Incivility Begets Incivility? Understanding the Contagion Dynamics of Uncivil Conversations on Social Media

Joshua Timm[*]
University of Southern California

Pablo Barberá[†]
University of Southern California

Abstract

Most people express frustration over the negativity of political conversations on social media platforms, which limits their potential as a space for political deliberation. We argue incivility is so prevalent on social media due to a cycle of incivility: a minority of "repeat offenders" are responsible for most of the uncivil comments, encouraging others to respond in similarly uncivil ways. We demonstrate the existence of this mechanism by analyzing Facebook comments on the pages of U.S. legislators using supervised learning and dictionaries enriched through word embeddings. We contribute to the literature on incivility by advocating for narrower classification tasks. We find that (1) over 40% of comments are uncivil, with 1% of highly active users disproportionally responsible for the uncivil content, (2) incivility receives higher user engagement, which increases its visibility, and (3) uncivil comments receive more uncivil replies, which supports the hypothesis that incivility on social media is contagious.

[*]Joshua Timm (corresponding author) is a PhD candidate in Political Science and International Relations at the University of Southern California. He can be reached at jtimm@usc.edu

[†]Pablo Barberá (www.pablobarbera.com) is an Assistant Professor of International Relations at the University of Southern California.

Motivation

Social media websites such as Twitter and Facebook – once heralded as new spaces that would revolutionize democratic politics and create opportunities for true public deliberation – have now become fertile ground for negativity, bitterness, and harassment. According to data from the Pew Research Center, nearly 40% of Americans have personally experienced online harassment and over 50% of social media users think that discussions on these platforms are angrier, less respectful, and less civil. As a result, in contrast to the initial optimism about the democratic opportunities of social media sites, many are now concerned that uncivil online interactions may be increasing affective polarization, exacerbating inequalities in political attitudes and civic engagement, and reducing the quality of political representation (see e.g. Suhay et al., 2018; Theocharis et al., 2016).

The goal of this paper is two-fold. First, we are interested in quantifying the prevalence of incivility in social media communication. While a large proportion of social media users report being exposed to uncivil messages, previous work trying to measure it directly has yielded quite different results. This variation is largely due to varying classification techniques and disagreements over how to define incivility, as well as how it is connected to other types of hateful langugage, such as hate speech, negativity, impoliteness, vitriol, etc. We devote a lot of our effort in this paper towards improving existing definitions of incivility. Second, we are also interested in offering potential solutions to this problem. We adopt a practical perspective and try to derive potential interventions that could reduce the degree of incivility on social media and improve the health of political exchanges that take place through these platforms.

With these goals in mind, we conduct an analysis of the prevalence of uncivil comments on the Facebook pages of U.S. Members of Congress. Legislators rely on these pages to connect with their constituents by sharing information about their policy positions, justifying their votes in Congress, and advertising their in-person events. While they generally use these platforms as another space to broadcast their messages, in practice many citizens flock to these pages as a place where they want their voices heard, and where they expect to engage in political discussions. For this reason, understanding whether incivility can be a deterrent for healthier exchanges can have an important impact in the political process.

In our analysis, we advocate for a broader and more nuanced conceptualization of incivility in a political context, which we divide into six (not mutually exclusive) categories: contempt, political threats, partisan vitriol, profanity, speech devaluation, and seditious language. We then explore two different methods to automatically classify comments: a supervised machine learning approach and a dictionary method that we enhance using word embeddings.

Our results confirm the conventional wisdom regarding the high prevalence of incivility in politicians' pages: we estimate that around 40% of comments in these pages can be considered uncivil, with pages by Republicans, Senators, and those with moderate ideological positions attracting a higher proportion of uncivil comments. However, contrary to survey evidence suggesting that people feel worn out after being exposed to this type of content, we find that uncivil comments receive a higher number of likes and responses, which likely increases its visibility due to Facebook's comment ranking algorithm. Finally, we also explore whether contagion dynamics could explain the high prevalence of incivility. We do so by providing evidence that uncivil comments are more likely to elicit additional uncivil responses, creating a cycle of incivility that may make other citizens who might be interested in participating in discussions less likely to engage. These results suggest that, for example, strategies that could downrank or hide uncivil comments could have an outsized effect by breaking this pernicious loop.

Defining incivility

Many scholars have contributed work that attempts to identify incivility in online settings. Conceptualizing and classifying such language is generally a difficult task, in large part due to the subjective nature of defining and operationalizing closely related yet distinct concepts, such as: uncivil speech, impolite speech, cyber-bullying, abusive speech, harmful speech, undesirable speech, harassment, etc.[1] Due to the lack of definitional and operational clarity in the literature, we attempt to cast some light on the debate by including a discussion of existing work that attempts to classify incivility online (as well as related concepts in the nexus of negative speech). In this review of existing work, we also include a discussion of the problems that accompany the task of

[1]For sake of clarity, throughout the rest of this paper we refer to the undefined aggregate of all these terms as "negative speech," with incivility being a subset of this broader category.

operationalizing incivility on social media.

What is incivility, anyway?

Among all the types of negative speech, incivility in particular faces measurement challenges because of the broadness of its definition. The first conceptual issue in this literature is the tension between the definitions of impoliteness and incivility as distinct or identical categories. Some have conceptualized it as general impoliteness or rudeness (Jamieson, 1997, 1998), while others argue that general rudeness is not incivility. Papacharissi (2004, p.267) argues for a much stricter definition of incivility, writing "Incivility can be defined as negative collective face; that is, disrespect for the collective traditions of democracy. Civility can then be operationalized as the set of behaviors that threaten democracy, deny people their personal freedoms, and stereotype social groups." In fact, Papacharissi adopts the definitions of incivility by Jamieson (1997) and Jamieson (1998) as her own definition of impoliteness, a concept she argues is distinct from incivility. While Papacharissi separates impoliteness from incivility, some scholars (Theocharis et al., 2016) conceptualize incivility closer to this 'impoliteness' definition. Still, other scholars (Rossini, 2019) conceptualize incivility broadly, including both 'impoliteness' and a more severe definition of incivility, such as intolerance.

This tension is related with a second conceptual dimension in the literature, which is related to a definition of incivility in relation to its consequences. Here we can find three broad groups, depending on whether the consequences are severe, common, or both.

Our conceptualization begins with severe incivility, which originated with Papacharissi (2004) but is expanded by Borah (2014), Blom et al. (2014), and Oz et al. (2017). Severe incivility includes disrespect for traditions of democracy, offensive stereotyping of vulnerable groups, threatening other individuals' rights, advocating/inciting violence, using racial/ethnic slurs, and hate speech. These are severe forms of incivility because their capacity to harm the public good is greater than that of common incivility.

In contrast, common incivility is a broader category that includes insulting language, name-calling, mockery, rude critiques, misrepresentative exaggeration, unnecessarily disrespectful tone

in conversation, hostility, aggression, intimidation, offensive language, unfriendly tone, and character criticism (Rosner et al., 2016; Kevin et al., 2014; Sobieraj and Berry, 2011; Jomini et al., 2015). In contrast with severe incivility, this more common form is sometimes considered acceptable in online contexts, particularly when it targets public officials, and may even contribute to increase political engagement (Brooks and Geer, 2007a; Rossini, 2019)

Finally, we view some past work as using a mixed definition of incivility, including anything from general impoliteness and rudeness to hate speech, slurs, and racism as incivility (Borah, 2014; Blom et al., 2014; Oz et al., 2017).

Given this lack of uniformity in the literature on what exactly incivility is, coming up with a universal definition of incivility may be a futile task. Thus, we argue a better goal in classifying incivility is not to use a single definition, but to build many conceptually narrow and precise definitions that can be analyzed separately or can be aggregated depending on the goal of the analysis. Using such a method will be more robust to challenges of conceptualization and more usable for automated classifying methods, as we discuss in the next section.

In addition, the more precise concepts and classification schemes are, the less subjectivity will be involved at the measurement stage, regardless of whether it is conducted by human annotators or using automated text analysis. We argue it is desirable to decrease subjectivity in classification tasks to increase accuracy in computer-automated classification, to improve robustness, and to improve transparency about the decisions taken not only by researchers, but also by major entities in content moderation.

For example, if one argues that three concepts satisfy the condition of being uncivil, such as: (1) racial slurs, (2) profanity, and (3) insults, it can be determined if a piece of text contains any/all of those three concepts/elements. For example, the sentence, "just shut up, idiot" contains an insult, but not a racial slur or profanity. By changing the sentence to "just shut up, you fu***ng idiot", we have added a satisfying condition for profanity, but not a racial slur. Thus, it would satisfy 2/3 index elements of incivility. While additive indices are not new in the classification literature, the difference in our approach is that we use narrower and more specific categories of incivility, as we show next.

Six dimensions of online political incivility

Figure 12 in the appendix provides precise operational definitions for our conceptualizations of political incivility in online settings. We argue that incivility can be decomposed in a series of six dimensions, which can then be aggregated in different ways depending on the scope of the analysis. These six elements of our definition of incivility were developed by both drawing on existing theoretical conceptualizations of incivility in the literature or inductively creating elements from reading thousands of social media comments on politicians' social media pages.

In practice, our definition implies a series of six binary indices, which we describe below. When it comes to human annotation of social media posts, these can be operationalized as a series of yes/no questions related to each of these dimensions that can be posed to human coders. Messages can be classified in a variety of ways, either via trained researchers or through online services such as MTurk or other crowd-sourcing platforms. We argue that in any coding task it is essential to include examples of each type of incivility in order to increase the reliability and validity of the coding process.

Index 1 (Profanity): Vulgarity / Profanity The 'Vulgar / Profanity' category is perhaps the most easily recognized and obvious form of incivility. Simply using curse words is a very common way of using uncivil language. Index elements for vulgarity / profanity appear in most of past work and is generally easier to classify due to its reliance on using certain taboo words, which can be detected via dictionary methods. A problem arises when considering nuance in language, however. For example, a black person using the n-word on social media carries a much different meaning than if a white person used the n-word (Gitari et al., 2015).

