

Measuring Offensive Speech in Online Political Discourse

Rishab Nithyanand¹, Brian Schaffner², Phillipa Gill¹

¹{rishab, phillipa}@cs.umass.edu, ²schaffne@polsci.umass.edu
University of Massachusetts, Amherst

Abstract

The Internet and online forums such as Reddit have become an increasingly popular medium for citizens to engage in political conversations. However, the online disinhibition effect resulting from the ability to use pseudonymous identities may manifest in the form of offensive speech, consequently making political discussions more aggressive and polarizing than they already are. Such environments may result in harassment and self-censorship from its targets. In this paper, we present preliminary results from a large-scale temporal measurement aimed at quantifying offensiveness in online political discussions.

To enable our measurements, we develop and evaluate an offensive speech classifier. We then use this classifier to quantify and compare offensiveness in the political and general contexts. We perform our study using a database of over 168M Reddit comments made by over 7M pseudonyms between January 2015 and January 2017 – a period covering several divisive political events including the 2016 US presidential elections.

1 Introduction

The apparent rise in political incivility has attracted substantial attention from scholars in recent years. These studies have largely focused on the extent to which politicians and elected officials are increasingly employing rhetoric that appears to violate norms of civility [4, 12]. For the purposes of our work, we use the incidence of offensive rhetoric as a stand in for incivility. The 2016 US presidential election was an especially noteworthy case in this regard, particularly in terms of Donald Trump’s campaign which frequently violated norms of civility both in how he spoke about broad groups in the public (such as Muslims, Mexicans, and African Americans) and the attacks he leveled at his opponents [2]. The consequences of incivility are thought to be crucial to the

functioning of democracy since “public civility and interpersonal politeness sustain social harmony and allow people who disagree with one another to maintain ongoing relationships” [17].

While political incivility appears to be on the rise among elites, it is less clear whether this is true among the mass public as well. Is political discourse particularly lacking in civility compared to discourse more generally? Does the incivility of mass political discourse respond to the dynamics of political campaigns? Addressing these questions has been difficult for political scientists because traditional tools for studying mass behavior, such as public opinion surveys, are ill-equipped to measure how citizens discuss politics with one another. Survey data does reveal that the public tends to perceive politics as becoming increasingly less civil during the course of a political campaign [18]. Yet, it is not clear whether these perceptions match the reality, particularly in terms of the types of discussions that citizens have with each other.

An additional question is how incivility is received by others. On one hand, violations of norms regarding offensive discourse may be policed by members of a community, rendering such speech ineffectual. On the other hand, offensive speech may be effective as a means for drawing attention to a particular argument. Indeed, there is evidence that increasing incivility in political speech results in higher levels of attention from the public [12]. During the 2016 campaign, the use of swearing in comments posted on Donald Trump’s YouTube channel tended to result in additional responses that mimicked such swearing [8]. Thus, offensive speech in online fora may attract more attention from the community and lead to the spread of even more offensive speech in subsequent posts.

To address these questions regarding political incivility, we examine the use of offensive speech in political discussions housed on Reddit. Scholars tend to define uncivil discourse as “communication that violates the norms of politeness” [12] a definition that clearly in-

cludes offensive remarks. Reddit fora represent a “most likely” case for the study of offensive political speech due its strong free speech culture [14] and the ability of participants to use pseudonymous identities. That is, if political incivility in the public did increase during the 2016 campaign, this should be especially evident on fora such as Reddit. Tracking Reddit discussions throughout all of 2015 and 2016, we find that online political discussions became increasingly more offensive as the general election campaign intensified. By comparison, discussions on non-political subreddits did not become increasingly offensive during this period. Additionally, we find that the presence of offensive comments did not subside even three months after the election.

2 Datasets

Our study makes use of multiple datasets in order to identify and characterize trends in offensive speech.

