# Foreign Influence Efforts (FIE) on Twitter: Identifying Signals to Detect Defamatory Strategies Before and After the 2016 U.S. Presidential Elections

K. Lomicka, C. Maine, D. Miftari

Text Mining for Social Sciences
Barcelona Graduate School of Economics
June 21, 2020

# Contents

# 1 Introduction

Political Influence Campaigns by governments to influence public opinion and election outcomes in another country represent just one way in which governments have effected regime change or intervened in the affairs of another country to domestic advantage. However, with growing evidence that the world's most powerful country (the U.S.) has been under attack more recently (especially in the 2016 U.S. presidential elections), it is unsurprising that world media attention and academic funding has increased its focus on this topic. Foreign influence is of importance because it threatens a nation's sense of sovereignty and right to self-determination, which are core democratic values. In some cases, a power imbalance may be exploited in order to achieve change favourable to the foreign power, and this can also lead to a sense of exploitation by the receiving state. According to Martin and Shapiro 2019, foreign influence efforts (FIE) can be categorized into 5 strategies: defamation, persuasion, polarization, shifting political agendas, and undermining institutions. This paper looks to further analyze defamation attempts by Russia through in the 2016 US elections by combining text analysis, unsupervised learning through topic modeling, and supervised learning through sentiment analysis.

We used Twitter's Election Integrity Data Archive to access public datasets suspected of conducting malicious activity in effort to influence foreign elections. Twitter is a social media platform which enables users to communicate short statements (a tweet) of 140-character limit to their audience and enables them to reach millions of individuals within seconds. The platform is accessible to everyone through its website or mobile application. The selected dataset for this paper originates from Russia and is believed to have malicious or fake account activity. The tweets were cleaned and preprocessed to fit the needs of our model with the final version consisting of 540,953 tweets by 320 accounts. The selected time frame of the tweets was between July 1st, 2014 and October 1st, 2017.

We first used MALLET (MAchineLearning for LanguagE Toolkit) by McCallum 2002 which is a quick topic model implementation giving a snapshot of topics in a corpus. Using MALLET, we could explore and identify the optimal number of topics given our unique dataset and verify the content was relevant to the 2016 US elections. To model defamation in particular, we created an index to measure the extent of defamation in tweets targeted to a specific person over time. We created the index by identifying key politicians relevant to our data and combining that with a defamation dictionary of negative, emotional and offensive language used in Donald Trump's tweets. We measured defamation by comparing language similarity between a tweet and the aforementioned defamation dictionary, followed by calculating the frequency of politician references in each tweet. Our results show many defamation attempts for the politicians identified in our corpus. Hillary Clinton had the highest defamation score with tweets targeted towards her having the second highest similarity to defamatory language.

We structured the rest of this paper as follows. Given the focus of US national elections, Chapter 2 reviews literature from a social science perspective. Chapter 3 describes the data used in the paper, the steps in the text prepocessing and cleaning strategy, and tools used for data exploration. Chapter 4 dives into the approach and measures taken to detect defamation in the data using a mixture of supervised and unsupervised learning. Chapter 5 concludes.

# 2 Literature Review

A systematic content analysis of news articles reporting on cyber interference activity concluded that social media manipulation is widespread, organised, permanent and funded. (Bradshaw and Howard 2018, Hindman and Barash 2018, Martin and Shapiro 2019) Coordinated manipulation campaigns are taking place domestically in every type of political regime. Foreign operations have targeted several Western and emerging democracies during recent elections. Additionally, almost every democracy in this sample has organized social media campaigns that target foreign public.

Although motivations for FIE are rarely transparent, the appeal of social media as a powerful tool for advertisers, politicians and foreign governments is undeniable. Social media reaches huge swathes – and a large diversity – of the population. The technology collects and infers intimate details about a person, allowing segmentation and micro-targeting of messages, which can be continuously iterated to improve their effectiveness through ongoing trials.

Revenue generation is an important motivation for those who use social media to pursue their goals; creating material as 'clickbait' or copying others' news stories (Cagé et al. 2019) to gain advertising revenue or tailoring news stories towards the tastes of audiences who will pay for content (Gentzkow and J. Shapiro 2006). In terms of motivation for foreign intervention however, it seems unlikely that a foreign power, presumably with far more direct and effective means to raise money, would be using social media for revenue generation. Influence of public opinion, and potentially election outcomes, appears a far more likely goal for intervention. However, it can be difficult to differentiate between interventions designed to cultivate 'clicks,' and those by foreign powers aiming to achieve political influence, given that the psychological mechanisms by which they work can be similar, and the outcomes the same.

Psychological processes underlie the mechanism by which social media influence campaigns work. (Shu et al. 2017) 'Naïve realism' is the process whereby people think that their own views are true, and those who disagree are misinformed, uninformed or biased. (Ward et al. 1997) Confirmation bias is the process by which individuals show preference for information which confirms their existing views. (Nickerson 1998) Both of these phenomena imply that people are more likely to be affected by content that broadly aligns with their views, and more likely to discount that with which they disagree. This idea that view reinforcement is more important than an aversion to opinion challenge is reinforced by an online behaviour-tracking study recruited from the readership of two partisan online news sites. (Garrett 2009)

Moving online, a survey of U.S. adults estimated that the average US adult might have seen perhaps one or several news stories in the months before the 2016 election. (Allcott and Gentzkow 2017) Of those who saw, and believed, fake news stories, those who are heavy media users and those with segregated social networks are more likely to believe articles that reflect their own beliefs.

In further support of this theory is evidence which indicates that the politically interested and those who consume a range of media avoid echo chambers (Dubois and Blank 2018), and the findings that local newspapers and non-partisan television news are central to the public's media environment. Additionally, partisan individuals do not seem to engage in active avoidance of disagreeable information and share media diets that are broadly similar to each other. (Weeks et al. 2016) These findings may imply that it is in the processing of information, not its selective exposure, in which social media campaigns exert their influence.

There is evidence to support the usage of the media in achieving real political outcomes (either wittingly or unwittingly.) The introduction of more traditional media sources in an area can affect that area's political outcomes, for example the introduction of Fox News into the cable market impacted vote share in Presidential elections and voter turnout. (Dellavigna and Kaplan 2007) Evidence that online content exposure impacts offline behaviour was inferred by Mmller and Schwarz 2017 looking at the co-occurrence of hate crimes and Facebook outages (but not other social media server outages) in Germany. Transference of non-relevant information to influence a person's judgement in another context and framing of issues and decisions by information favouring one conclusion over another could be responsible for these kinds of impacts of social media absorption in more general life and political contexts.