Index 2 (Contempt): Name-Calling / Insults / Attacks' This category comes from a combination of Jamieson (1997) and Jamieson (1998)'s "name-calling" and "casts aspersions" categories. This encompasses calling various types of common incivility: calling somebody a nasty or unflattering name, insulting somebody or some group, or otherwise attacking a group or individual. This is by far the broadest category of incivility, but we attempting to separate name-calling, insults, and attacks on individuals into further smaller categories is complex. The concepts are all fairly similar conceptually, and coders had a difficult time determining which category a comment

most belonged to. In combining these concepts, we boosted accuracy but decreased conceptual specificity.

Index 3 (Seditious language): Claims of Un-American Activity This index element was developed inductively from reading thousands of Facebook comments. Many comments during the period of the 114th Congress called MCs or other members of government 'treasonous', 'traitorous', or called for their impeachment. We believe these technically fall into the "Name-Calling / Insults / Attacks" category, but argue there is something unique about these specific claims of anti-American or anti-country activity that warrants a separate category. We do not believe comments claiming a president has committed treason or violated the constitution conceptually belong with comments that simply call somebody a 'jerk' or 'idiot.' Naturally, this classification will depend largely on whether or not a government official has committed treasonous acts or impeachable offenses, which is often unknowable by a researcher.

Index 4 (Speech devaluation): Calls somebody a liar or devalues their speech This index element was adapted from the "liar" and "pejorative for speech" categories from Jamieson (1998). We argue that calling somebody a liar and referring to somebody's speech in a pejorative way (ex: "that's nonsense; gibberish; BS") are conceptually similar enough to warrant collapsing the two concepts into one category. Accusing somebody of being a liar and referring to somebody's speech as "nonsense" are quite similar concepts; both assert that the speech of somebody else is not valid.

Index 5 (Partisan vitriol): Negative stereotypes or negative assessments related to political party / ideology This element was adapted from Papacharissi (2004)'s third element of incivility related to stereotypes. While Papacharissi's classification of stereotypes was broad, we argue that stereotypes or negative assessments of people related to political party or ideology are inherently different and less severe than stereotypes or negative assessments of people related to race / ethnicity. Members of political parties are not protected classes of individuals like other groups are in the US, and thus stereotypes surrounding political parties are less severe than stereotypes surrounding racial/ethnic groups.

Index 6 (Political threats): Threatens or calls for electoral consequences for a member of

Congress This is an element that was generated inductively. Many comments that we read called for MCs to resign, retire, be "stopped", or directly threatened to kick somebody out of office/ not vote for them. Advocating or threatening electoral consequences for an MC is conceptually different than other index elements. It is distinct from both simple insults and claims of treasonous, traitorous, or constitution-violating activity.

One could argue that citizens logging onto social media to publicly threaten to use their voting power is actually an example of civic action, and not incivility at all. To this point, we reiterate our stance on a binary index method. By separating out these categories of incivility and analyzing all messages for each type of incivility, criticisms of any index element are alleviated. With conceptually distinct index elements of incivility, the validity of results is not put at risk by potential disagreement over any single category of incivility.

We believe these are six appropriate categories for classifying instances of common incivility on social media. These do not cover the entire range of incivility as they do not include severe instances of incivility such as hate speech or advocacy for violence, but we believe these categories adequately capture all instances of common incivility on social media.[2]

The primary advantage of an index method of classification, such as the one we use here, is that they are robust to contestations of conceptualization, which (Silva et al., 2016) and others suggest is inevitable with this type of task; indeed, we have disagreed with conceptualizations/operationalizations of work we have come across. Because all conceptualizations of incivility may perhaps be legitimately contested, we argue the best solution is to use a classification method that is highly tailored to a specific task, dividing a concept as broad as 'incivility' across multiple categories/index elements. If all index elements of our incivility conceptualization are binary in nature and have clear definitions for each condition or "index element," then we can test if any disputes with a specific index element change overall results.

[2]While we would have liked to include hate speech in our analysis, we found that it was exceedingly rare in our sample of public Facebook comments, most likely due to Facebook's automated detection systems, which delete comments that violate their Community Standards. In addition, since legislators' pages are likely to be moderated by campaign staff, most of the blatantly harmful comments that would qualify as hate speech are probably removed. In our sample, examples of hate speech, racial slurs, threats to democracy, advocating violence, and other such severe types of incivility were very rare. This is a common problem for other studies of hate speech. For example, Davidson et al. (2017) use a lexicon of hate speech-related terms to search Twitter, and they find only 5% of tweets in their sample of messages that had already been pre-filtered using a dictionary of hate speech words.

For example, if one defines "name-calling" as any piece of text that refers to another individual or group using a negative term such as 'jerk(s)' or 'idiot(s)' and that definition is challenged, it is easy to exclude that index from analysis and observe how it changes the results. Further, contestations of any operationalization of an index element before publication can be seriously considered and actually changed. This benefit improves as the number of categories one uses grows. For example, having two categories to classify still allows one to remove a category and re-run the analysis, but the benefits increase as index elements are added.

Understanding the prevalence of incivility on social media

Why is uncivil behavior so prevalent on social media? In this article we explore an argument that combines demand- and supply-side mechanisms. The first set of mechanisms is related to how citizens react to being exposed to incivility: based on past work, we claim that bitter and vitriolic content receives higher engagement (more likes, retweets, etc.) on social media because it generates anger or even outrage, as well as emotional arousal (Crockett, 2017; Mutz, 2016; Ryan, 2012). This may be particularly true when it comes to political incivility in a context of increased political disaffection and lower levels of trust in politicians (Torcal and Montero, 2006). Especially when incivility targets politicians, many citizens may even see it as a form of entertainment. For these reasons, we expect to observe that:

Hypothesis 1: *uncivil messages on social media receive higher engagement.*

This demand-side mechanism, in turn, is likely to create incentives for an increase in the supply of uncivil messages. Since most social media platforms rank content in reply or comment threads in part taking users' engagement as a key signal, we should expect uncivil messages to receive more visibility. This weakness can be exploited by minority of users (commonly characterized as "trolls") whose purpose is generate discord, either with a financial or political motivation. This is a phenomenon that is not unique to incivility – for example, Barberá and Rivero (2015) find that a minority of hyperpartisan users on Twitter are responsible for most political content, which as a result becomes more polarizing. Based on this, we hypothesize that:

Hypothesis 2: *a minority of users is disproportionally responsible for most uncivil content on social*

media.

At the intersection of supply and demand we find what we think is a key explanation to why political interactions on social media can become so bitter: incivility is contagious. In other words, "trolls" play a key role not only in driving up the prevalence of incivility through their own messages, which are amplified by ranking algorithms, but also because other social media users are likely to engage with them and respond with additional uncivil messages. This could happen either because other individuals may feel compelled to respond to what they see as an attack or simply because seeing that an uncivil message is already posted may affect their perception of the norms regulating what is permissible in that particular context.

To our knowledge, this type of second-order effects of incivility have not been identified empirically in past work, but we believe they represent an important mechanism whose existence has important implications for content moderation practices. For this reason, in this paper we also seek to test whether:

Hypothesis 3: *uncivil messages on social media are more likely to receive uncivil responses.*

Research design and Data

To test our hypotheses we use publicly available comments on the Facebook pages of members of the U.S. Congress. We believe this is a good case study because Facebook is the most prominent social media site in the US. We selected political pages of legislators as our population of interest because, as opposed to other similarly public online spaces such as the pages of media outlets or interest groups, an overwhelming majority of conversations that take place are related to politics. In addition, these pages represent an important mechanism for citizens to exercise the political accountability of their political representatives, as well as an additional source of information for politicians to learn about their constituents' preferences. Measuring the prevalence of incivility in these spaces is thus normatively and empirically relevant.

We gathered data from the Facebook pages of 453 members of the 114th US Congress (MCs) from January 2015 until December 2016. For each MC, we collected all posts and comments using Facebook's public Graph API, which until 2018 allowed researchers to access publicly available

data. In this paper we focus our attention on the nearly 7.5 million comments that were published on the MCs' pages during this period. As we describe in the following sections, we will a combination of supervised machine learning and dictionary methods to predict whether each of these comments can be considered uncivil, and which type of incivility each of them represents.

Finally, we enrich this dataset with legislator-level covariates: their gender, political party, and chamber (available through the Congressional Biographical Directory), and estimates of their political ideology using DW-NOMINATE scores (Lewis et al., 2018).

Measuring incivility

Given the complexity of conceptualizing incivility, ideally researchers would rely on trained human annotators to classify sets of documents according to whether they can be considered uncivil or not. In practice, however, this is simply not feasible given the size of the datasets, particularly when we consider sets of social media messages. It is for this reason that in this paper we rely on automated text analysis methods in order to identify incivility at scale.

Our work joins the efforts of many scholars that have previously attempted to conceptualize and measure negative speech at scale. When trying to derive insights from this literature, one key challenge is the large variation in how incivility is characterized. Using the distinction introduced in earlier sections in this paper, we find that some scholars focus on severe incivility, such as hate speech (Silva et al., 2016; Davidson et al., 2017; Burnap and Williams, 2015) or "offensive language" (Davidson et al., 2017), while even more still study general incivility (Papacharissi, 2004; Borah, 2014; Rosner et al., 2016; Blom et al., 2014; Oz et al., 2017; Jomini et al., 2015; Kevin et al., 2014). While they all focus on similar topics, differences in how negative speech is defined and operationalized makes it difficult to compare results across studies.

By further specifying and narrowing classification tasks, social media platforms may both avoid political issues and improve the number of harmful posts removed by developing and improving on a number of very narrow classifiers, making them highly specific but also (ideally) very accurate.

Figure 1 below includes a table depicting the performances of some of the incivility classifiers

we encountered in the literature. Because some values are reported as a score out of 100 and other scores were reported as a score from 0-1, values have been normalized to represent values from 0-1. Accuracy, precision, recall, and F1 scores for incivility classification tasks are often fairly low. The exceptions to this general pattern are papers whose definition of incivility is narrow.