The Crowdfunder hate speech dataset. The Crowdfunder hate speech dataset [1] contains 14.5K tweets, each receiving labels from at least three contributors. Contributors were allowed to classify each tweet into one of three classes: Not Offensive (NO), Offensive but not hateful (O), and Offensive and hateful (OH). Of the 14.5K tweets, only 37.6% had a decisive class – *i.e.*, the same class was assigned by all contributors. For indecisive cases, the majority class was selected and a class confidence score (fraction of contributors that selected the majority class) was made available. Using this approach, 50.4%, 33.1%, and 16.5% of the tweets were categorized as NO, O, and OH, respectively. Since our goal is to identify any offensive speech (not just hate speech), we consolidate assigned classes into Offensive and Not Offensive by relabeling OH tweets as Offensive. We use this modified dataset to train, validate, and test our offensive speech classifier. To the best of our knowledge, this is the only dataset that provides *offensive* and *not offensive* annotations to a large dataset.

Offensive word lists. We also use two offensive word lists as auxiliary input to our classifier: (1) The Hatebase hate speech vocabulary [3] consisting of 1122 hateful words and (2) 422 offensive words banned from Google’s What Do You Love project [7].

Reddit comments dataset. Finally, after building our offensive speech classifier using the above datasets, we use it to classify comments made on Reddit. While the complete Reddit dataset contains 2B comments made between the period of January 2015 and January 2017, we only analyze only 168M. We select comments to be analyzed using the following process: (1) we exclude comments shorter than 10 characters in length, (2) we exclude comments made by [deleted] authors, and (3)

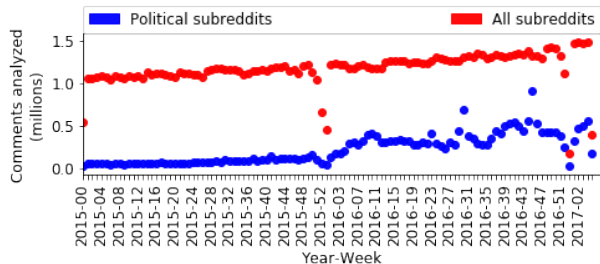


Figure 1: Number of analyzed political and apolitical comments belonging to each week between January 2015 and January 2017.

we randomly sample and include 10% of all remaining comments. We categorize comments made in any of 21 popular political subreddits as *political* and the remainder as *apolitical*. Our final dataset contains 129M apolitical and 39M political comments. Figure 1 shows the number of comments in our dataset that were made during each week included in our study. We see an increasing number of political comments per week starting in February 2016 – the start of the 2016 US presidential primaries.

3 Offensive Speech Classification

In order to identify offensive speech, we propose a fully automated technique that classifies comments into two classes: Offensive and Not Offensive.

3.1 Classification approach

At a high-level, our approach works as follows:

- **Build a word embedding.** We construct a 100-dimensional word embedding using all comments from our complete Reddit dataset (2B comments).
- **Construct a hate vector.** We construct a list of offensive and hateful words identified from external data and map them into a single vector within the high-dimensional word embedding.
- **Text transformation and classification.** Finally, we transform text to be classified into scalars representing their distance from the constructed hate vector and use these as input to a Random Forest classifier.

Building a word embedding. At a high-level, a word embedding maps words into a high-dimensional continuous vector space in such a way that semantic similarities between words are maintained. This mapping is achieved

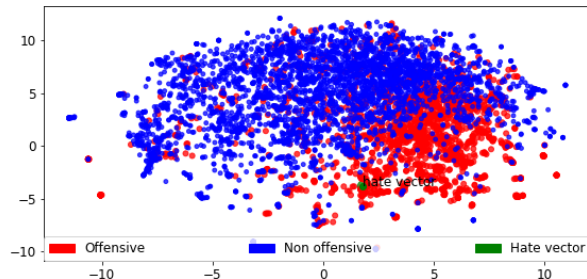


Figure 2: Proximity of offensive and non-offensive comments to the hate vector. Dimension reduction is performed using t-SNE.

by exploiting the distributional properties of words and their occurrences in the input corpus.