In terms of how disinformation spreads through a network, data collected from Twitter in September and November 2016 sourced from accounts identified as Russian bots, showed that conservatives retweeted Russian Trolls significantly more often than Democrats and produced 36 times more tweets. (Badawy et al. 2018) Given that higher numbers of bots were estimated to be conservative than liberal (4.9 percent to 6.2 percent,) this could just be a reflection that the conservatives had been fed higher volumes of tweets to which they are agree rather than a propensity amongst those who are differently ideologically aligned to retweet at different rates.

Defamation, the focus of this paper, is not a new phenomenon. Defamation is defined as 'A direct attack against a person, intended to discredit him or her.' The Roman Empire created laws to protect defamation; to strike a balance between the need to protect personal character and public institutions from destructive attack whilst also protecting freedom of thought and the benefit of public discussion. The movement to the online realm has caused huge challenges to those wishing to defend against defamation due to the protection of anonymity afforded by the internet. There is a legal and PR industry dedicated to the protection of reputation online. High profile defamation cases, such as Susan B. Anthony List v. Driehaus in the U.S., and the U.K. Defamation Act 2013 have placed defamation laws in the spotlight, including controversial discussions as to the responsibilities of internet platforms.

It is true to say that defamation laws in the U.K. and U.S. largely favour the defendant. The prosecution bears the responsibility to prove serious harm, whilst the defendant may claim honest opinion, truth or even substantive truth in defence. This position reflects the value western countries place on Freedom of Speech. However, it also provides no real legal framework to defend against the type of semi-truth or exaggerated truth, continuous feed of information typical of defamatory social media campaigns. This is despite the fact that the political impact of these campaigns can be very real over the medium or long term.

In this context, characterisation and identification of defamation at the individual tweet (or even account level) is difficult. Instead, defamatory language content and topic patterns can be analysed over time to identify likely defamatory content. Even so, this approach is complicated by the lack of clarity as to which accounts are FIE and not-FIE, and therefore the lack of any kind of certainty of features which distinguish the two. Plus the overlap between the two sets of content in the first place, which is exacerbated by the use of real people, 'trolls,' to generate FIE content.

While the social science and legal literature on defamation is more developed, there is little technical literature specifically exploring measures to identify online defamation strategies and defamatory language. That said, the literature dedicated to identifying and understanding political sentiment or ideology is more robust. For example, Gentzkow and J. Shapiro 2010 introduce a new index of media

slant that measures the similarity of a news outlet's language to that of congressional Republicans and Democrats. In order to measure slant, common phrases used by both parties were identified by analyzing speeches from the *2005 Congressional Record*. Term frequency vectors were created for the two and three word phrases by speaker and each speech was labeled with the speaker's political party. Feature selection was done using phrase coocurrance and Pearson's $\chi^2$ coefficient to identify the most impactful phrases. The observed term-political party relationships were used to infer the ideology of newspapers by looking at whether the newspaper tends to favor phrases used by one party or another. Similarly, Nouh et al. 2019 also used the approach of analyzing a proxy text of known ideology to develop a means to identify Islamic extremist content on Twitter. The authors first built a radical language model using both TF-IDF and word2vec word embeddings based on articles from Dabiq extremist magazines. Complementing to the language model, additional features were crafted to pick up the emotional and behavioral signals. Examples include characteristic emotional or offensive language from existing NLP dictionaries and behaviors such as frequency of tweets posted. A binary random forest classifier was then trained to identify radical content on datasets of known extremist and non-extremist tweets. Lastly, Miller 2019 also leveraged emotional dictionaries in conjunction with an LDA topic model to characterize the topics and temporal changes in emotional sentiment of Russian-government sponsored twitter activity around the 2016 election.

# 3 Data

## 3.1 Political Context and Twitter's Elections Integrity Archive

The 2016 US elections were a turning point for American politics when Donald Trump, a New York real estate mogul and reality TV star, become the 45th President of the United States and defeated Hillary Clinton, the former first lady and Secretary of State. Trump led a populist, nationalist campaign and still won 304 votes in the Electoral College, compared to 227 votes for Clinton. In contrast, the popular vote showed different results, with Clinton winning almost 3 million more votes than Trump. These results made Clinton the 5th US presidential candidate in history to win the popular vote and lose the Electoral College. The elections were not important only because of the controversial results but were also infamously known for Trump's late nights on Twitter and the role of social media in influencing the eventual vote of swing voters. Although it is not clear how voters are influenced online, one method widely used is attempting to damage the reputation of running candidates, defamation, in favour of another.

A prominent channel for defamation during elections is Twitter, a social media platform which allows users to communicate brief messages (i.e. tweets) to millions of individuals within seconds. Twitter, like many social media platforms, actively attempts to diminish the spread of false information by malicious accounts within its platform. To stimulate collective action on building stronger tools for this issue, the platform has created an election integrity archive (Twitter 2018). This archive stores public datasets of Twitter accounts and tweet information believed to be associated to state-backed information operations. The data used in this paper derives from this archive and comprises accounts believed to originate from Russia active between August 3rd, 2010 and November 6th 2018. Considering the suspicion of Russian interference in the 2016 US federal elections, we decided to further subset the data to tweets generated between July 1st, 2014 and October 1st, 2017. The initial dataset contained 920,761 tweets by 361 accounts. After subsetting the data for only English tweets and selecting dates relevant to the US election, it shortened to 540,953 tweets by 320 accounts.