Table 1: Model performances for classifying various types of negative speech

Authors	Model	Domain	Classifier	Precision	Recall	F1
Badjatiya et al 201	LSTM+Random Embedding + GBDT	16k Tweets	sexism, racism	0.93	0.93	0.93
Davidson et al 2017	logistic regression with L2 regularization	25k tweets	offensive language but not hate speech	0.96	0.91	0.93
Davidson et al 2017	logistic regression with L2 regularization	25k tweets	neither hate speech nor offensive language	0.83	0.94	0.88
Nobata et al 2016	Vowpal Wabbit Regression; multiple models	448k Yahoo! Finance comments and 726k Yahoo! News comments	abusive language	0.773	0.794	0.783
Burnap and Williams 2015	ensemble vote: bayeslian logistic regression, random-forest decision trees, and support vector machines on data transformed to n-gram reduced typed dependencies and hateful terms	450k tweets	hateful or antagonistic responses toward minority groups	0.89	0.69	0.77
waseem and hovy 2016	character n-grams	16k tweets	sexism, racism	0.7287	0.7775	0.7389
Warner and Hirschberg 2012	template-based feature extraction	9000 paragraphs from Yahoo! news group posts and American Jewish Congress-identified offensive websites	anti-semitism	0.68	0.6	0.6375
Sadeque et al 2019	Neural Network (GRU)	Newspaper comment sections; Russian troll comments	Vulgarity	0.4872	0.5757	0.5277
Davidson et al 2017	logistic regression with L2 regularization	25k tweets	hate speech	0.44	0.59	0.51
Sadeque et al 2019	Neural Network (GRU)	Newspaper comment sections; Russian troll comments	Name-calling	0.4576	0.5063	0.4807
Silva et al 2016	novel sentence structure classifier: I {intensity} {user intent} {any word}	27.55 million whispers	targets of hate speech	100	low; not reported	NA
Silva et al 2016	novel sentence structure classifier: I {intensity} {user intent} {any word}	512 million tweets	targets of hate speech	100	low; not reported	NA

For example, Badjatiya et al. (2017) classify hate speech in tweets using deep learning methods trained with data labeled by Waseem and Hovy (2016). They boast precision, recall, and F1-score measures of 0.93, which is among the highest in the literature. However, they focus only on sexism and racism, which they measuring using an index that includes "problematic hashtags", which limits its generalizability. testing this greater sign.

Although it is difficult to draw generalizations from this set of results because the classification tasks differ, as do the definitions of each type of negative speech, the key takeaway point for us is that narrower definitions of negative speech lead to better performance. We believe this supports our suggestion of using a set of binary indices instead of broader definition of incivility.

Building a training dataset

The first step in order to build a system to automate any coding task is to construct a dataset that contains comments labeled according to the type of incivility they contain. These labels will be assigned to a random sample of comments by a set of trained human annotators. The training set will be then be then be used both for training purposes – in other words, to try to extrapolate human coding to the entire corpus of comments using automated methods – and to evaluate the performance of those automated text analysis methods.

We coded a random sample of approximately 5,000 comments (stratified by MC) using Figure Eight (formerly called Crowdflower), a crowdsourcing classification platform (Benoit et al., 2016). As we described earlier, our codebook (shown in Figure 12) was developed both based on theory and inductively motivated by reading Facebook comments and iteratively improving the codebook through our initial rounds of coding, in line with Grimmer and Stewart (2013)'s recommendations on codebook development.

In order to improve the reliability and cost-efficiency of our coding approach, we introduced two key innovations with respect to previous studies:

(1) First, instead of coding incivility for each single comment, we only do so for comments that our coders identified as having a negative valence. In other words, we rely on a two-step coding workflow. The first step is a sentiment question that asked coders to answer a question regarding

whether the main tone of the comment was positive, neutral or negative. This approach partially builds on work by Gitari et al. (2015) and has as its main advantage the fact that it is faster and cheaper since it avoids paying coders to classify a significant amount of positive or neutral tone messages for incivility, which by definition are not uncivil. Approximately 2,300 out of the 5,000 comments were classified as negative and thus will be coded by incivility. In our analysis, we will also build a two-step classifier: first we detect which comments are negative and then we will classify them into each of our incivility categories.

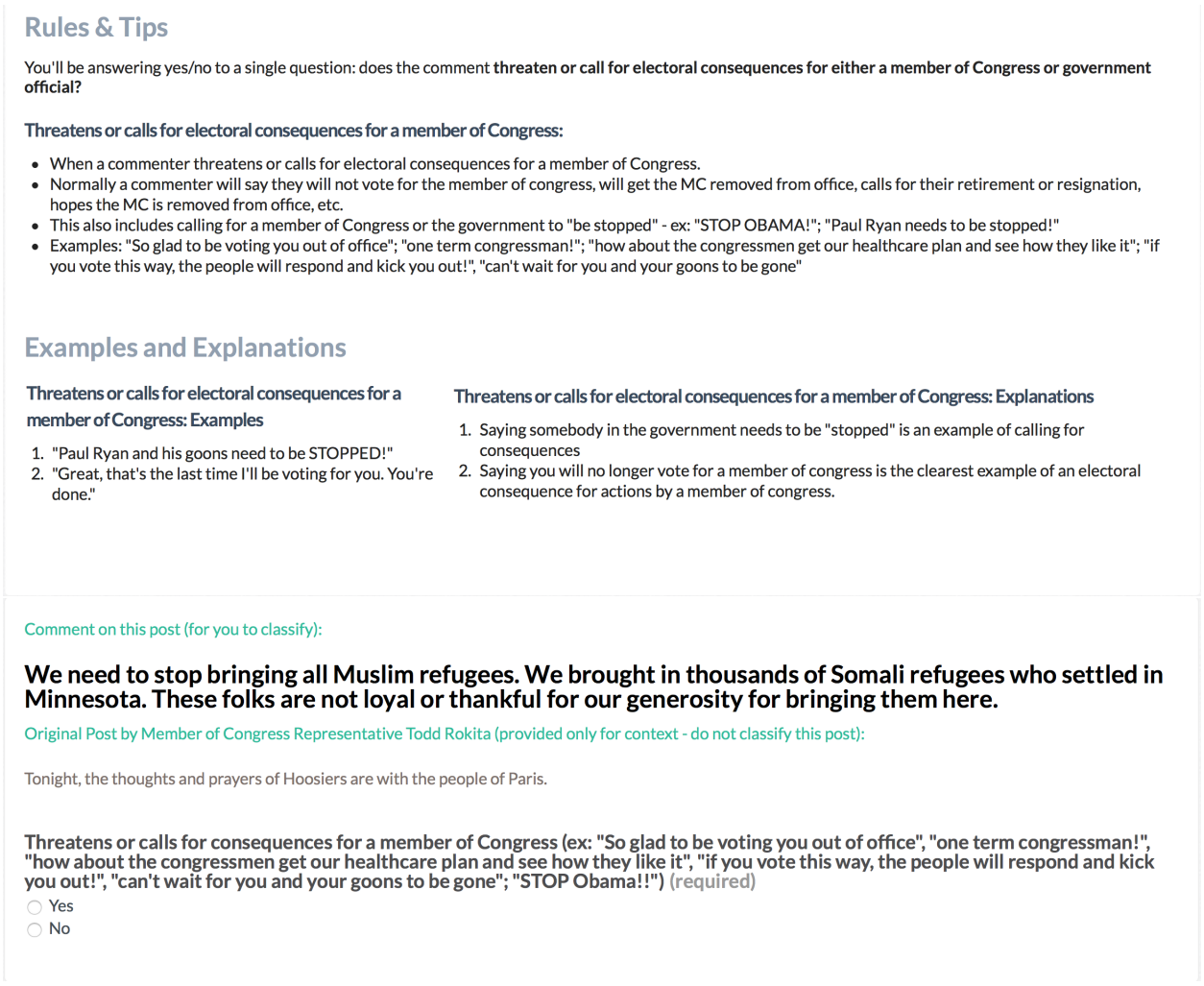
(2) Building upon our conceptualization of incivility, we designed the coding task as a series of binary questions regarding each of the incivility categories for each comment separately. Instead of showing coders a single post and asking them to identify which of the six categories were present, this approach breaks down the classification task into a series of six discrete tasks. In each task, only one question is being asked at a time, as shown in Figure 1. While this approach may appear to be inefficient (after coders have already read one post, why not code the rest?), in practice we found that it achieved much greater intercoder reliability and speed at only slightly higher cost. Our view is that this approach reduces mental fatigue because it avoids constantly switching across categories, and allows coder to develop an expertise in each of the categories.

We combined these two decisions with Figure Eight's built-in data quality system, which asks the workers to first code a set of "gold questions" manually labeled by the authors, and only allows them to continue labeling data as long as their responses to these questions are at least 70% in agreement with our own labels.

Automated text classification

Once we have obtained a labeled set, we consider two different approaches to generalize the coding conducted by our sample of workers to our entire database of 7.5 million comments. First, we explore a supervised learning classifier that learns which features are associated with each of these categories in the training set and then extrapolate the labels to the uncoded data. Second, we also experiment with a dictionary method, combined with query expansion using word embeddings, that manually tries to identify these words based on the authors' substantive knowledge.

Figure 1: Preview of a Coder's Actual Classification Job for a single 1-item Text



Rules & Tips

You'll be answering yes/no to a single question: does the comment **threaten or call for electoral consequences for either a member of Congress or government official?**

Threatens or calls for electoral consequences for a member of Congress:

- When a commenter threatens or calls for electoral consequences for a member of Congress.
- Normally a commenter will say they will not vote for the member of congress, will get the MC removed from office, calls for their retirement or resignation, hopes the MC is removed from office, etc.
- This also includes calling for a member of Congress or the government to "be stopped" - ex: "STOP OBAMA!"; "Paul Ryan needs to be stopped!"
- Examples: "So glad to be voting you out of office"; "one term congressman!"; "how about the congressmen get our healthcare plan and see how they like it"; "if you vote this way, the people will respond and kick you out!", "can't wait for you and your goons to be gone"

Examples and Explanations

Threatens or calls for electoral consequences for a member of Congress: Examples

1. "Paul Ryan and his goons need to be STOPPED!"
2. "Great, that's the last time I'll be voting for you. You're done."

