitter is related to their prominence more generally, and that their engagement with Twitter might get them more attention, both on Twitter and beyond it. It is unavoidably only a partial attempt to describe why a person receives the abuse they do, since it is hard to capture factors specific to that person, such as any recent allegations concerning them, in a measure. The model was fitted using Lavaan,[8] resulting in a chi-square with a p-value of 0.403 (considered satisfactory, see Hox and Bechger (2007)), and shows a number of significant findings (indicated with a bold line and asterisks against the regression figure). Positive numbers indicate a positive relationship, and negative ones, a negative relationship.

The model shows that a strong pathway to receiving more abuse on Twitter is simply that if a person is well-known, they receive a lot of tweets (*attention* relates positively and significantly with *Twitter attention*), and if they receive a lot of tweets, they receive a lot of abusive tweets, in absolute terms (*Twitter attention* relates positively and significantly with *abusive tweets received*). However, an additional pathway shows that having removed this numbers effect from consideration, being very well known leads to a person being *less* likely to receive abuse on Twitter (*attention* relates negatively and significantly with *abusive tweets received*). Perhaps to certain senders of abuse, a large target is a less attractive one. A further pathway positively relates Twitter engagement to abuse received, supporting Theocharis *et al*'s (2016) findings. In this case, perhaps it is what the person said that provides an attractive target for abuse. The suggestion is of different types of abuse.

The effects that are not significant are also interesting.

[8]`http://lavaan.ugent.be/`

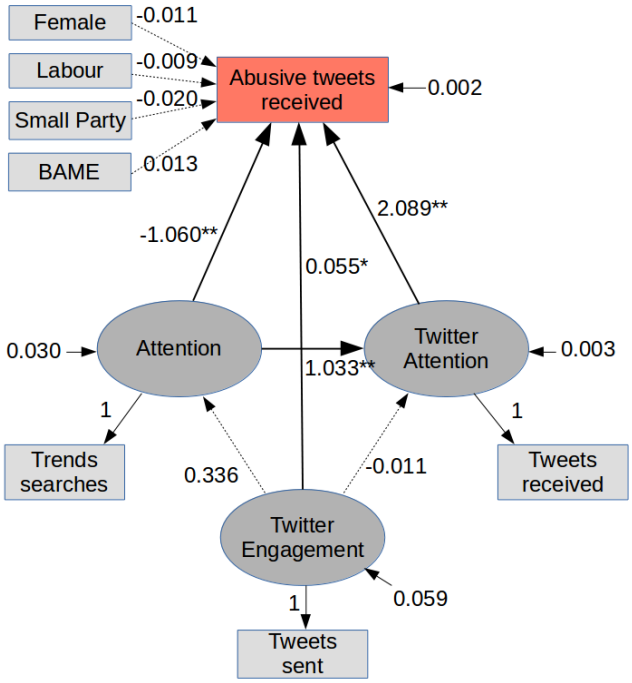


Figure 1: Abuse per MP in 2017

For example, engaging more with Twitter does not relate with getting more attention on Twitter. The impact of gender, party and ethnicity is, though somewhat telling, uncompelling; being a male, a Conservative or of an ethnic minority may tend to increase abuse received. The data are available in the form of an interactive graph [9] covering all politicians that went on to be elected, that the reader is recommended to explore. A tendency for males and Conservatives to receive more abuse is evident in these visualisations, and indeed t-tests find both of these relationships significant ($p<.001$), but the SEM analysis suggests that other factors may account for the bulk of it. A larger sample size may also be necessary for a significant result.

Who is sending the abuse?

To examine the behaviour of those who send abuse, 2506 Twitter accounts were selected from our 2017 dataset who have sent at least three abuse-containing tweets. A random sample of 2500 tweeters for whom we found no abusive tweets were then selected to form a contrast group.