Figure 1: Number of tweets per Twitter account

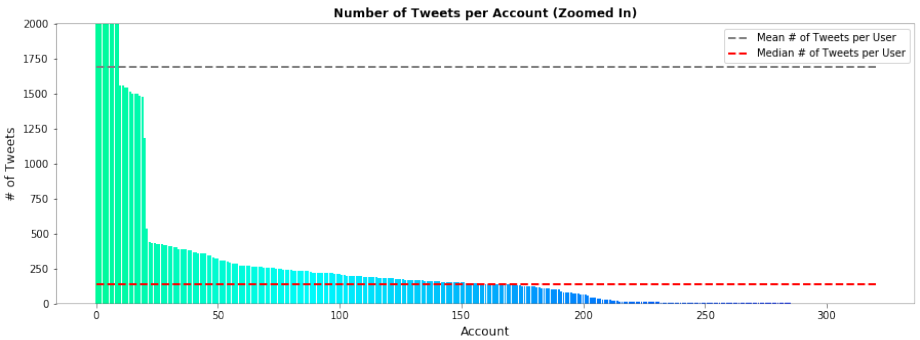
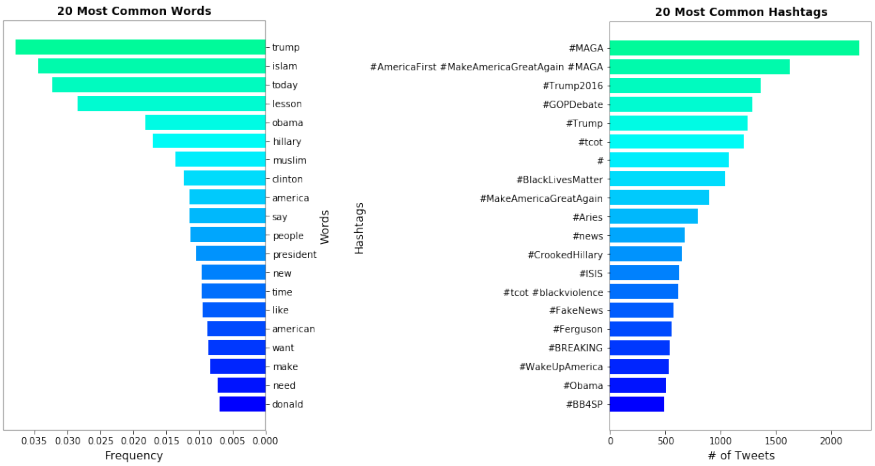


Figure 2: Top words and hashtags in the data




## 3.2 Data Cleaning and Preprocessing

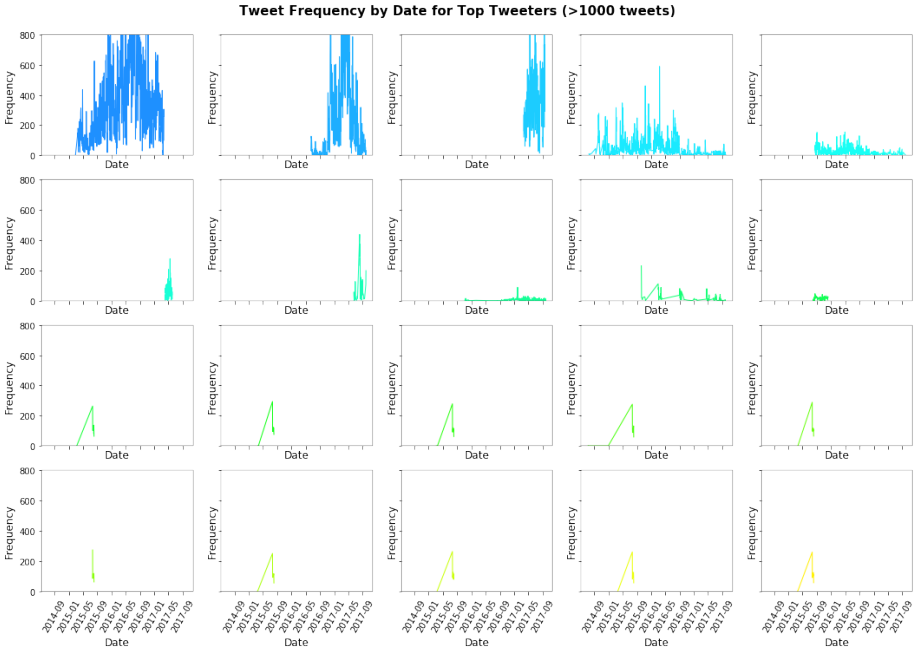
Working with text requires undertaking a thorough data cleaning process as the first and most crucial step. Twitter users have diverse writing styles. To ensure a fair representation of words or phrases expressing similar things, we changed parts of the data such as correcting misspelled words or expanding contractions. Given its strict 140 character limit per tweet, Twitter motivates people to use short-form word representation, acronyms, emojis or slang to deliver their message while staying within the character limit. Because of this, our data cleaning was essential to having high-quality input data for our model.

The preprocessing, outlined below, allowed us to explore and analyze the data in an easier way, such as in Figure 2 where we can see the top words and hashtags used in our text. This information confirms the main target of the users was the US election with a particular focus on Donald Trump, Barack Obama, and Hilary Clinton.

Data Preprocessing Steps:

1. Using the 'Tweet Preprocessor' library, we selected the tweet texts from our dataset and stored all hashtags, emojis, and mentions for later use.
2. After storing separately the elements we need, we removed all hashtags, mentions, emojis, numbers, and URLs from the text.
3. To avoid case-sensitive processes, we changed the entire tweets to lowercase representation.
4. Removal of all symbols and punctuation (e.g. ! ? , . / )
5. To ensure that words such as "you're" and "you are" are treated equally, we expanded all contractions in the text.

Figure 3: Frequency of tweets throughout collection period




6. In text analysis, individual words that form a text need to be represented numerically. We ensured that the tweets were represented as a list of individual words as opposed to a single sentence.

7. Stop words are words that are very common across all text (e.g. and, the, to, with, from). These words do not add value to text analysis and were removed from all tweets.

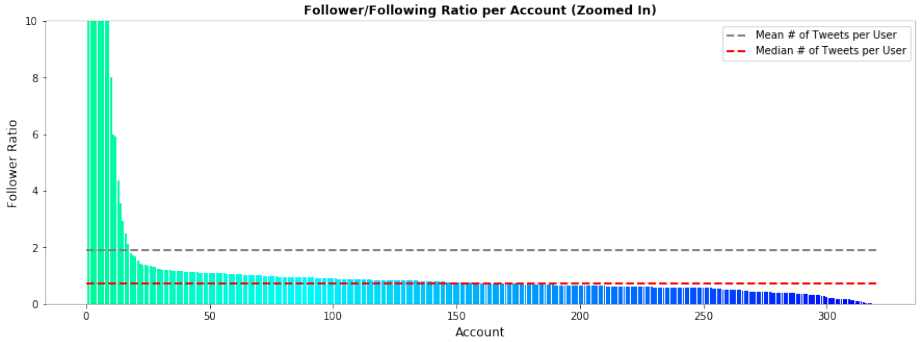
8. When words only appear in a few tweets from a 500,000+ corpus, they only contribute to making a word vector matrix more sparse, thus removing rare words helps condense the data.

9. The final step before the finish line was substituting words that have extra unnecessary letters (e.g. caaaat) to its correct spelling (e.g. cat).