Threatens or calls for electoral consequences for a member of Congress: Explanations

1. Saying somebody in the government needs to be "stopped" is an example of calling for consequences
2. Saying you will no longer vote for a member of congress is the clearest example of an electoral consequence for actions by a member of congress.

Comment on this post (for you to classify):

We need to stop bringing all Muslim refugees. We brought in thousands of Somali refugees who settled in Minnesota. These folks are not loyal or thankful for our generosity for bringing them here.

Original Post by Member of Congress Representative Todd Rokita (provided only for context - do not classify this post):

Tonight, the thoughts and prayers of Hoosiers are with the people of Paris.

Threatens or calls for consequences for a member of Congress (ex: "So glad to be voting you out of office", "one term congressman!", "how about the congressmen get our healthcare plan and see how they like it", "if you vote this way, the people will respond and kick you out!", "can't wait for you and your goons to be gone"; "STOP Obama!!") (required)

- ○ Yes
- ○ No

We describe each of these methods below.

For the supervised learning approach, we used *xgboost* (Chen and Guestrin, 2016), a state-of-the-art machine classification method that relies on gradient boosting (an ensemble of decision trees), and which has been recently found to maximize classification accuracy in most tasks (Olson et al., 2017). We trained this classifier using bag-of-words (unigrams and bigrams) and 100-dimensional word2vec embeddings (Mikolov et al., 2013) trained on the full corpus of comments and then taking the average comment embedding as features. We used 5-fold cross-validation to identify the parameters that maximize in-sample performance, and then measure how well it performs on a 20% sample of the training dataset left out of the estimation.

Table 2 reports the results of our supervised learning classifiers. While our first-level classifier (sentiment) achieves an excellent level of performance that is comparable to human coding, we generally find that the other models perform rather poorly, with precision and recall below 10% in many cases. This table also offers some descriptive statistics regarding the relative prevalence of each type of incivility – contempt (name-calling, insults, attacks...) is by far the most frequent, with nearly 60% of comments being coded as such; whereas the other categories only are present in around 2–10% of comments. This smaller sample size likely explains the worse performance of these models, given that we only have around 100 positive labels for each case. When the sample size is larger, as in the case of contempt, both precision and recall are high and above acceptable levels, close to intercoder reliability.

Given its poor performance, we also explore building improved dictionaries as an alternative method to supervised learning classification. To do so, we start with a set of seed words that based on theory and our own reading of thousands of comments we identified as being related to each concept. Then, we used the trained word embeddings to identify other words that were semantically related to our seed words. Semantic similarity here is based on these words appearing in similar contexts, and we computed using cosine similarity on the word embedding space. For more details about word embeddings and a political science application, see Gurciullo and Mikhaylov (2017). Figure 2 shows an example of how we started with seed words and then found other potentially relevant words. After doing this, we would select words that based on our do-

Table 2: Out-of-sample performance of machine learning classifiers and dictionary methods

			Machine learning			Dictionary method		
Category	Prop.	IR	Prec.	Rec.	F1	Prec.	Rec.	F1
Negative	48%	86%	.755	.843	.797	–	–	–
Profanity	4%	94%	.100	.125	.111	**.578**	**.488**	**.529**
Contempt	60%	82%	**.679**	**.738**	**.707**	.757	.131	.223
Seditious language	2%	95%	.000	.000	–	**.111**	**.378**	**.172**
Speech devaluation	2%	96%	1.00	.053	.100	**.702**	**.745**	**.723**
Partisan vitriol	5%	85%	.429	.115	.181	**.251**	**.452**	**.323**
Political threats	8%	90%	.250	.069	.108	**.296**	**.462**	**.361**

Notes: *Prop.* is the proportion of comments in each category in the training set (a stratified random sample of posts); *IR* is the average pairwise agreement on the coders' responses; *precision* is the % of comments predicted to be in that category that are correctly classified; *recall* is the % of comments in that category that are correctly classified. Precision and recall are computed for the positive value of each category.

main knowledge could be relevant to the latent concept being measured.

Table 2 also reports the performance of this dictionary approach, again measured using the training dataset labeled by crowd workers. In this case, we find levels of precision and recall that are higher than with supervised learning. Although they are still below what could be considered desirable, they are in line with the average performance of existing classifiers in past work (see Figure 1). Given the difficulty of identifying some of these categories, we consider these results acceptable for our task. In the rest of our analysis we will use the estimates from the dictionary methods, with the only exceptions being sentiment and the contempt category, where we still use the supervised classifier given that it performed better. Using dictionaries in this way may provide a way to flag or identify potentially harmful comments that supervised models do not pick up, possibly due to small sample size.

Results

Before testing our main hypotheses, we first offer a descriptive analysis of incivility on Facebook pages of U.S. politicians. Here we define as "uncivil" any comment that is predicted to fall in at least one of the six categories of incivility we use. We find that nearly 51% of all comments are classified as negative. Of these, a large majority (40% of all comments) fall into at least one of the six categories of incivility.

Figure 2: Example: word2vec-enhanced dictionaries

```
> distance(file_name = "FBvec.bin",
+          search_word = "libtard",
+          num = 10)
Entered word or sentence: libtard

Word: libtard  Position in vocabulary: 5753
          word              dist
1          lib 0.798957586288452
2        lefty 0.771853387355804
3      libturd 0.762575328350067
4    teabagger 0.744283258914948
5     teabilly 0.715277075767517
6      liberal 0.709996342658997
7       retard 0.690707504749298
8      dumbass 0.690422177314758
9         rwnj 0.684058785438538
10 republitard 0.678197801113129
```

```
> distance(file_name = "FBvec.bin",
+          search_word = "idiot",
+          num = 10)
Entered word or sentence: idiot

Word: idiot  Position in vocabulary: 646
        word              dist
1   imbecile 0.867565214633942
2    asshole 0.848560094833374
3      moron 0.781079053878784
4     asshat 0.772150039672852
5     a-hole 0.765781462192535
6      ahole 0.760824918746948
7    asswipe 0.742586553096771
8  ignoramus 0.735219776630402
9   arsehole 0.732272684574127
10     idoit 0.720151424407959
```

Figure 3 breaks down the proportion of negative and uncivil comments across party and gender groups, with the vertical axis displaying the proportion of comments in each category on a legislator's page. Both panels point in the same direction: pages of female legislators and Republican Members of Congress attract a much higher number of comments that our methods classified as being negative and uncivil.

Figure 4 offers a different visualization of the data using violin plots and labeling the three legislators with the highest and lowest values for each category. This allows us to observe that many of the Members of Congress whose pages feature the highest negativity and incivility values appear to be prominent legislators, such as Harry Reid, Paul Ryan, and Lindsay Graham. However, incivility is relatively high across all legislators' pages: even for the pages with most civil comments, the proportion that is classified as uncivil in at least one of the six dimensions we consider is around 10%.

Figure 5 disaggregates incivility into the six dimensions we consider, and across the gender of the Member of Congress. Contempt (name-calling, insults, and attacks) represents by far the most frequent category of incivility we observe, and appears to be where we find the largest difference between gender. Across the other categories, seditious language, partisan vitriol and speech

Figure 3: Prevalence of Negative and Uncivil Comments, by Party and Gender

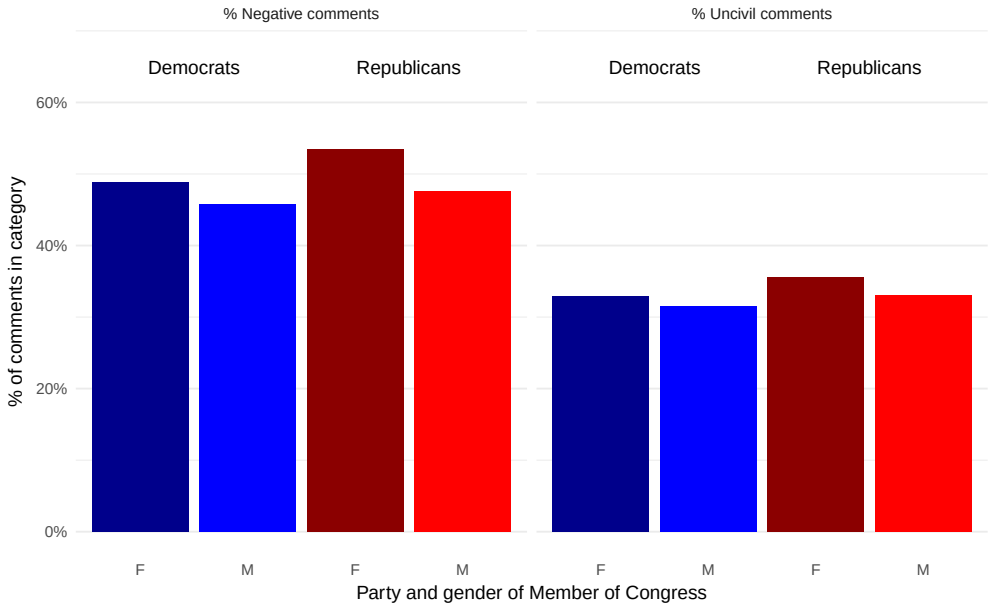
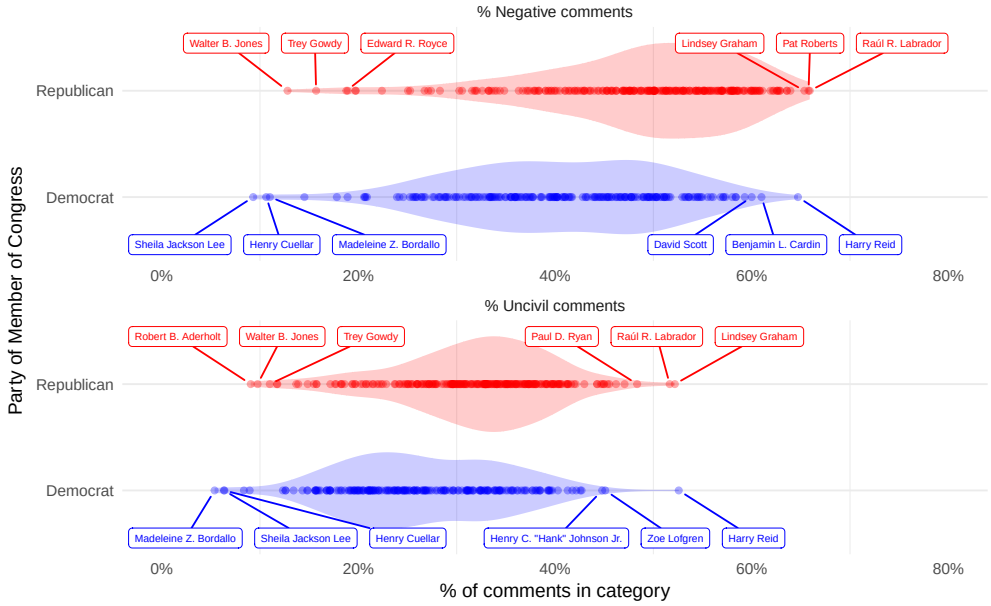