Rather than using an off-the-shelf word embedding (e.g., the GloVe embeddings [13] trained using public domain data sources such as Wikipedia and news articles), we construct a 100-dimensional embedding using the complete Reddit dataset (2B comments) as the input corpus. The constructed embedding consists of over 400M unique words (words occurring less than 25 times in the entire corpus are excluded) using the Word2Vec [10] implementation provided by the Gensim library [15]. Prior to constructing the embedding, we perform stop-word removal and lemmatize each word in the input corpus using the SpaCy NLP framework [5]. The main reason for building a custom embedding is to ensure that our embeddings capture semantics specific to the data being measured (Reddit) – e.g., while the word “*karma*” in the non-Reddit context may be associated with spirituality, it is associated with points (comment and submission scores) on Reddit.

Constructing a hate vector. We use two lists of words associated with hate [3] and offensive [7] speech to construct a hate vector in our word embedding. This is done by mapping each word in the list into the 100-dimensional embedding and computing the mean vector. This vector represents the average of all known offensive words. The main idea behind creating a hate vector is to capture the point (in our embedding) to which the most offensive observed comments are likely to be near. Although clustering our offensive word lists into similar groups and constructing multiple hate vectors – one for each cluster – results in marginally better accuracy for our classifier, we use this approach due to the fact that our classification cost grows linearly with the number of hate vectors – i.e., we need to perform $O(|S|)$ distance computations per hate vector to classify string S .

Transforming and classifying text. We first remove stop-words and perform lemmatization of each word in

the text to be classified. We then obtain the vector representing each word in the text and compute its similarity to the constructed hate vector using the cosine similarity metric. A 0-vector is used to represent words in the text that are not present in the embedding. Finally, the maximum cosine similarity score is used to represent the comment. Equation 1 shows the transformation function on a string $S = \{s_1, \dots, s_n\}$ where s_i is the vector representing the i^{th} lemmatized non-stop-word, \cos is the cosine-similarity function, and H is the hate vector.

$$T(S) = \max_{1 \leq i \leq n} [\cos(s_i, H)] \quad (1)$$

In words, the numerical value assigned to a text is the cosine similarity between the hate vector and the vector representing the word (in the text) closest to the hate vector. This approach allows us to transform a string of text into a single numerical value that captures its semantic similarity to the most offensive comment. We use these scalars as input to a random forest classifier to perform classification into Offensive and Not Offensive classes. Figure 2 shows the proximity of Offensive and Non Offensive comments to our constructed hate vector after using t-distributed Stochastic Neighbor Embedding (t-SNE) [9] to reduce our 100-dimension vector space into 2 dimensions.

3.2 Classifier evaluation

We now present results to (1) validate our choice of classifier and (2) demonstrate the impact of training/validation sample count on our classifiers performance.

Classifier	Accuracy (%)	F1-Score (%)
Stochastic Gradient Descent	80.7	80.0
Naive Bayes	81.5	81.2
Decision Tree	91.8	91.4
Random Forest	92.0	91.9

Table 1: Average classifier performance during 10-fold cross-validation on the training/validation set. Results shown are for the best performing parameters obtained using a grid search.

Classifier selection methodology. To identify the most suitable classifier for classifying the scalars associated with each text, we perform evaluations using the stochastic gradient descent, naive bayes, decision tree, and random forest classifiers. For each classifier, we split the CrowdFlower hate speech dataset into a training/validation set (75%), and a holdout set (25%). We perform 10-fold cross-validation on the training/validation set to identify the best classifier model and