## 3.3 Data Exploration

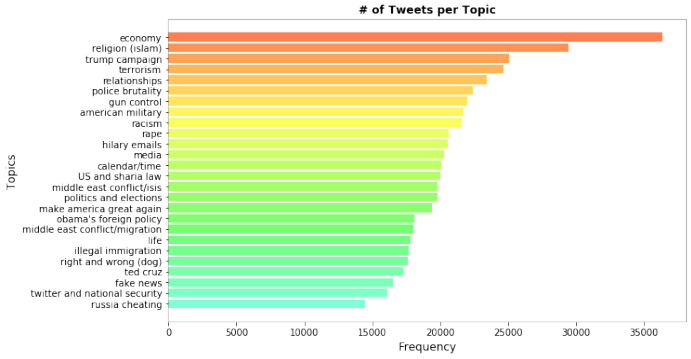
The initial exploration of our data comprised basic descriptive statistics. Figure 1 shows a scaled version of the tweet distribution per user where we can see that the data is quite skewed with the mean number of tweets being 15 times larger than the median. Additionally, 4 out of 320 users have generated over 80 percent of the data. We were curious about the consistency of top users (i.e. having a total tweet count of over 1000) and if there were trends among their defamation strategies. Figure 1 shows how active users were depending on the date from 2014 through 2017. We can see that the top five users' tweet patterns are unrelated; however, the bottom half of the top users seem to have tweeted at similar dates, with the same frequency. This suggests there was a common

Figure 4: Follower to following ratio of users




strategy and potentially an underlying connection between these accounts. Lastly, according to Gurajala et al. 2016, typical fake accounts created on Twitter, or social media, in general, usually have a higher ratio of account followers to accounts they follow. Figure 4 shows a zoomed plot of follower ratios of 320 accounts in our dataset. Although there are a number of accounts that have a larger ratio, we can see the median ratio is 1, which suggests that the accounts involved in this influence strategy were potentially real individuals and not necessarily 'bots'.

Figure 5: Number of tweets per 26 topics identified in the MALLET topic model

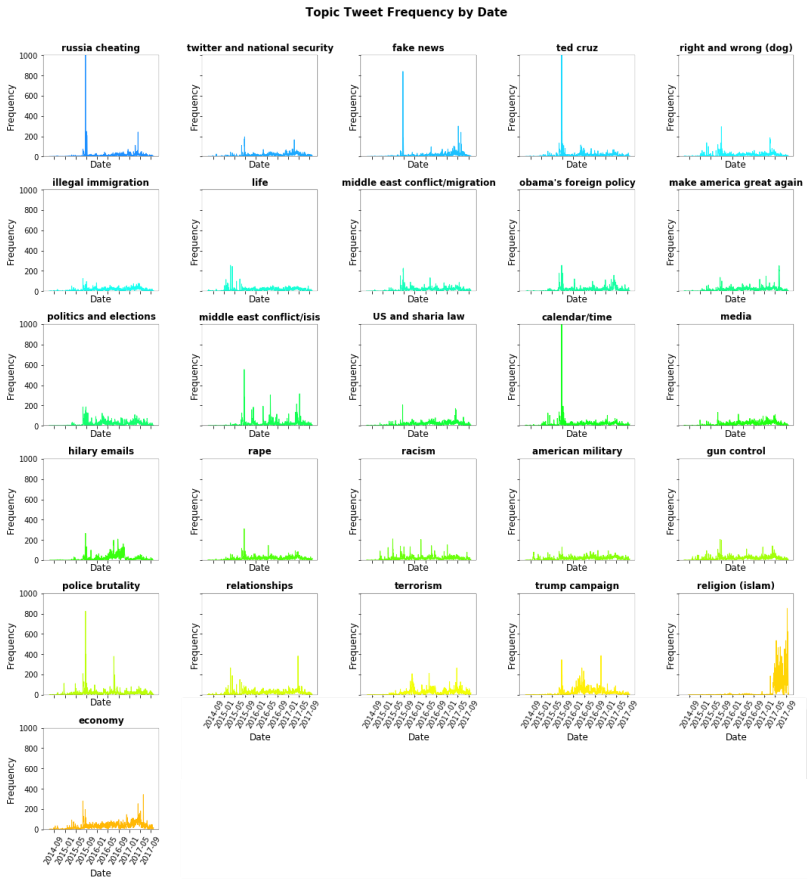



## 3.4 MALLET Topic Model

Prior to identifying an approach for defamation detection in tweets, we aimed to understand the key topics discussed in the tweets using topic modeling. Topic models are a ubiquitous tool in text analysis because of their versatility in handling massive quantities of unlabeled text. Here, a topic refers to a group of words which seem to have a higher probability of appearing together. By running our corpus through a topic model, we were able to preview potential areas of influence in the strategies of our users in relation to defamation attempts.

Although there are many variants of topic models in literature, we used the MALLET (MAchine Learning for LanguagE Toolkit) (McCallum 2002) model for our analysis. MALLET is an excellent tool for when one aims to use topic modeling as an initial exploratory tool in the data because it is a scalable implementation of Gibbs sampling which works very well for expediting clustering. We first ran the model to do an efficient search for the optimal amount of topics in our corpus. The search range spanned from 5 to 35 topics with a step size of 3. As a result, the optimal number of topics was 26 with a highest coherence value of 39.32. The primary results of the model are shown in Figure 5 where we see a labeled 26 topics along the y-axis, and the number of tweets pertaining to each topic along the x-axis. Given the context of the US elections, and American current events in general, it is reasonable that the main topics include Trump, Clinton, Barack Obama, terrorism, economics, and gun control. Overall, majority of the identified topics are sensitive to the US and are a clear choice for attempts at at influencing opinions. Furthermore, in Figure 6, for each of the topics, we can see how tweet frequency changes by date between 2014 and 2017. What stands out is the peaks of tweets around Summer 2015, which is coincidentally also the starting time of the Trump campaign.

Figure 6: Frequency of topics throughout collection period



# 4 Model and Results

We created an index to measure the extent of defamation in tweets targeted to a specific person over a period of time. Our approach considered three criteria. The first was that the content must be directed towards a specific target, which in this case would be a prominent political figure. Second, the statements must be intended to cause injury or harm to the target's reputation with the general public. Lastly, the volume of defamatory tweets must be substantial enough over the given time period to indicate a strategy.

We accomplished this by identifying a list of target political figures and separately creating a defamation dictionary incorporating negative sentiment, emotional and offensive language dictionaries, as well common phrases used in Trump's tweets towards his political rivals. The corpus and dictionary were tokenized with TF-IDF vectors and the cosine similarity between each tweet and the defamation dictionary was used to score the level of defamatory language in each tweet. The average similarity score of tweets tagged to each target was combined with a measure of the volume of tweets per target to create a score indicative of the relative level of defamatory content directed towards each target.