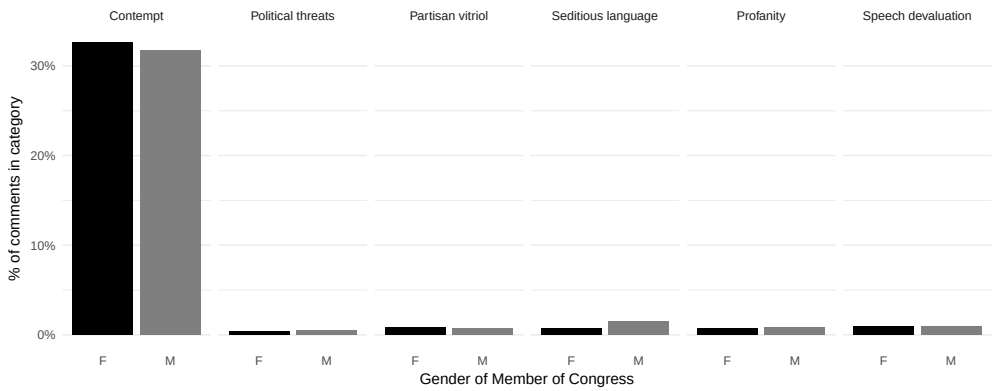
Figure 4: Distribution of Prevalence of Incivility, by Party



devaluation appear to be somewhat more prevalent, although without large differences across gender groups.

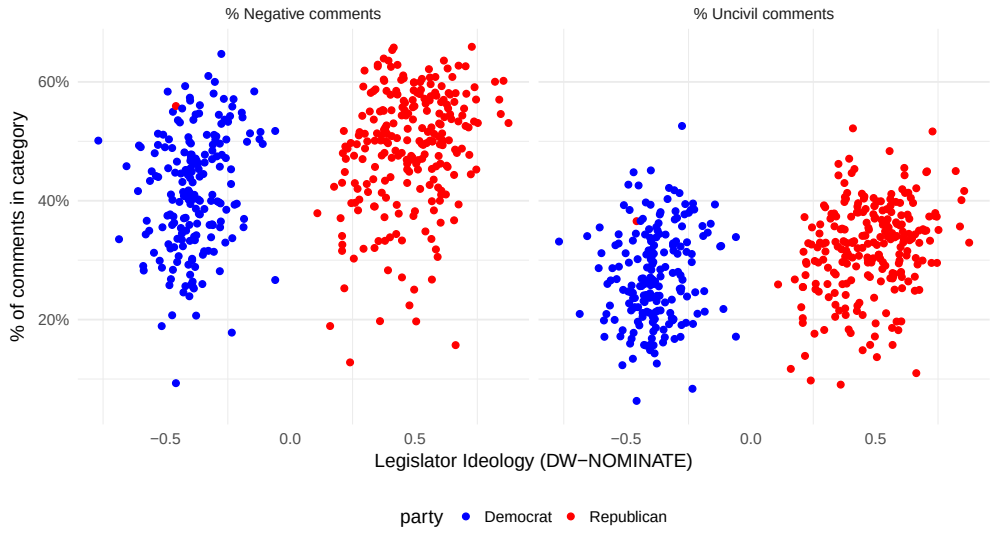
As a final type of bivariate analysis, we examine how political ideology – measured using

Figure 5: Prevalence of Uncivil Comments, by Gender and Type of Incivility



NOMINATE scores collected from VoteView (Lewis et al., 2017) – correlates with the degree to which a legislator's page attracts negative and uncivil comments. The results, shown in Figure 6, do not reveal any clear pattern regarding this relationship.

Figure 6: Prevalence of Uncivil Comments, by Ideology



Of course, one key limitation of these bivariate analyses is that our results could be driven by confounding variables. For example, since most female legislators are Democrats, our results regarding gender could simply due to a party effect. To better understand the prevalence of incivility on legislators' pages, we also estimate multivariate linear regressions at the MC level (using weights based on comment count). Here, the dependent variable is the proportion of comments

on a given legislator's page that fall into each of the six categories of incivility we consider, as well as negativity overall and an aggregate index of incivility measured as the proportion of posts that fall into at least one category. As independent variables, we consider party ID, gender, chamber, and extremity (measured as the absolute value of each legislator's NOMINATE score).

Table 3: WLS regression of % comments in category on a legislator's page

	Neg.	Uncivil	Contempt	Threats	Deval.	Vitriol	Profan.	Seditious
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Republican	4.16^{***}	3.06^{***}	2.82^{***}	0.16^{***}	0.06	-0.001	-0.26^{***}	0.91^{***}
	(1.38)	(1.08)	(1.03)	(0.03)	(0.06)	(0.06)	(0.07)	(0.11)
Male	-3.87^{**}	-1.57	-1.58	0.03	0.07	-0.08	0.24^{***}	0.31^{***}
	(1.52)	(1.19)	(1.14)	(0.04)	(0.07)	(0.06)	(0.07)	(0.12)
Senator	11.09^{***}	8.53^{***}	8.21^{***}	0.11^{***}	0.31^{***}	0.20^{***}	0.28^{***}	0.60^{***}
	(1.17)	(0.91)	(0.87)	(0.03)	(0.05)	(0.05)	(0.06)	(0.09)
Extremity	-11.03^{***}	-6.18^{**}	-5.94^{**}	-0.32^{***}	-0.54^{***}	-0.72^{***}	-0.26	-0.30
	(3.45)	(2.70)	(2.58)	(0.08)	(0.15)	(0.14)	(0.17)	(0.27)
Intercept	49.50^{***}	31.86^{***}	30.61^{***}	0.54^{***}	1.05^{***}	1.15^{***}	0.76^{***}	0.41^{**}
	(2.20)	(1.72)	(1.65)	(0.05)	(0.10)	(0.09)	(0.11)	(0.17)
N	432	432	432	432	432	432	432	432
Adjusted R^2	0.18	0.17	0.17	0.10	0.09	0.09	0.10	0.26

Note: *p<.10; **p<.05; ***p<.01. The dependent variable is the proportion (0-100) of comments that are classified in each category over the total of comments on each legislator's page. Extremity is defined as the absolute value of the legislator's DW-NOMINATE score (dimension 1).

We show our results in Table 3. We find three clear patterns. First, Senators systematically attract much more negativity and incivility than Representatives, which is perhaps not surprising given that they represent an entire state and as such are more likely to be prominent and also to attract heterogeneous (and thus more critical) audiences. Second, the pages of ideologically extreme legislators actually feature a *lower* level of negativity and incivility. And third, name-calling and insults (contempt) are significantly more frequent (around 3 percentage points higher) in the pages of Republican legislators, but other types of incivility are similarly common across both groups of pages. Contrary to our earlier result, we do not find that gender is a predictor of incivility once we control for these other variables.

An important note here is that if stating that certain types of legislators "attract" more uncivil or negative comments may be misleading because we do not know who were the specific targets of negativity or incivility in the messages analyzed, only the pages on which those messages were posted. It is possible that animosity in these pages is addressed towards other commenters, and

not necessarily the Member of Congress. However, we argue that both types of incivility may be equally damaging to the quality of the public conversations that takes place in these pages.

Now that we have demonstrated that negativity and incivility are highly prevalent on legislators' Facebook pages, we now turn to examine the evidence regarding each of our three hypotheses.

Uncivil comments receive higher engagement

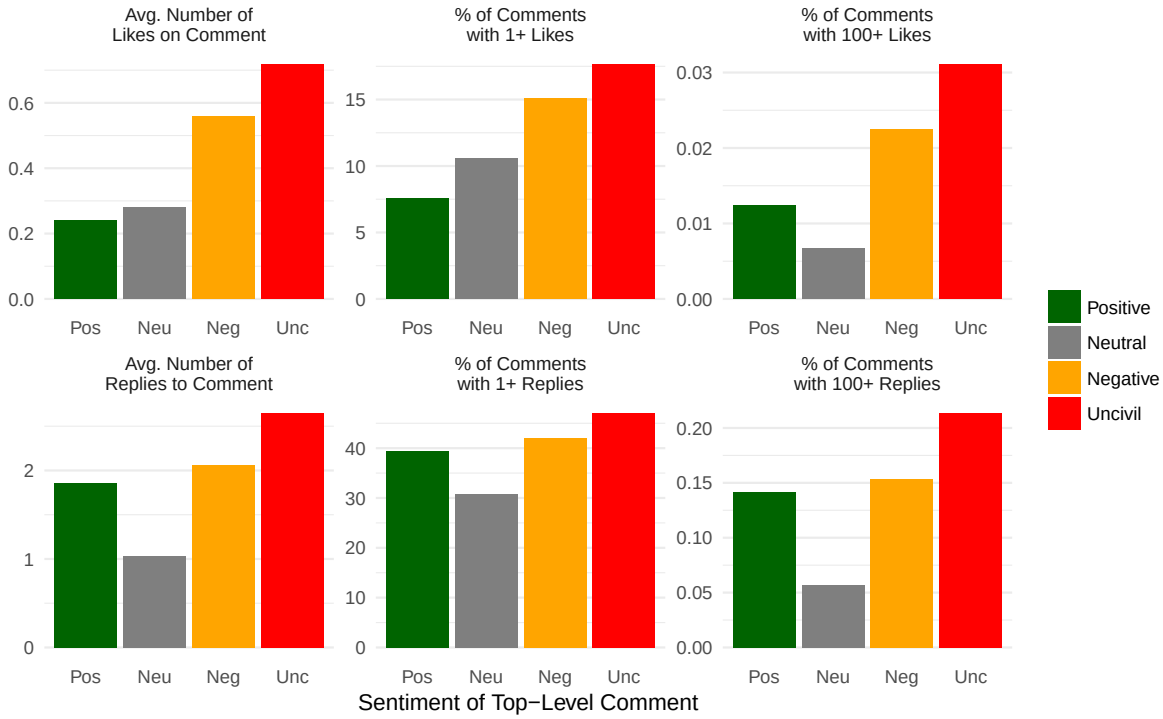
How do people react to uncivil comments? As we discussed in our literature review, one of the defining characteristics of incivility is that it undermines citizens' right to freely express their opinions. But the answer to this question is also important from a more mechanical perspective – the current system that Facebook uses to rank comments appears to be based at least partially on user engagement. In other words, when there are more than two or three comments on a post, the comments are not displayed chronologically. Instead, Facebook chooses to show the "Most relevant" ("comments with most views, reactions, and replies," according to how it is described on the website). It is thus important to understand whether uncivil comments raise to the top according to this metric, which would increase their visibility.