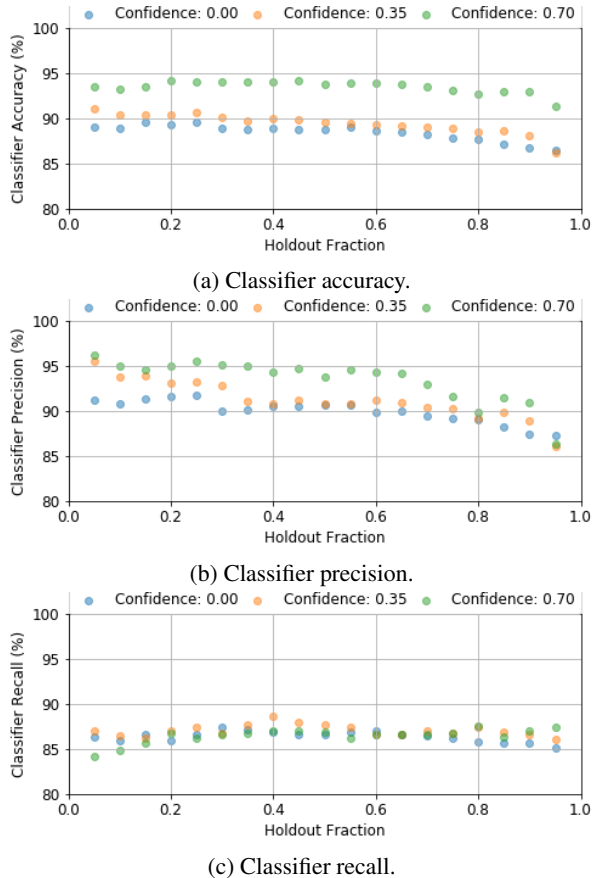


Figure 3: Classifier performance on holdout sets while varying holdout set sizes and minimum confidence thresholds.

parameters (using a grid search). Based on the results of this evaluation, we select a 100-estimator entropy-based splitting random forest model as our classifier. Table 1 shows the mean accuracy and F1-score for each evaluated classifier during the 10-fold cross-validation.

Real-world classifier performance. To evaluate real-world performance of our selected classifier (*i.e.*, performance in the absence of model and parameter bias), we perform classification of the holdout set. On this set, our classifier had an accuracy and F1-score of 89.6% and 89.2%, respectively. These results show that in addition to superior accuracy during training and validation, our chosen classifier is also robust against over-fitting.

Impact of dataset quality and size. To understand how the performance of our chosen classifier model and parameters are impacted by: (1) the quality and consistency of manually assigned classes in the CrowdFlower dataset and (2) the size of the dataset, we re-evaluate the classifier while only considering tweets having a minimum confidence score and varying the size of the holdout set. Specifically, our experiments considered confidence

thresholds of 0 (all tweets considered), .35 (only tweets where at least 35% of contributors agreed on a class were considered), and .70 (only tweets where at least 70% of the contributors agreed on a class were considered) and varied the holdout set sizes between 5% and 95% of all tweets meeting the confidence threshold set for the experiment.

The results illustrated in Figure 3 show the performance of the classifier while evaluating the corresponding holdout set. We make several conclusions from these results:

- Beyond a (fairly low) threshold, the size of the training and validation set has little impact on classifier performance. We see that the accuracy, precision, and recall have, at best, marginal improvements with holdout set sizes smaller than 60%. This implies that the CrowdFlower dataset is sufficient for building an offensive speech classifier.
- Quality of manual labeling has a significant impact on the accuracy and precision of the classifier. Using only tweets which had at least 70% of contributors agreeing on a class resulted in between 5-7% higher accuracy and up to 5% higher precision.
- Our classifier achieves precision of over 95% and recall of over 85% when considering only high confidence samples. This implies that the classifier is more likely to underestimate the presence of offensive speech – *i.e.*, our results likely provide a lower-bound on the quantity of observed offensive speech.

4 Measurements

In this section we quantify and characterize offensiveness in the political and general contexts using our offensive speech classifier and the Reddit comments dataset which considers a random sample of comments made between January 2015 and January 2017.

Offensiveness over time. We find that on average 8.4% of all political comments are offensive compared to 7.8% of all apolitical comments. Figure 4 illustrates the fraction of offensive political and apolitical comments made during each week in our study. We see that while the fraction of apolitical offensive comments has stayed steady, there has been an increase in the fraction of offensive political comments starting in July 2016. Notably, this increase is observed after the conclusion of the US presidential primaries and during the period of the Democratic and Republican National Conventions and does not reduce even after the conclusion of the US presidential elections held on November 8. Participants in political subreddits were 2.6% more likely to observe offen-

sive comments prior to July 2016 but 14.9% more likely to observe offensive comments from July 2016 onwards.