## 4.1 Dictionary Creation

Given that there are not any pre-existing dictionaries or labeled datasets to identify defamatory language, we looked to alternative sources given contextual knowledge about the type of strategies taken by Russian trolls during the pre- and post-2016 election time period, legal definitions of defamation and the general nature of online defamation. We first created a list of potential targets by identifying the common political figures in 2014 - 2017 by reviewing online news articles. We refined the list by only including those identified in our corpus, for a total of 21.[1]

Next, we created a dictionary of defamatory words. As discussed in Nieto 2020, displaying emotions such as contempt for or anger at someone with the intent of causing that person anger, shame or pain, could indicate a defamation strategy. To that end, we created a dictionary of negative sentiment words[2], words reflecting anger and disgust[3] and offensive language[4]. We also supplemented our dictionary with common words from Donald Trump's tweets as a proxy for defamatory language. Trump's style of tweeting has largely been found to be insulting and attacking in nature by major news sources. In an analysis by the Washington Post [Schwartzman and Johnson 2015] of more than 6000 tweets posted by Donald Trump between June and December 2015, 11% were found to be insulting. Similarly, a 2019 article in the New York Times [McIntire and Confessore 2019] stated, "Over half of the president's more than 11,000 tweets are attacks... But in more than 2,000 tweets, Mr. Trump has cited one person for praise: himself." Compared to Hillary's tweets, during the same time period, a blog article by the London School of Economics (Evans et al. 2016) found that over 17 percent of Trump's total tweets were spent criticizing government or other public figures, while

[1] Given that political figures could be referenced in multiple ways, we included both first names, last names and concatenated names as features (e.g., Tweets that included "hillary", "clinton", "hillary's" and "hillaryclinton" were tagged for Hillary Clinton).

[2] Negative sentiment words sourced from NLTK's opinion lexicon

[3] from the WordNetEffect Emotion lists: http://wndomains.fbk.eu/wnaffect.html

[4] Sourced from: https://www.cs.cmu.edu/biglou/resources/bad-words.txt. Note that the offensive language wordlist was manually refined to exclude words of a sexual or violent nature

less than 1 percent of Clinton's tweets were negative about these groups. The Trump tweets were pre-processed using the techniques described in Section 3.2. We used Trump's tweets[5] from July 2014 to October 2017 mentioning the identified targets[6]. We used TD-IDF vectorizer to create a dictionary of common terms that occurred in at least ten and at most 80%[7] of tweets. The final list was manually screened to select terms that imply negativity, contempt for, anger or disgust towards a person[8]. The final dictionary included a total of 153 words, refined down from 291 to include only those coocurring in our corpus.

## 4.2 Additional Features

Each tweet in the corpus was tagged with the target politician or politicians mentioned in the tweet. A boolean indicator was also added to reflect whether more than one politician was mentioned in the tweet.

## 4.3 Model and Measures

Our model and measures were designed to answer the following three questions: (1) How similar is the language in the tweets in our corpus to that of the defamation dictionary, (2) How many of the tweets are targeted to a political figure and (3) What is the concentration of tweets that contain defamatory language targeted towards a political figure compared to the overall corpus?

In order to determine how similar the language in each tweet was to our defamation dictionary, we first tokenized the corpus by turning each tweet into a tf-idf (term frequency-inverse document frequency) vector, which is a vector containing the weighted term frequency for each unique word in the corpus.

$$\text{TFIDF}(v, d) = f_{v,d} * log\left(\frac{D}{df_v}\right)$$

$f_{v,d}$ is the number of times term $v$ appears in document $d$, $D$ is the number of documents in the corpus and $df_v$ is the number of documents that contain $v$. The term frequency is multiplied by the term's inverse document frequency, a weighting factor representing a word's importance in the corpus. Words that occur in fewer documents will have a higher weight than those occurring in many documents. Thus, the level of importance of words occurring in a majority of documents would be scaled down. After tokenization using Gensim's TF-IDF module, each of the 540,953 tweets in the corpus was represented by a vector of tuples with an index number representing the term and it's corresponding tf-idf value. In total, there were 31,811 unique terms in the corpus. The defamation dictionary was also tokenized using the same model.

Next, we used the cosine similarity between the defamation dictionary and each tweet in the corpus to measure how similar the language in each tweet is to the defamation dictionary. The

[5]Sourced from the Trump Twitter Archive: http://www.trumptwitterarchive.com/
[6]We excluded tweets about Trump himself as those would be laudatory in nature
[7]In fact there were no terms that occurred in more than 50% of documents
[8]65 of 440 terms sourced from Trump's tweets were included in the dictionary.

cosine similarity measures the cosine angle between two non-zero vectors in an inner product space.

$$\text{cosine similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

As opposed to the Euclidian distance, the cosine similarity does not take the length of the vector into account. As a result, is a popular measure of similarity for text data, where documents are often of different lengths.

We then filtered the results for tweets aligned to a single target and calculated the average similarity score for all tweets aligned to each target. Tweets associated with multiple targets were excluded in order to ensure certainty as to the subject of the tweet. The values of the average similarity scores are low, and not meaningful on their own, as the defamatory dictionary only represents .05% of the unique words in the corpus. However, the rank of the scores against each other is meaningful to indicate the level of defamatory statements associated with each politician.

Lastly, a defamation score was created by weighting the average similarity score with the ratio of tweets in the corpus associated with each politician.

$$\text{defamation score} = \text{average similarity score} * \frac{\text{number of tweets tagged to target}}{\text{total number of tweets in corpus}} * 1000$$

A scaling factor of 1000 was added to center the score between 0 and 1 for readability purposes.

## 4.4 Results

Our initial data exploration revealed topics related to political figures, such as Donald Trump, Hillary Clinton and Barack Obama. Given this information and the time period of July 2014 to October 2017, we hypothesized that we would find an election-related defamation strategy targeting Hillary Clinton. As a result, we expected the defamation score to be high for Hillary Clinton, moderate for Barack Obama and low for Donald Trump. Table 1 presents the average similarity defamation scores by politician. In line with our expectations, Hillary Clinton received the second highest defamation score and second highest similarity score. Donald Trump, one the other hand, received one of the lowest average similarity scores (15th out of 21 and a level equivalent to half of Hillary Clinton's), but the second highest defamation score as he was tagged in 38,502 tweets (7% of the corpus), the highest number of tweets compared to the other targets. This indicates that the sentiment in the tweets pertaining to Donald Trump is likely mixed, but given the sheer number of tweets, they could potentially reflect a polarization strategy, which could be an opportunity for further expansion on this work. Obama's defamation strategy score placed third. His average similarity score placed 7th out of 21 observations but with 3.7% of tweets in the corpus targeting him, there are indications of a defamation strategy. In contrast, Mitch McConnell had the highest average similarity score but the 15th lowest defamation score as there were only 105 tweets targeting him. This result could be interpreted as a short-lived, weak defamation strategy.