To test this possibility, we collected data on the number of likes and replies that each comment received. As we show in Figure 7, each individual comment on Facebook can be liked by other users, and it can also receive a reply, which would lead to a nested comment. We will refer to to the comments that respond to a legislator's post as "top-level" or "root" comments, whereas any response to a top-level comment will be defined here as a "reply."

Figure 8 shows the average of three alternative metrics for two different types of engagement (number of likes on comments and number of replies to comments). We compute these metrics for each of the four potential top-level comment classifications: positive, neutral, negative (but civil) and negative (and uncivil according to a least one dimension). We find systematic evidence that uncivil comments receive higher engagement: they obtain three times more likes and around 50% more replies than positive or neutral comments. This result is not simply due to potential negativity bias, since uncivil comments also have higher engagement than negative (but civil comments).

Figure 7: Example of different types of Facebook comment responses

Figure 8: Uncivil comments receive higher engagement



Note that the estimated level of engagement with uncivil comments is likely an underestimate because not all Facebook users respond to a comment via the reply button. Users may reply to another comment by writing their message in the general thread, usually writing the name of the

person or tagging them. Figure 7 illustrates this point. These sort of general-thread responses appear to be fairly common from our observations of Facebook comment threads.

A minority of users is responsible for most incivility

Who are the perpetrators of incivility? An analysis of the characteristics that predict whether someone will post an uncivil messages is not feasible, given our lack of individual-level data. However, we can use the unique identifiers (available through the Facebook API) of the users posting the uncivil comments to examine the extent to which most users send uncivil messages or whether it is only a small and loud minority.

Figure 9 displays our attempt at testing this second hypotheses. Here, each curve shows the cumulative distribution of the proportion of users sending a proportion of messages within each category – positive, negative but civil, and negative and uncivil.

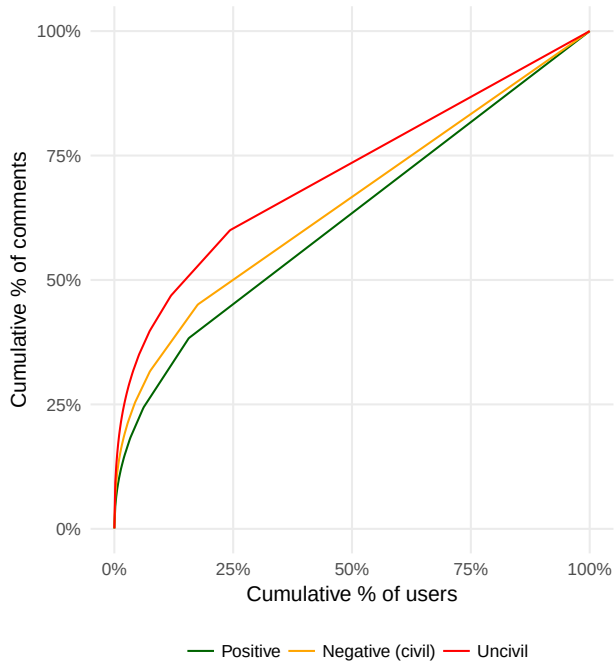
We find that that 19% of all uncivil comments were generated by only 1% of users; and that 60% of all uncivil comments were posted by 25% of users. In other words, a small minority of users is responsible for a large majority of incivility. Furthermore, as demonstrated by the fact that the corresponding Lorenz curves for positive and negative civil comments are *below* the curve for uncivil comments, this concentration in the production of content among a minority of users is larger for uncivil comments.

The cycle of incivility

The final step in our analysis consists on testing our hypothesis regarding diffusion of incivility. Our main interest here is to understand whether "incivility begets incivility," with at least part of the existing vitriol on social media being a consequence of citizens simply responding or reacting to already existing incivility, with a negative compound effect on the quality of political deliberation that could take place within Facebook pages. To address this question, we again exploit the nested structure of comments and analyze whether uncivil top-level comments are more likely to receive uncivil replies than civil top-level comments.

Figure 10 displays the results of a bivariate analysis exploring this hypothesis. We find that

Figure 9: Lorenz curve showing amount of users responsible for types of speech

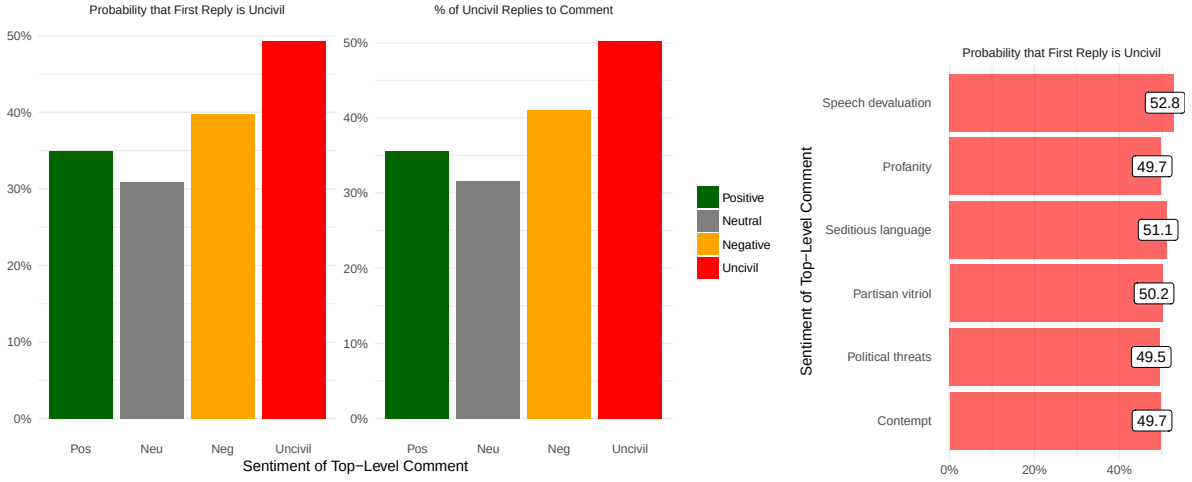


uncivil comments are between 25% and 4% more likely to receive uncivil replies, either as the first response to the comment (left panel) or as the overall volume of uncivil replies in the entire thread (rigth panel). We can also disaggregate across types of incivility, as shown on the right panel. Although the differences here are not as large, we do find that speech devaluation (calling someone a liar) is the type of incivility that appears to be somewhat more "contagious."

One limitation of this bivariate analysis is that other confounding factors may affect both the probability that a top-level comment is uncivil and that it receives uncivil replies. For example, posts about controversial topics may elicit more vitriolic responses overall. Or specific pages of some legislators, as we found in our earlier analyses, may attract more incivility in general. To address this concern, we also fit multivariate regression models that control for a wide range of confounding factors.

Table 4 displays the results of two sets of multivariate regression models where the unit of analysis is the comment thread (that is, a top-level comment and all the subsequent replies). We consider two different dependent variables: the total number of replies to a top-level comment

Figure 10: Uncivil comments receive higher engagement



(Models 1–3) and the proportion of uncivil replies to a top-level comment that received at least one reply (Models 4–5). Our main independent variables will top-level comment sentiment (positive or negative, with neutral as the base category) and whether the root comment is uncivil.

For each dependent variable, we explore three different model specifications. The first model includes the same set of legislator-level control variables as those in Table 3. The second model tries to account for other legislator-specific effects not accounted by these covariates. It does so by including legislator-fixed effects, as well as time-fixed effects. Since all legislator covariates are constant within legislator, they are dropped from the regression. Finally, the third model tries to control for any topic-specific effects by including post-level topic probabilities from an Latent Dirichlet Allocation model. The LDA model is computed using the text of each legislator's post as document and extracting 100 topics. As we show in Figure 11 in the Appendix, most of these topics validly capture political issues (e.g. health care, terrorism, national security, gun control, budget discussions, environmental policy, etc.) and indeed generate varying degrees of uncivil responses.

The results here offer strong empirical evidence for the existence of contagion effects. Compared to neutral top-level comments, comments that are negative and uncivil are estimated to receive around 8.5% more replies.[3] Part of this effect is due to the fact that negativity in gen-

[3]Note that since we consider all uncivil comments to be negative, we need to interpret the first coefficient in the

eral is more likely to elicit additional responses: negative (but civil) comments receive 6.5% more replies. However, incivility receives a response boost of around 2 percentage points beyond that. This effect may seem small, but note that it is over four times larger than the estimated difference between Senators and Representatives – with the former predicted to receive only around 2% the large difference in their potential audience size. Positive comments, in contrast, are less likely to receive replies (around 3% less). These results are robust to the inclusion of legislator- and time-fixed effects and topic controls.