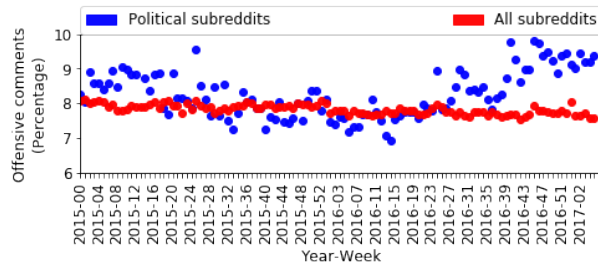


Figure 4: Fraction of offensive comments identified in political and all subreddits.

Reactions to offensive comments. We use the comment *score*, roughly the difference between up-votes and down-votes received, as a proxy for understanding how users reacted to offensive comments. We find that comments that were offensive: (1) on average, had a higher score than non-offensive comments (average scores: 8.9 vs. 6.7) and (2) were better received when they were posted in the general context than in the political context (average scores: 8.6 vs. 9.0). To understand how people's reactions to offensive comments evolved over time, Figure 5 shows the average scores received by offensive comments over time. Again, we observe an increasing trend in average scores received by offensive and political comments after July 2016.

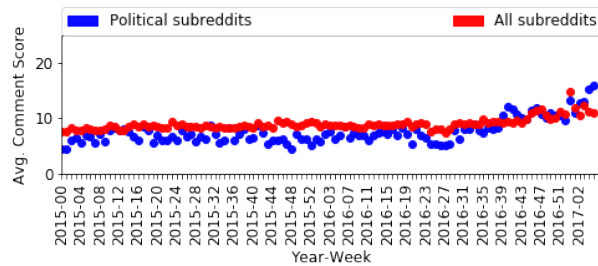


Figure 5: Average scores of offensive comments identified in political and all subreddits.

Characteristics of offensive authors. We now focus on understanding characteristics of authors of offensive comments. Specifically, we are interested in identifying the use of *throwaway* and *troll* accounts. For the purposes of this study, we characterize *throwaway* accounts as those with less than five total comments – *i.e.*, accounts that are used to make a small number of comments. Similarly, we define *troll* accounts as those with over 15 comments of which over 75% are classified as offensive – *i.e.*, accounts that are used to make a larger number of comments, of which a significant majority are

offensive. We find that 93.7% of the accounts which have over 75% of their comments tagged as offensive are throwaways and 1.3% are trolls. Complete results are illustrated in Figure 6.

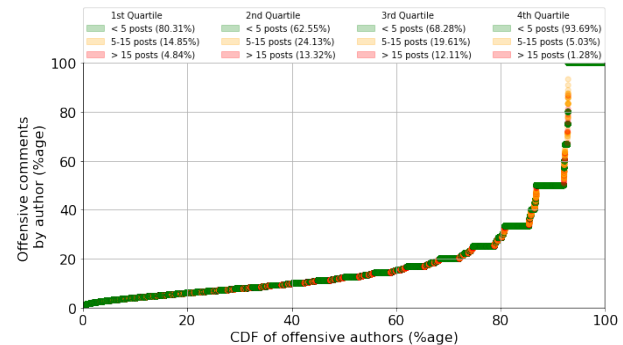


Figure 6: CDF of the fraction of each author's comments that were identified as offensive. Green, orange, and red dots are used to represent authors with <5, 5-15, and >15 total comments, respectively. The legend provides a breakdown per quartile.

Characteristics of offensive communities. We breakdown subreddits by their category (default, political, and other) and identify the most and least offensive communities in each. Figure 7 shows the distribution of the fraction of offensive comments in each category and Table 2 shows the most and least offensive subreddits in the political and default categories (we exclude the “other” category due to the inappropriateness of their names). We find that less than 19% of all subreddits (that account for over 23% of all comments) have over 10% offensive comments. Further, several default and political subreddits fall in this category, including *r/the_donald* – the most offensive political subreddit and the subreddit dedicated to the US President.