As discussed in the literature review, the characterization and identification of defamation at the individual tweet level (or account level) is difficult. Our measure provides a means of identifying likely defamatory content. In order to take the analysis one step further, we added a temporal view by plotting the volume of tweets per politician over time. A defamation strategy usually has a

Table 1: Average Similarity Score and Defamation Score by Target

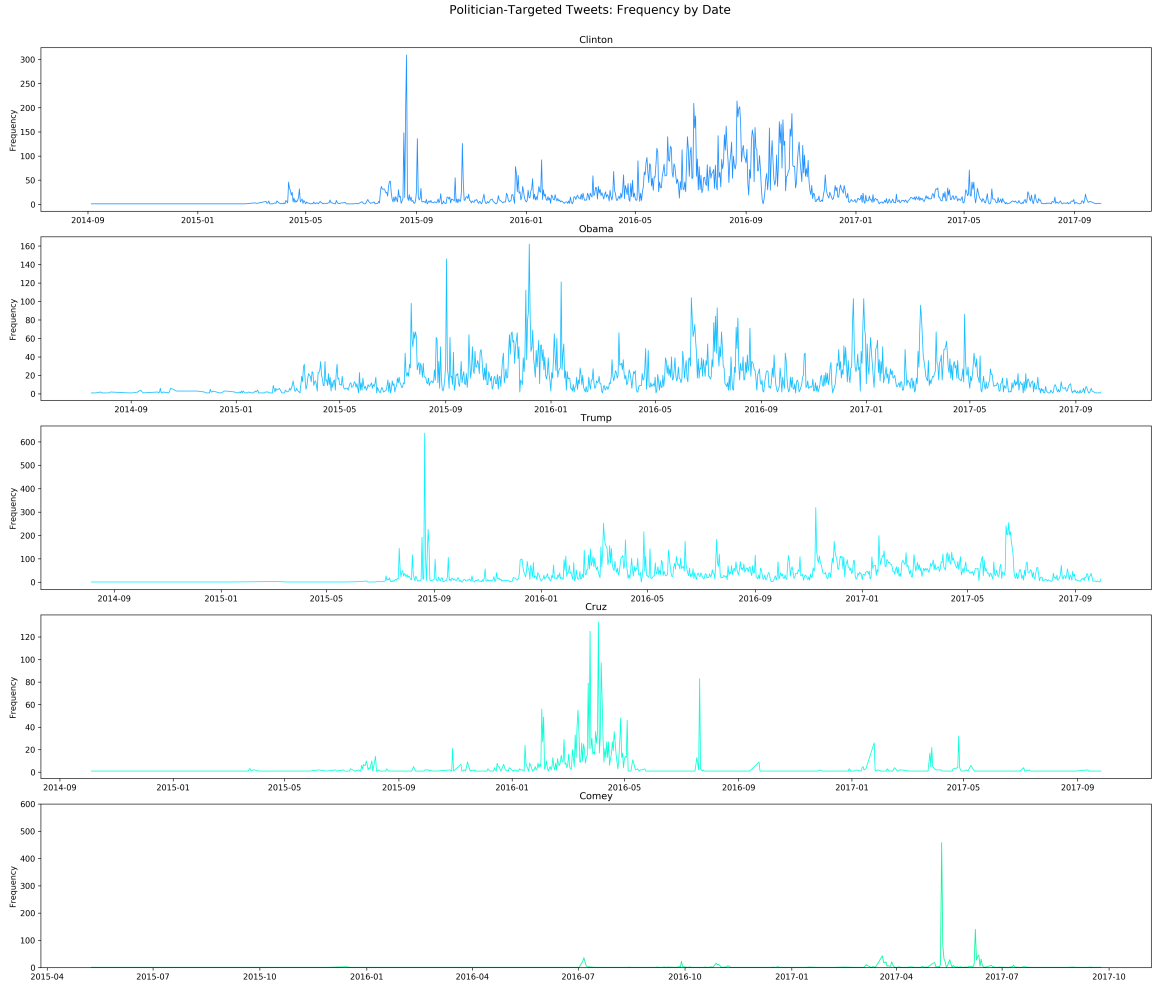
| **Target** | **Similarity Score** | **Similarity Rank** | **No. Tweets** | **Tweets Ratio** | **Defamation Score** |
|---|---|---|---|---|---|
| Hillary Clinton | .0069 | 2 | 24405 | .0451 | .3018 |
| Donald Trump | .0034 | 15 | 38502 | .0711 | .2418 |
| Barak Obama | .0049 | 7 | 19756 | .0365 | .1771 |
| Ted Cruz | .0056 | 3 | 2786 | .0051 | .0290 |
| James Comey | .0055 | 4 | 1643 | .0030 | .0167 |
| Chuck Schumer | .0070 | 0 | 732 | .0014 | .0095 |
| Bernie Sanders | .0047 | 8 | 881 | .0016 | .0076 |
| Paul Ryan | .0035 | 13 | 1040 | .0019 | .0068 |
| Jeb Bush | .0039 | 11 | 812 | .0015 | .0058 |
| Mario Rubio | .0045 | 9 | 598 | .0007 | .0050 |
| Michael Flynn | .0054 | 5 | 392 | .0007 | .0039 |
| John Kasich | .0042 | 10 | 375 | .0006 | .0029 |
| Steve Bannon | .0030 | 16 | 328 | .0003 | .0018 |
| Robert Mueller | .0049 | 6 | 189 | .0002 | .0017 |
| Mitch McConnell | .0030 | 1 | 105 | .0003 | .0014 |
| Sean Spicer | .0035 | 14 | 186 | .0003 | .0011 |
| Rodney Davis | .0036 | 12 | 153 | .0003 | .0010 |
| Mike Huckabee | .0021 | 19 | 143 | .0003 | .0006 |
| Reince Priebus | .0021 | 17 | 85 | .0002 | .0003 |
| John Brennan | .0018 | 20 | 66 | .0001 | .0002 |
| Dick Cheney | .0021 | 18 | 29 | .0001 | .0001 |