A top-level comment that features any of the six dimensions of incivility is likely to garner more attention - but is the additional set of responses also more likely to be uncivil? Our results also support this "contagion effect" hypothesis: the proportion of uncivil replies to an uncivil top-level comment is 11 percentage points compared to a neutral top-level comment. Confirming the results of our bivariate analysis, we again find that negativity incites incivility, and that negativity combined with incivility attracts even more uncivil replies. As earlier, these findings are larger in magnitude: they represent around a quarter of the difference between the legislators with the lowest and highest proportions of incivility in their Facebook pages. And they are robust to the inclusion of all legislator- and post-level controls.

Discussion and Conclusion

The overarching goal of this paper was to advance our understanding of how incivility on social media can be conceptualized and measured, and to try to offer new mechanisms that explain its high prevalence. To achieve this goal, we offered a descriptive analysis of the extent to which incivility is an exception or the norm in the Facebook pages of U.S. Members of Congress.

Our results show that over 40% of comments on these pages can be categorized into at least one of the six dimensions of incivility we considered, and that a mechanism explaining such high prevalence could be that uncivil comments receive more visibility because users tend to engage more with them. In turn, this creates an incentive for people to post more uncivil content, because it is more likely to receive visibility than generic positive content. And since uncivil comments

regression is as if it were the multiplicative term in an interaction effect. In contrast, the coefficient for whether the top-level comment is negative corresponds to the difference between neutral and negative (but civil) comments.

Table 4: OLS regression: the cycle of incivility

	Dependent variable:					
	log(replies)	log(replies)	log(replies)	% uncivil	% uncivil	% uncivil
	(1)	(2)	(3)	(4)	(5)	(6)
Comm. is negative+uncivil	1.9^{***}	2.1^{***}	2.2^{***}	8.7^{***}	7.9^{***}	7.0^{***}
	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Comment is negative	6.4^{***}	6.6^{***}	6.4^{***}	8.3^{***}	4.1^{***}	4.0^{***}
	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.2)
Comment is positive	-3.7^{***}	-3.3^{***}	-3.1^{***}	3.1^{***}	1.5^{***}	1.3^{***}
	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.2)
Republican	-1.3^{***}			-0.5^{***}		
	(0.1)			(0.1)		
Extremity	2.3^{***}			-1.8^{***}		
	(0.2)			(0.3)		
Senator	1.8^{***}			5.9^{***}		
	(0.1)			(0.1)		
Male	-3.0^{***}			0.1		
	(0.1)			(0.1)		
Intercept	11.5^{***}			25.4^{***}		
	(0.2)			(0.2)		
Legislator fixed effects	No	Yes	Yes	No	Yes	Yes
Year-month fixed effects	No	Yes	Yes	No	Yes	Yes
Topic controls	No	No	Yes	No	No	Yes
N	1,000,000	1,000,000	1,000,000	528,079	528,079	528,079
Adjusted R^2	0.02	0.02	0.02	0.04	0.02	0.03

Note: *p<.10; **p<.05; ***p<.01. The dependent variable in Models 1–3 is the log of replies to a root comment (multiplied by 100 in order to facilitate interpretation). The dependent variable in Models 4–6 is the proportion (0-100) of uncivil replies to a root comment (only for root comments with at least one reply). Models 1–3 were fit with a random sample of comments for computational reasons.

elicit additional uncivil responses, the outcome of this process is a vicious cycle that reduces the quality of political conversations on Facebook and limits its potential as a space for public deliberation.

While we think our analysis has found strong evidence in support of our argument, we also acknowledge limitations in our approach. First, we are not able to examine who is the target of uncivil comments. In many cases, the attacks could be addressed not towards the Member of Congress but instead towards another user on Facebook. These two types of incivility are qualitatively very different and they have also varying normative implications: whereas incivility towards politicians could perhaps even facilitate accountability, harassment of another user could mean that he or she decides to stop talking about politics. If such harassment is concentrated

on specific segments of the population, such as racial or political minorities, its impact could be undoubtedly pernicious. Unfortunately our analysis does not allow us to make this important distinction.

Second, there is room for improvement regarding the performance of our automated classification methods. As we have shown, we don't do much better or worse than other existing approaches, but despite our best efforts to rely on recent innovations in text analysis, it is possible that we may be under- or over-estimating the prevalence of incivility. However, given that the categories on which we do worse are relatively rare, we believe that this shouldn't affect our results much.

And third, we note that some aspects of common incivility may have positive effects from a normative perspective. For example, Brooks and Geer (2007b) show that incivility does not appear to have large detrimental effects among the public and, in fact, may have some modest positive consequences for the political engagement of the electorate. Of course, as explained earlier, we don't claim that's the case for the most extreme forms of incivility, such as hate speech or harassment. But in our case it is possible that some people counter-intuitively become more interested in politics by finding a politician's Facebook page precisely because they followed some uncivil controversy taking place.

To conclude, what we see as one important implication of our analysis is that it clearly suggests potential feature changes that social media platforms could implement in order to reduce the prevalence of incivility. Given that uncivil comments appear to receive more engagement, we advocate for the end of ranking systems that blindly give visibility to comments that receive more likes or comments. Instead, we argue in favor of using additional signals, such as whether the comment is considered as informative or polite, which could be developed using a similar set of text analysis methods as we use here, in order to decide which comments are displayed at the top.

In addition, based on our finding regarding the concentrated production of uncivil comments by a small minority of "trolls" and its subsequent second-order effect via contagion, we believe that user-level actions should be used by social media sites. This could be in the form of "shadow banning" (allowing people to post comments, but then hiding them from other users; a technique

already used by companies such as Reddit) or "cool-off" periods that delay uncivil comments from being visible to other users.

Appendix: Additional results

Figure 11: Prevalence of Uncivil Comments, by Topic of Legislator's Post (represented with six most likely words per topic)

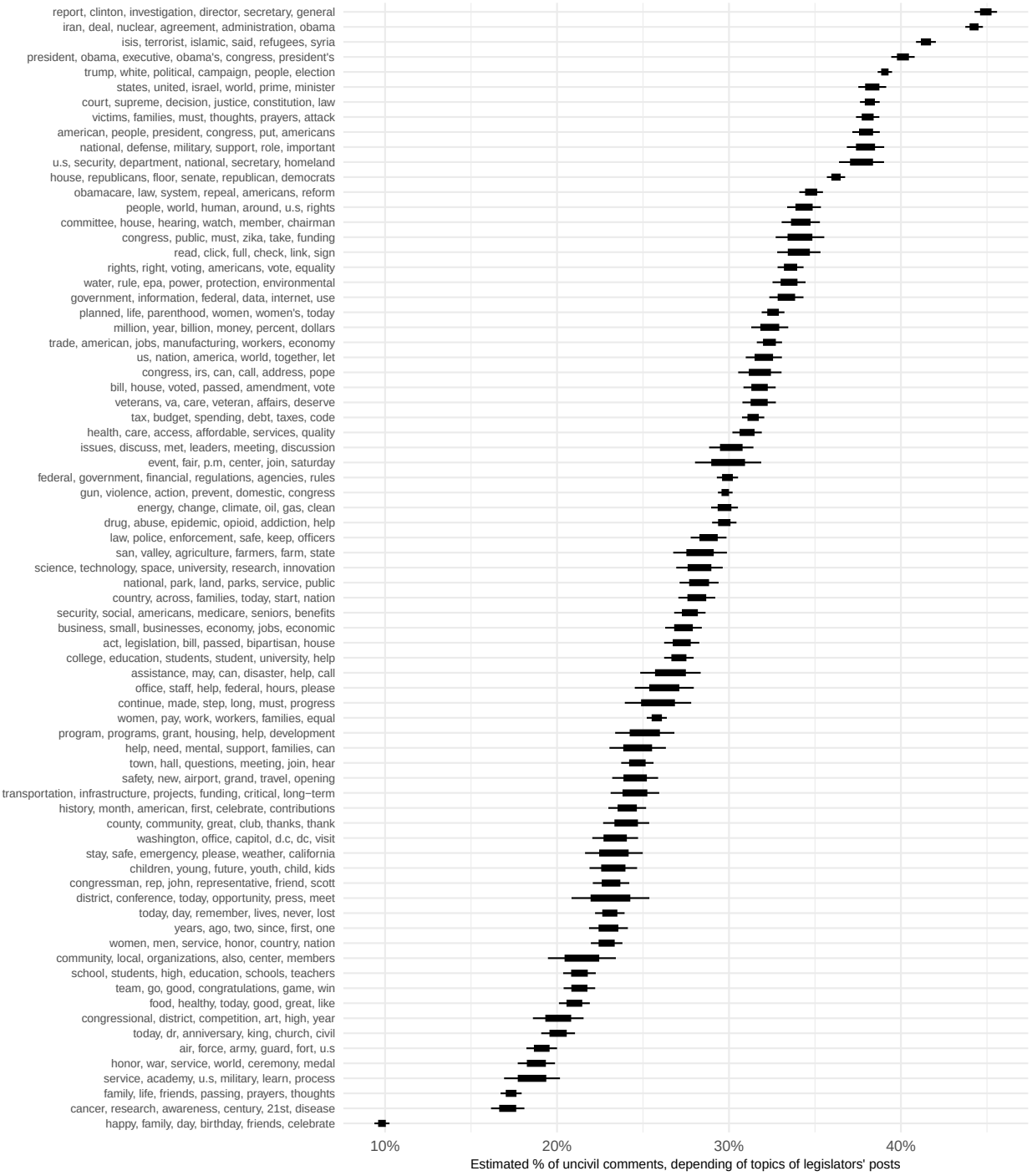
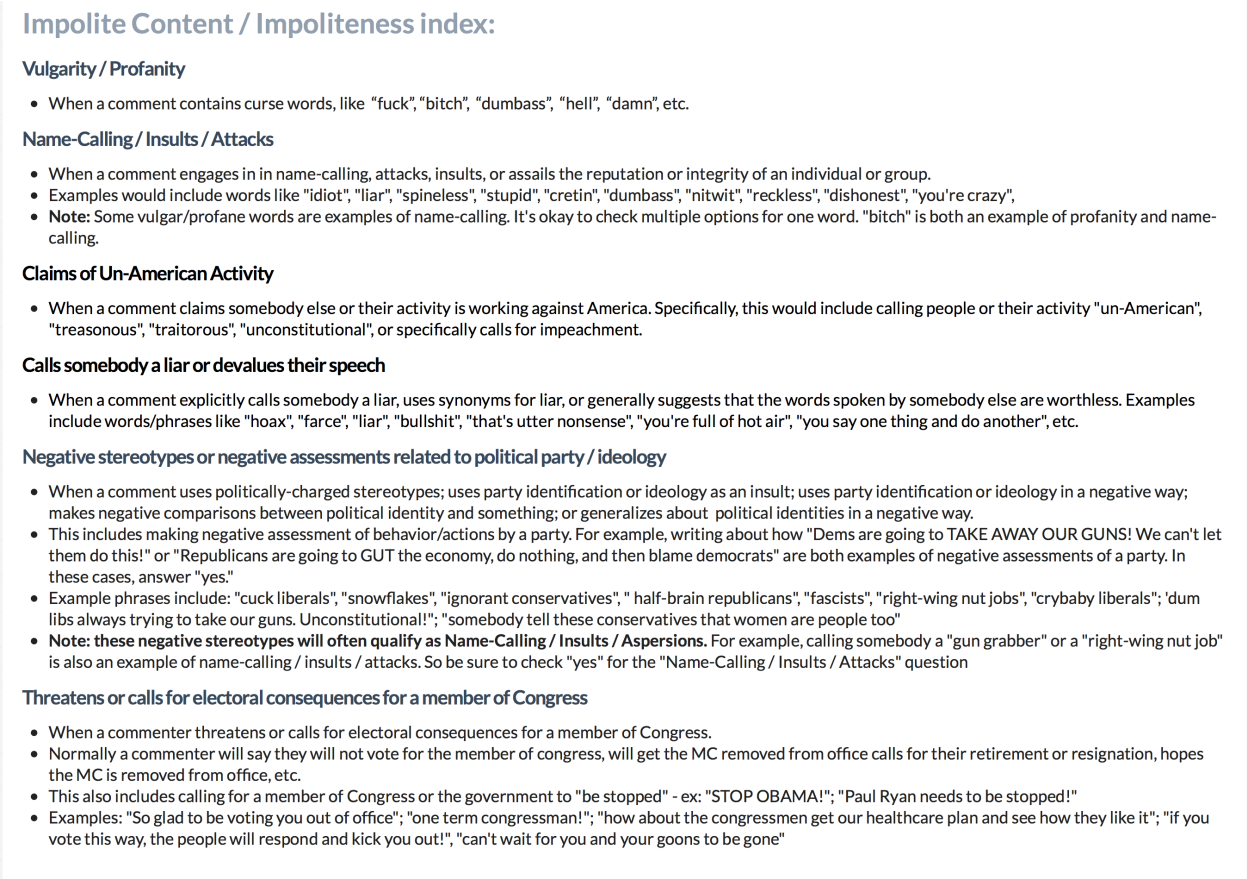