Independent samples t-tests revealed that those who tweeted abusively have more recent Twitter accounts by a few months (1533 days on average vs 1608, $p<.001$), smaller numbers of favourited tweets (7379 vs 14596, $p<.001$), fewer followers (1085 vs 3260, $p<.05$), follow fewer accounts (923 vs 1472, $p<.05$), are featured in fewer lists (23 vs 67, $p<.001$) and have fewer posts (16445 vs 25258, $p<.001$). After partialing out account age, number of abusive tweets still correlated significantly with number of favourites ($p<.001$), number of followed accounts ($p<.001$), number of times listed ($p<.01$) and number of posts ($p<.001$), demonstrating that with the exception of follower number, these relationships cannot be explained by account age. One explanation for these findings would be that a certain number of accounts are being created for the purpose of sending anonymous abuse.

A similar analysis of accounts that sent abusive tweets in the lead-up to the 2015 general election revealed some differences compared with 2017. Firstly, whilst abuse-posting accounts were again younger, in 2015 they posted more (12732 statuses on average vs 8752, $p<.001$) and favourited more (2499 vs 1517, $p<.001$). Reviewing the data reveals that in 2015 more abuse was sent by a smaller number of individuals, including a substantial number of what might be termed serial offenders, who perhaps explicitly seek attention by posting and favouriting. The greater quantity of abuse found in the 2017 data is more thinly spread across a larger number of lesser offenders. Twitter's commitment to reviewing and potentially blocking abusive users might account for the difference in the two years. In fact, given that Twitter is now more engaged in blocking users, the increase in abuse between the two years may be even greater than that indicated in this work. Manual review of the data shows no evidence of "bots" (automated accounts) in the sample. Though bot activity is common in Twitter political contexts (Kollanyi, Howard, and Woolley 2016), we suggest that bots are perhaps unlikely to use abusive language.

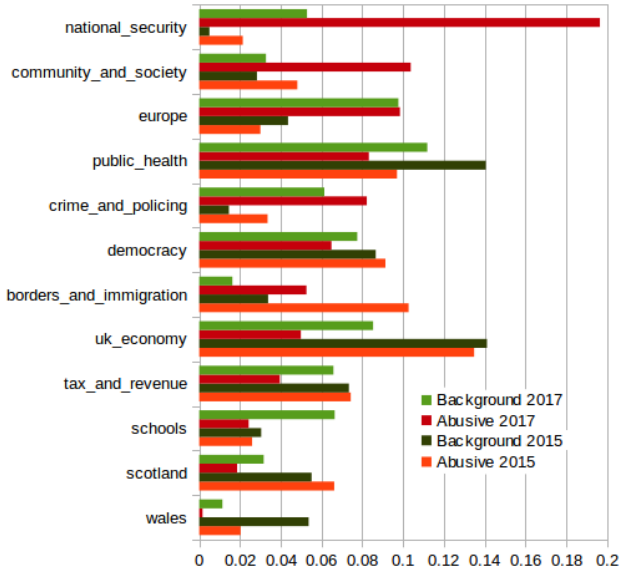
In both 2015 and 2017 datasets we found that significantly more accounts had been closed from the group that sent abusive tweets; 16% rather than 6% ($p<.01$) in 2015 and 8% rather than 2% ($p<.001$) in 2017 (Fisher's exact test).

Topics Triggering Abusive Replies

Examining topics mentioned in abusive tweets may provide insights into what is motivating the abuse. Mentions of the predetermined topics described earlier were counted in the tweets. Figure 2 presents topics accounting for at least 5% of total topic mentions in at least one of four sets; abusive tweets in 2015, abusive tweets in 2017, all tweets in 2015 or all tweets in 2017. Topic titles are generally self-explanatory, but a few require clarification. "Community and society" refers to issues pertaining to minorities and inclusion, and includes religious groups and different sexual identities. "Democracy" includes references to the workings of political power, such as "eurocrats". "National security" mainly refers to terrorism, where "crime and policing" does not include terrorism. "Public health" in the UK focuses on the National Health Service (NHS). The graph shows that national security dominates abusive tweets in 2017, despite attracting much less attention in tweets generally. A similar result on a smaller scale is visible for "community and society". Note that in the month preceding the 2017 election the UK witnessed its two deadliest terrorist attacks of the decade, both attributed to ISIS. In 2015, the most common topic in abusive tweets was the economy. However, this reflects the general level of interest in the economy at

[9] http://demos.gate.ac.uk/politics/buzzfeed/sunburst.html

Figure 2: Topics in Abusive vs All Responses



that time, and isn't disproportionate. Borders and immigration, however, is the second most prominent topic in abusive tweets, and is much less prominent in tweets generally. Note that a key 2015 election topic was the holding of an EU membership referendum, considered to have implications for immigration.