purpose. In the case of our corpus, the objective could be to injure Hillary Clinton's reputation and sway votes towards Donald Trump in advance of the 2016 presidential election or to defame a critic of Trump in order to discredit the critic's statements. Figure 7 illustrates the frequency of tweets for the politicians with the five highest defamation scores. Hillary Clinton's plot shows two peaks. The first is a sharp peak in September 2015, a few months after the start of the 2016 presidential election campaigns. The second is a slower build from the spring of 2016, peaking just before the election in October, and dropping sharply after the election. These patterns further support the claim that there was a defamation strategy targeting Hillary Clinton. The plots for Ted Cruz and James Comey are also indicative of defamation strategies. In the case of Ted Cruz, the volume of tweets peaked in the spring of 2016, and dropped sharply when he dropped out of the Repbulican primaries on May 3, 2016. James Comey is the ex-FBI director, fired by Trump in May 2017. His plot shows activity just before and after that date, further indicating that there was a light defamation campaign targeting him in order to support Trump's stance at that time.

Interestingly, the tweet patterns for Obama and Trump differ from the other three targets as the volume of tweets exhibit more stationarity throughout the time period. This pattern could be indicative of a different strategy, rather than a targeted defamation strategy and could be another opportunity to expand on this analysis.

We also explored the source of the tweets targeting politicians. There were only four out of 194 twitter user IDs with more than 1000 tweets tagging the targets. This could indicate that other users were focusing on different strategies. Of the four, one of the accounts clearly dominated with

Figure 7: Tweet Frequency Targeting Politicians with Top Defamation Scores




over 20K, 18K and 13K tweets about Donald Trump, Hillary Clinton and Barak Obama respectively. Another potential extension of this analysis would be to specifically analyze the tweets associated with each of these users to see if one specifically had a pro-Trump strategy, whereas another had a pro-Clinton strategy.

One surprise was that none of the topics identified in data exploration showed a clear defamation strategy. The expectation was that Topic 25: Hillary Emails would have the highest average similarity score to the defamation dictionary and a high percentage of tweets tagging Hillary Clinton. While it did in fact have the second highest average similarity score, only 2.5% of the tweets mentioned Hillary Clinton, precluding us from concluding that the entire topic represents a defamation strategy. Additional details on the results of the topic-level analysis can be found in Table 2.

Figure 8: Top Twitter User IDs Targeting Politicians (> 1000 Tweets)

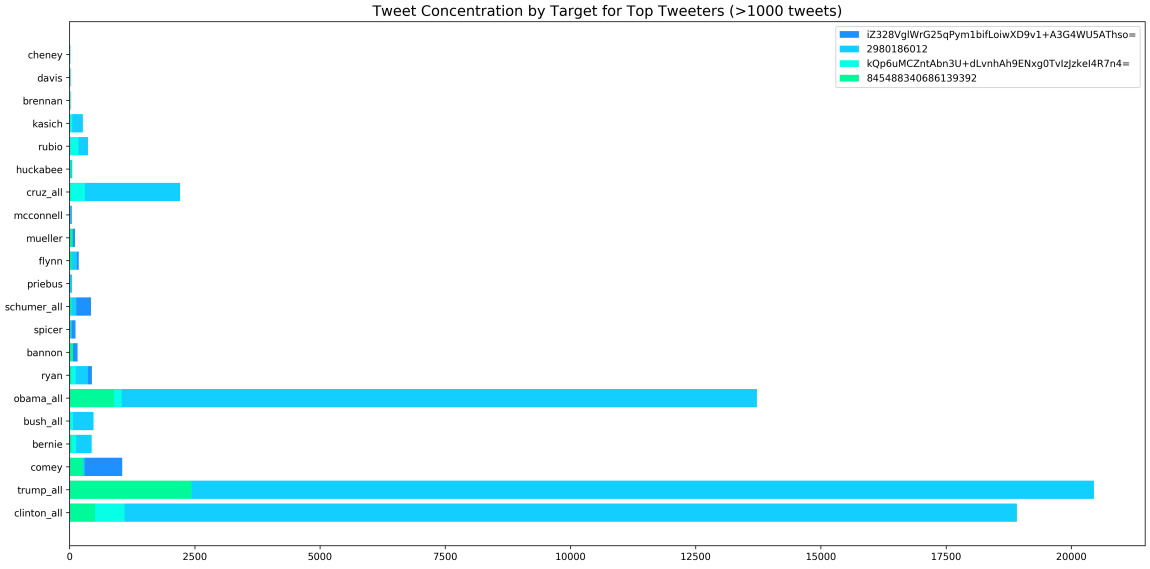



Table 2: Top Five Topics with the Highest Ave. Similarity Score

| **Dominant Topic** | **Topic Label** | **Similarity Score** | **No. Tweets** | **Politician Ratio** |
|---|---|---|---|---|
| 14 | US & Sharia Law | .0095 | 19681 | .034 |
| 25 | Hillary Emails | .0066 | 18561 | .038 |
| 24 | Fake News | .0065 | 15431 | .031 |
| 6 | Racism | .0061 | 21330 | .040 |
| 20 | Russian Cheating | .0054 | 14152 | .027 |

## 4.5 Other Exploratory Experiments

Given the open nature of the topic, we tried several exploratory experiments before landing on our final approach of an unsupervised model using cosine similarity. We initially tried to measure the similarity between the tweets in the corpus and the defamation dictionary by creating TF-IDF vectors for the corpus tweets based on the words in the dictionary. The limitation of this approach is only the words in each tweet corresponding to the words in the dictionary were embedded, thus losing the additional context around the proportion of the defamatory words in relation to the rest of the words in the corpus. Complementing the word embedding, we added features for the count of target politicians mentioned in the tweet and a boolean flag indicating whether multiple targets are mentioned. We then used a K-means clustering algorithm to classify the tweets into two categories of defamatory and non-defamatory. The objective of K-means is to assign each data point to one of $k$ clusters by minimizing the in-cluster sum of squares. The algorithm works by identifying $k$ centroids, initialized at random, and iteratively assigning each each data point to the closest centroid. After each iteration, each cluster's new centroid is found and each cluster is again assigned to the cluster

with the closest centroid. Convergence occurs when the centroid does not change from one iteration to the next. In order to score each tweet, we calculated the euclidian distance to the centroid of its assigned cluster.