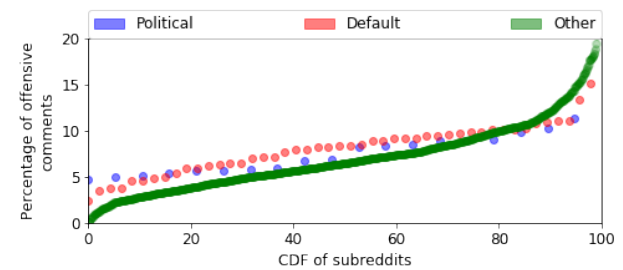


Figure 7: Distribution of the fraction of offensive comments observed in each subreddit category. Only subreddits with over 1000 comments are considered.

Flow of offensive authors. Finally, we uncover patterns in the movement of offensive authors between communities. In Figure 8 we show the communities

Category	Most offensive (%)	Least offensive (%)
Default	r/tifu (15.1%)	r/askscience (2.4%)
	r/announcements (13.2%)	r/personalfinance (3.4%)
	r/askreddit (11.0%)	r/science (3.8%)
Political	r/the_donald (11.4%)	r/republican (4.4%)
	r/elections (10.2%)	r/sandersforpresident (4.9%)
	r/worldpolitics (9.8%)	r/tedcruz (5.1%)

Table 2: Subreddits in the default and political categories with the highest and lowest fraction of offensive comments.

in which large number of authors of offensive content on the r/politics subreddit had previously made offensive comments (we refer to these communities as sources). Unsurprisingly, the most popular sources belonged to the default subreddits (e.g., r/worldnews, r/wtf, r/videos, r/askreddit, and r/news). We find that several other political subreddits also serve as large sources of offensive authors. In fact, the subreddits dedicated to the three most popular US presidential candidates – r/the_donald, r/sandersforpresident, and r/hillaryclinton rank in the top three. Finally, outside of the default and political subreddits, we find that r/nfl, r/conspiracy, r/dota2, r/reactiongifs, r/blackpeopletwitter, and r/imgoingtohellforthis were the largest sources of offensive political authors.

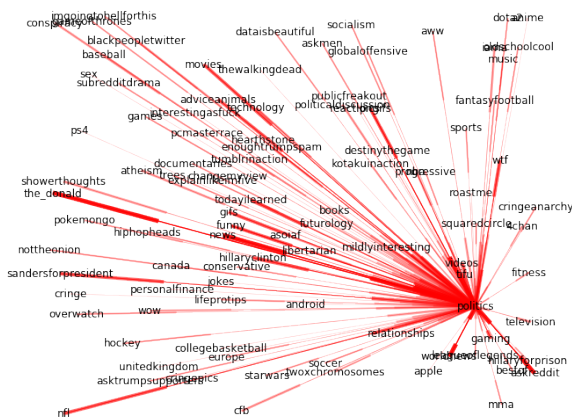


Figure 8: Flow of offensive authors. An edge between two subreddits indicates that authors made offensive comments in the source subreddit before the first time they made offensive comments in the destination subreddit. Darker and thicker edges indicate larger flow sizes (only flows ≥ 200 authors are shown).

5 Conclusions and Future Work

We develop and validate an offensive speech classifier to quantify the presence of offensive online comments from January 2015 through January 2017. We find that political discussions on Reddit became increasingly less civil – as measured by the incidence of offensive comments – during the 2016 general election campaign. In fact, during the height of the campaign, nearly one of every 10 comments posted on a political subreddit were classified as offensive. Offensive comments also received more positive feedback from the community, even though most of the accounts responsible for such comments appear to be throwaway accounts. While offensive posts were increasingly common on political subreddits as the campaign wore on, there was no such increase in non-political fora. This contrast provides additional evidence that the increasing use of offensive speech was directly related to the ramping up of the general election campaign for president.