Discussion and Future Work

This work provides an empirical contribution to the current debate on abuse of politicians online. Abuse directed at politicians has increased in recent years, both in volume and proportionally, despite Twitter's greater activity in banning abusive use. Abuse received relates strongly with tweets received, suggesting such behaviour is for the most part impersonal. However, whilst more prominent politicians receive more tweets and therefore more abusive tweets by volume, within that there is a tendency for more prominent politicians to receive *less* abuse, suggesting a certain motivation to abuse that prefers smaller targets. Male MPs and Conservatives may receive more abuse. Users who send abuse show more evidence of using throwaway accounts. In 2015, immigration was a major topic of concern among those sending abuse, whereas in 2017, terrorism was.

The lack of evidence of increased abuse toward women politicians is in keeping with the result for the general population reported by Pew Internet Research,[10] who note that whilst men receive more abuse, women are more likely to be subject to online stalking and sexual harrassment; a distinction that wasn't made in this work.

The data from this study are available in the SoBigData catalogue,[11] entitled "UK election abuse data". More extensive analysis of the data can be found in a longer related work, also by Gorrell *et al* (2018).

[10] http://www.pewinternet.org/2014/10/22/online-harassment/

Acknowledgments

This work was partially supported by the European Union under grant agreements No. 610829 DecarboNet and 654024 SoBigData, the UK Engineering and Physical Sciences Research Council (grant EP/I004327/1), and by the Nesta-funded Political Futures Tracker project.[12]

References

Bartlett, J.; Reffin, J.; Rumball, N.; and Williamson, S. 2017. Anti-social media. Technical report, Demos.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial behavior in online discussion communities. In *ICWSM*, 61–70.

Coe, K.; Kenski, K.; and Rains, S. A. 2014. Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication* 64(4):658–679.

Colleoni, E.; Rozza, A.; and Arvidsson, A. 2014. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication* 64(2):317–332.

Gorrell, G.; Greenwood, M.; Roberts, I.; Maynard, D.; and Bontcheva, K. 2018. Online abuse of uk mps in 2015 and 2017: Perpetrators, targets, and topics. *arXiv preprint arXiv:1804.01498* 1804.

Hox, J. J., and Bechger, T. M. 2007. An introduction to structural equation modeling.

Kollanyi, B.; Howard, P. N.; and Woolley, S. C. 2016. Bots and automation over twitter during the first us presidential debate. *Comprop Data Memo*.

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

Nulty, P.; Theocharis, Y.; Popa, S. A.; arnet, O.; and Benoit, K. 2016. Social media and political communication in the 2014 elections to the european parliament. *Electoral studies* 44:429–444.

Rüsel, J. T. 2017. Bringing civility back to internet-based political discourse on twitter research into the determinants of uncivil behavior during online political discourse. Master's thesis, University of Twente.

Stambolieva, E. 2017. Methodology: detecting online abuse against women mps on twitter. Technical report, Amnesty International.

Theocharis, Y.; Barberá, P.; Fazekas, Z.; Popa, S. A.; and Parnet, O. 2016. A bad workman blames his tweets: the consequences of citizens' uncivil twitter use when interacting with party candidates. *Journal of communication* 66(6):1007–1031.

Vargo, C. J., and Hopp, T. 2017. Socioeconomic status, social capital, and partisan polarity as predictors of political incivility on twitter: A congressional district-level analysis. *Social Science Computer Review* 35(1):10–32.

[11] https://sobigdata.d4science.org/

[12] http://www.nesta.org.uk/news/political-futures-tracker