Incivility Begets Incivility? Understanding the Contagion Dynamics of Uncivil Conversations on Facebook.*

Joshua Timm[†] University of Southern California

Pablo Barberá[‡] London School of Economics

Paper prepared for the 2018 EPSA Conference

Abstract

Most social media users express frustration over the negative and bitter tone of the political conversations that take place through these platforms, which limit their potential as a space for open political deliberation. This paper seeks to explain why incivility is so prevalent on social media. Our main argument is the existence of a negative cycle of incivility: a minority of "repeat offenders" are responsible for initiating most of the uncivil comments, which makes the rest of users more likely to respond in similarly uncivil ways. We provide preliminary, descriptive evidence for the existence of this mechanism through an analysis of comments on the public Facebook pages of Members of Congress in the U.S., and by relying on a combination of supervised learning methods and dictionaries of uncivil words enriched through word embeddings. We find that (1) over 40% of comments are uncivil but also that 1% of highly active users are responsive for a majority of the uncivil content, (2) incivility receives higher user engagement, which increases its visibility given the ranking algorithms used to display content, and (3) uncivil comments are likely to receive more uncivil replies, which supports the theory of the cycle of incivility.

*We gratefully acknowledge support from a Facebook Research Faculty Grant.

[†]Joshua Timm is a PhD student in Political Science and International Relations at the University of Southern California.

[‡]Pablo Barberá (www.pablobarbera.com) is an Assistant Professor of Computational Social Science at the London School of Economics. He can be reached at P.Barbera@lse.ac.uk.

1 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. Part of this variation is likely due to disagreements over how to define incivility, and 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. And 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 the political exchanges that take place through these platforms.

With these goals in mind, we conduct an analysis of the prevalence of uncivil comments in the Facebook pages of U.S. Members of Congress. Legislators rely on these pages to connect with their constituents. 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 could have an important impact in

the poiltical process.

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 different methods to try 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 these type of content, we find that uncivil comments receive a higher number of likes and responses, which is likely to increase 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 hateful comments are more likely to elicit additional uncivil responses, creating a cycle of incivility that may make other citizens who'd be interested in participating in the discussion 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.

2 Defining incivility

Many scholars have contributed work that attempts to classify hate speech and incivility online. Classifying such language online 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, and other such terms. [1] Exemplifying the conceptual clarity problem, Silva et al. (2016, p.6) write,

[1] For sake of clarity, throughout the rest of this paper we refer to the undefined aggregate of all these terms as "negative speech."

"Hate speech lies in a complex nexus with freedom of expression, group rights, as well as concepts of dignity, liberty, and equality. For this reason, any objective definition (i.e., that can be easily implemented in a computer program) can be contested." This illustrates the high degree of subjectivity in the task of negative speech classification.

Due to the lack of definitional and operational clarity in the literature, we attempt to provide some clarity by including a discussion of existing work that attempts to classify hate speech and 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 such a task.

2.1 Existing Literature on Classifying Incivility and Hate Speech

Many scholars have attempted to conceptualize and measure different types of negative speech. Research that studies hate speech include Silva et al. (2016); Davidson et al. (2017); Burnap and Williams (2015). Some scholars attempt to classify '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 study similar topics, differences in how negative speech is defined and operationalized makes it difficult to compare results across studies and to derive general conclusions.

A key limitation of existing work is that some methodologically-driven papers lack clearly bounded, precise definitions of the concepts being studied. For instance, Badjatiya et al. (2017) classify hate speech in tweets using advanced deep learning methods using 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 measure hate speech use an index of racism and sexism that include "problematic hashtags," limits its generalizability. This example illustrates that the problem of incivility classification cannot be solved by computational methods alone, and requires more precise operationalization for accurate measurement.

2.1.1 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; Jamieson and Falk, 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 and Falk (1998) as her own definition of impoliteness, a concept she argues is distinct from incivility. While Papacharissi separates impoliteness from incivility, some scholars conceptualize incivility closer to this 'impoliteness' definition. Still, other scholars conceptualize incivility broadly, including both 'impoliteness' and a more severe definition of incivility, closer to Papacharissi's.

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 common, severe 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. We define these as severe forms of incivility because we argue their capacity to harm the public good is greater than that of common incivility. We conceptualize common incivility broadly, including 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). Finally, we view some works 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, a better goal in classifying incivility may not be to build a single perfect classifier, but to build many conceptually narrow and precise classifiers that can be applied inexpensively to all subjects within a dataset. Using such a method of multiple micro-classifiers 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 a classification scheme is, the less subjectivity will be involved in the overall classification task. We argue it is desirable to decrease subjectivity in classification tasks to both increase accuracy in computer-automated classification and to improve robustness.