Upon reviewing the results, we decided not pursue the approach further. While the concentration of defamatory terms was higher in the defamatory cluster at 25%, it was not much higher than the non-defamatory cluster at 18%. In addition, the number of tweets in the defamatory cluster was exactly the same as the number of tweets with politicians tagged, indicating that the politician count feature, rather than the language, drove the clustering. Lastly, the average distance from the centroid for the defamatory cluster is .68 and the maximum distance is 6.7, compared to .19 and .18 for the non-defamatory cluster. This points to vast differences in the level of defamatory words in each tweet of the defamatory cluster. While there could have been potential for further exploration of the defamatory cluster, the alternative approach described in Section 4.3 provided more interpretable signals for the scale of a defamation strategy.

Table 3: K-Means Clustering Results

| | **Non-Defamatory** | **Defamatory** |
|---|---|---|
| | Cluster 0 | Cluster 1 |
| No. Tweets | 432,728 | 108,225 |
| Ave. Defamatory Word Count | 18% | 25% |
| Maximum Distance from Centroid | 1.0 | 6.7 |
| Minimum Distance from Centroid | .03 | .33 |
| Ave. Distance from Centroid | .19 | .68 |
| Ave. No. Politicians Tagged | 0 | 1.3 |

# 5 Conclusions

There are several challenges in identifying defamatory campaigns by FIE on Twitter. First, confident classification of twitter accounts as FIE. Second, the ambiguity surrounding the definition and legal definition of defamation, which allows for substantially true attacks against a person under Freedom of Expression protections. The combination of these challenges makes it almost impossible to identify for certain whether an individual tweet is a bot, a troll, a legitimate twitter user with grievances, and whether that tweet is true, causing severe harm (as per the defamation definition.)

Our approach overcame these challenges by looking for evidence of defamatory attacks at a population-level. We created a measure of defamation using comparisons to a defamatory language dictionary and combining this with volume measures and filters for attacks mentioning specific individuals. The results identified Hilary Clinton as having been subject to more defamatory attacks compared to any other politician during the 2016 Presidential campaign. Given the auspicious timing of the higher volumes of tweets, we felt this provided clear evidence of defamatory attacks. In contrast, Barack Obama and Donald Trump showed more consistent defamation scores over time, and Donald Trump's defamation score was largely caused by the large volume of tweets mentioning him, which we believe demonstrated weaker evidence of defamatory attacks against these two personalities.

Our methodology allows for the comparison of strength of defamatory campaigns against different politicians within the context of U.S. Presidential politics. However by adjusting the defamatory language dictionaries, we believe this method could be more widely applied to look for evidence of defamation on twitter more broadly.

Despite the approach we took to tackle the challenges laid out in the first paragraph of the conclusion, there are further measures that could be taken to establish the robustness of our approach. Additional work could explore the comparisons of defamatory language between a corpus of identified legitimate twitter users and FIE databases in order to identify signals that are distinctive to attacks by FIE. Manual labelling of a sample of tweets could be used to test the consistency of our defamation scoring with a score from a human - which could be achieved in collaboration with a fact-checking organisation with an existing database of fact-checked tweets.

# References


Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. https://doi.org/10.3386/w23089

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. https://doi.org/10.1109/asonam.2018.8508646

Bradshaw, S., & Howard, P. N. (2018). Challenging truth and trust: A global inventory of organized social media manipulation. *The Computational Propaganda Project*, *1*.

Cagé, J., Hervé, N., & Viaud, M.-L. (2019). The production of information in an online world: Is copy right? *Available at SSRN 2672050*.

Dellavigna, S., & Kaplan, E. (2007). The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, *122*(3), 1187–1234. https://doi.org/10.1162/qjec.122.3.1187

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication Society*, *21*(5), 729–745. https://doi.org/10.1080/1369118x.2018.1428656

Evans, H., Brown, K., & Wimberly, T. (2016). Hillary clinton is tweeting more than donald trump and attacks him more often than he does her. London School of Economics. https://blogs.lse.ac.uk/usappblog/2016/07/09/hillary-clinton-is-tweeting-more-than-donald-trump-and-attacks-him-more-often-than-he-does-her/

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, *14*(2), 265–285.

Gentzkow, M., & Shapiro, J. (2006). What drives media slant? evidence from u.s. daily newspapers. https://doi.org/10.3386/w12707

Gentzkow, M., & Shapiro, J. (2010). What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, *78*(1), 35–71. https://doi.org/10.3982/ecta7195

Gurajala, S., White, J., Hudson, B., Voter, B., & Matthews, J. (2016). Profile characteristics of fake twitter accounts. *Big Data Society*, *3*. https://doi.org/10.1177/2053951716674236

Hindman, M., & Barash, V. (2018). Disinformation, and influence campaigns on twitter.

Martin, & Shapiro. (2019). Trends in online foreign ináuence efforts.

McCallum, A. K. (2002). *Mallet: A machine learning for language toolkit* [http://mallet.cs.umass.edu]. http://mallet.cs.umass.edu.

McIntire, M., & Confessore, N. (2019). Trump's twitter presidency: 9 key takeaways. *The New York Times*. Retrieved November 2, 2019, from https://www.nytimes.com/2019/11/02/us/trump-twitter-takeaways.html

Miller, D. T. (2019). Topics and emotions in russian twitter propaganda. *First Monday*. https://doi.org/10.5210/fm.v24i5.9638

Mmller, K., & Schwarz, C. (2017). Fanning the flames of hate: Social media and hate crime. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3082972

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nieto, V. G. (2020). Defamation as a language crime - a sociopragmatic approach to defamation cases in the high courts of justice of spain. *International Journal of Language and Law*.

Nouh, M., Nurse, R. J., & Goldsmith, M. (2019). Understanding the radical mind: Identifying signals to detect extremist content on twitter. *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*. https://doi.org/10.1109/isi.2019.8823548

Schwartzman, P., & Johnson, J. (2015). It's not chaos. it's trump's campaign strategy. *Washington Post*. Retrieved December 9, 2015, from https://www.washingtonpost.com/politics/its-not-chaos-its-trumps-campaign-strategy/2015/12/09/9005a5be-9d68-11e5-8728-1af6af208198%20story.html

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Twitter. (2018). Elections integrity data archive. Twitter. https://about.twitter.com/en_us/advocacy/elections-integrity.html#data

Ward, A., Ross, L., Reed, E., Turiel, E., & Brown, T. (1997). Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and knowledge*, 103–135.

Weeks, B. E., Ksiazek, T. B., & Holbert, R. L. (2016). Partisan enclaves or shared media experiences? a network approach to understanding citizens' political news environments. *Journal of Broadcasting & Electronic Media*, *60*(2), 248–268.