Figure 12: Combined Codebook: 6 Categories or Index Elements of Incivility



Impolite Content / Impoliteness index:

Vulgarity / Profanity

- When a comment contains curse words, like "fuck", "bitch", "dumbass", "hell", "damn", etc.

Name-Calling / Insults / Attacks

- When a comment engages in in name-calling, attacks, insults, or assails the reputation or integrity of an individual or group.
- Examples would include words like "idiot", "liar", "spineless", "stupid", "cretin", "dumbass", "nitwit", "reckless", "dishonest", "you're crazy",
- **Note:** Some vulgar/profane words are examples of name-calling. It's okay to check multiple options for one word. "bitch" is both an example of profanity and name-calling.

Claims of Un-American Activity

- When a comment claims somebody else or their activity is working against America. Specifically, this would include calling people or their activity "un-American", "treasonous", "traitorous", "unconstitutional", or specifically calls for impeachment.

Calls somebody a liar or devalues their speech

- When a comment explicitly calls somebody a liar, uses synonyms for liar, or generally suggests that the words spoken by somebody else are worthless. Examples include words/phrases like "hoax", "farce", "liar", "bullshit", "that's utter nonsense", "you're full of hot air", "you say one thing and do another", etc.

Negative stereotypes or negative assessments related to political party / ideology

- When a comment uses politically-charged stereotypes; uses party identification or ideology as an insult; uses party identification or ideology in a negative way; makes negative comparisons between political identity and something; or generalizes about political identities in a negative way.
- This includes making negative assessment of behavior/actions by a party. For example, writing about how "Dems are going to TAKE AWAY OUR GUNS! We can't let them do this!" or "Republicans are going to GUT the economy, do nothing, and then blame democrats" are both examples of negative assessments of a party. In these cases, answer "yes."
- Example phrases include: "cuck liberals", "snowflakes", "ignorant conservatives", " half-brain republicans", "fascists", "right-wing nut jobs", "crybaby liberals"; 'dum libs always trying to take our guns. Unconstitutional!"; "somebody tell these conservatives that women are people too"
- **Note: these negative stereotypes will often qualify as Name-Calling / Insults / Aspersions.** For example, calling somebody a "gun grabber" or a "right-wing nut job" is also an example of name-calling / insults / attacks. So be sure to check "yes" for the "Name-Calling / Insults / Attacks" question

Threatens or calls for electoral consequences for a member of Congress

- When a commenter threatens or calls for electoral consequences for a member of Congress.
- Normally a commenter will say they will not vote for the member of congress, will get the MC removed from office calls for their retirement or resignation, hopes the MC is removed from office, etc.
- This also includes calling for a member of Congress or the government to "be stopped" - ex: "STOP OBAMA!"; "Paul Ryan needs to be stopped!"
- Examples: "So glad to be voting you out of office"; "one term congressman!"; "how about the congressmen get our healthcare plan and see how they like it"; "if you vote this way, the people will respond and kick you out!", "can't wait for you and your goons to be gone"

References

Badjatiya, P., S. Gupta, M. Gupta, and V. Varma (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760. International World Wide Web Conferences Steering Committee.

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

Benoit, K., D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review 110*(2), 278–295.

Blom, R., S. Carpenter, B. J. Bowe, and R. Lange (2014). Frequent contributors within u.s. newspaper comment forums: An examination of their civility and information value. *American Behav-*

ioral Scientist 58(10), 1314–1328.

Borah, P. (2014). Does it matter where you read the news story? interaction of incivility and news frames in the political blogosphere. *Communication Research 41*(6), 809–827.

Brooks, D. J. and J. G. Geer (2007a). Beyond negativity: The effects of incivility on the electorate. *American Journal of Political Science 51*(1), 1–16.

Brooks, D. J. and J. G. Geer (2007b). Beyond negativity: The effects of incivility on the electorate. *American Journal of Political Science 51*(1), 1–16.

Burnap, P. and M. L. Williams (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet 7*(2), 223–242.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. *arXiv preprint*.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour 1*(11), 769.

Davidson, T., D. Warmsley, M. Macy, and I. Weber (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pp. 512–515.

Gitari, N. D., Z. Zuping, H. Damien, and J. Long (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering 10*(4), 215–230.

Grimmer, J. and B. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis 21*(3), 267–297.

Gurciullo, S. and S. Mikhaylov (2017). Detecting policy preferences and dynamics in the un general debate with neural word embeddings. *arXiv preprint arXiv:1707.03490*.

Jamieson, K. H. (1997). Civility in the house of representatives: A background report. *APPC Report 10'*.

Jamieson, K. H. (1998). *Civility in the House of Representatives: An update*, Volume https://www.annenbergpublicpolicycenter.org/Downloads/Civility/Old Annenberg Public Policy Center of the University of Pennsylvania.

Jomini, S. N., S. J. M., M. Ashley, and C. A. L. (2015). Changing deliberative norms on news organizations' facebook sites. *Journal of Computer-Mediated Communication 20*(2), 188–203.

Kevin, C., K. Kate, and R. S. A. (2014). Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication 64*(4), 658–679.

Lewis, J. B., K. Poole, H. Rosenthal, A. Boche, A. Rudkin, and L. Sonnet (2017). Voteview: Congressional roll-call votes database. Technical report, https://voteview.com/.

Lewis, J. B., K. Poole, H. Rosenthal, A. Boche, A. Rudkin, and L. Sonnet (2018). Voteview: Congressional roll-call votes database.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mutz, D. C. (2016). *In-your-face politics: The consequences of uncivil media*. Princeton University Press.

Olson, R. S., W. La Cava, Z. Mustahsan, A. Varik, and J. H. Moore (2017). Data-driven advice for applying machine learning to bioinformatics problems. *arXiv preprint arXiv:1708.05070*.

Oz, M., P. Zheng, and G. M. Chen (2017). Twitter versus facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media & Society 0*(0), 1461444817749516.

Papacharissi, Z. (2004). Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society 6*(2), 259–283.

Rosner, L., S. Winter, and N. C. Krämer (2016). Dangerous minds? effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior 58*, 461 – 470.

Rossini, P. (2019). Disentangling uncivil and intolerant discourse. *A crisis of civility*, 142–158.

Ryan, T. J. (2012). What makes us click? demonstrating incentives for angry discourse with digital-age field experiments. *The Journal of Politics 74*(4), 1138–1152.

Silva, L. A., M. Mondal, D. Correa, F. Benevenuto, and I. Weber (2016). Analyzing the targets of hate in online social media. *CoRR abs/1603.07709*.

Sobieraj, S. and J. Berry (2011). From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication 28*(1), 19–41.

Suhay, E., E. Bello-Pardo, and B. Maurer (2018). The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics 23*(1), 95–115.

Theocharis, Y., P. Barberá, Z. Fazekas, S. A. Popa, and O. Parnet (2016). A bad workman blames his tweets: The consequences of citizens' uncivil twitter use when interacting with party candidates. *Journal of Communication*.

Torcal, M. and J. R. Montero (2006). *Political disaffection in contemporary democracies: social capital, institutions and politics*. Routledge.

Waseem, Z. and D. Hovy (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pp. 88–93.