Even though our study relies on just a single source of online political discussions – Reddit, we believe that our findings generally present an upper-bound on the incidence of offensiveness in online political discussions for the following reasons: First, Reddit allows the use of pseudonymous identities that enables the online disinhibition effect (unlike social-media platforms such as Facebook). Second, Reddit enables users to engage in complex discussions that are unrestricted in length (unlike Twitter). Finally, Reddit is known for enabling a general culture of free speech and delegating content regulation to moderators of individual subreddits. This provides users holding fringe views a variety of subreddits in which their content is welcome.

Our findings provide a unique and important mapping of the increasing incivility of online political discourse during the 2016 campaign. Such an investigation is important because scholars have outlined many consequences for incivility in political discourse. Incivility tends to “turn off” political moderates, leading to increasing polarization among those who are actively engaged in politics [18]. More importantly, a lack of civility in political discussions generally reduces the degree to which people view opponents as holding legitimate viewpoints. This dynamic makes it difficult for people to find common ground with those who disagree with them [11] and it may ultimately lead citizens to view electoral victories by opponents as lacking legitimacy [12]. Thus, from a normative standpoint, the fact that the 2016 campaign sparked a marked increase in the offensiveness of political comments posted to Reddit is of concern in its own right; that the incidence of offensive political comments has remained high even three months after the election is all the more troubling.

In future work, we will extend our analysis of Reddit back to 2007 with the aim of formulating a more complete understanding of the dynamics of political incivility. For example, we seek to understand whether the high incidence of offensive speech we find in 2016 is unique to this particular campaign or if previous presidential campaigns witnessed similar spikes in incivility. We will also examine whether there is a more general long-term trend toward offensive online political speech, which would be consistent with what scholars have found when studying political elites [6, 16].

References

- [1] CrowdFlower Blog. Hate speech identification. URL: <https://www.crowdfLOWER.com/data/hate-speech-identification/> (Accessed May 24, 2017).
- [2] Justin H Gross and Kaylee T Johnson. Twitter taunts and tirades: Negative campaigning in the age of trump. *PS: Political Science & Politics*, 49(4):748–754, 2016.
- [3] Hatebase. Meet the Hatebase API. URL: https://www.hatebase.org/connect_api (Accessed May 24, 2017).
- [4] Susan Herbst. *Rude democracy: Civility and incivility in American politics*. Temple University Press, 2010.
- [5] Matthew Honnibal and Mark Johnson. An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, 2015. Association for Computational Linguistics.
- [6] Kathleen Hall Jamieson and Erika Falk. Continuity and change in civility in the house. In *Polarized politics: Congress and the president in a partisan era*, pages 96–108. Washington, DC: CQ Press, 2000.
- [7] Jamiew. All the dirty words from Google’s “what do you love” project. URL: <https://gist.github.com/jamiew/1112488> (Accessed May 24, 2017).
- [8] K Hazel Kwon and Anatoliy Gruzd. Is aggression contagious online? a case of swearing on donald trump’s campaign videos on youtube. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.
- [9] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.
- [10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [11] Diana C Mutz. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press, 2006.
- [12] Diana C Mutz. *In-your-face politics: The consequences of uncivil media*. Princeton University Press, 2015.
- [13] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [14] Reddit CEO Speaks Out On Violentacrez In Leaked Memo: ‘We Stand for Free Speech’. Hate speech identification. URL: <https://web.archive.org/web/20170710074932/http://gawker.com/5952349/reddit-ceo-speak-s-out-on-violentacrez-in-leaked-memo-we-stand-for-free-speech> (Accessed May 24, 2017).
- [15] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 2010. ELRA. <http://is.muni.cz/publication/884893/en>.
- [16] Daniel M Shea and Alex Sproveri. The rise and fall of nasty politics in america. *PS: Political Science & Politics*, 45(03):416–421, 2012.
- [17] J Cherie Strachan and Michael R Wolf. Political civility. *PS: Political Science & Politics*, 45(03):401–404, 2012.
- [18] Michael R Wolf, J Cherie Strachan, and Daniel M Shea. Incivility and standing firm: A second layer of partisan division. *PS: Political Science & Politics*, 45(03):428–434, 2012.