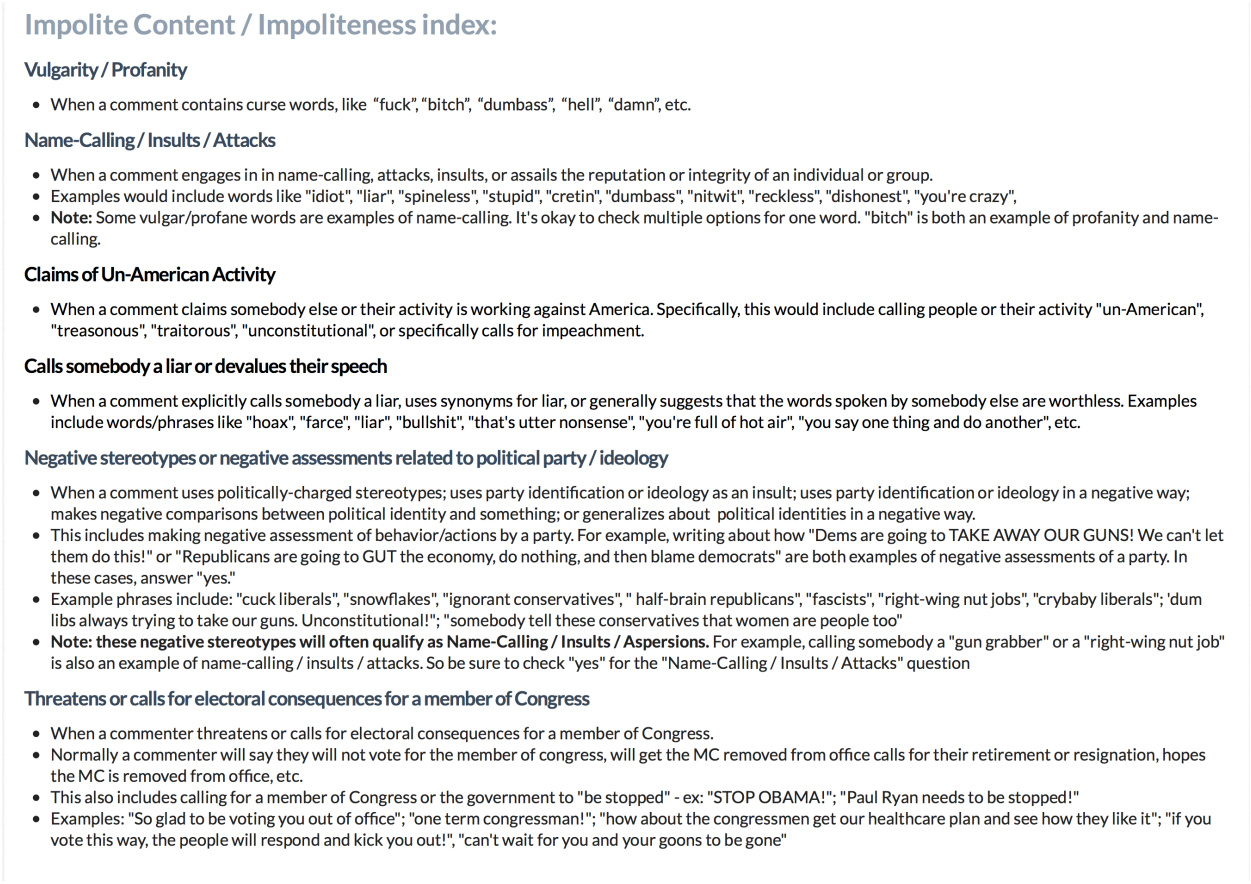
For example, if one argues that three concepts satisfy the condition of being uncivil, such as: (1) racial slurs, (2) profanity, and (3) insults, it can be determined if a piece of text contains any/all of those 3 concepts/elements. For example, the sentence, "just shut up, idiot" contains an insult, 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. And while this is not new in the literature, the difference in our approach is that we use narrower and more specific categories of incivility, as we show next.

2.2 Conceptualizing incivility using a binary index

Figure 1 depicts our binary index of incivility. Elements of the incivility index were created by both adapting theoretical conceptualizations of incivility from existing scholars' work or inductively creating elements from reading thousands of Facebook comments. The "binary" in binary index method comes in the methodology of how incivility is operationalized. In practice, we will label social media messages by asking a single yes/no question related to each of these dimensions. Messages can be classified in a variety of ways, mainly either via trained researchers or through

online services such as MTurk or other crowd-sourcing platforms. In our coding task, we include examples of each time of incivility, which contribute to our goal of coming up with narrow and specific categories to ensure our operationalization makes the construct of incivility consistent with how it is measured.

Figure 1: 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"

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 a wide variety of papers and is generally very easy to classify due to its reliance on using certain taboo words.

Index 2 (Contempt): Name-Calling / Insults / Attacks' This category comes from a combination of Jamieson (1997) and Jamieson and Falk (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 encountered accuracy problems when attempting to separate name-calling, insults, and attacks on individuals. 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.'

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 and Falk (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.

Inded 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. Facebook users publicly threatening use of their voting power is a strong example of common incivility.

One could argue that citizens logging onto Facebook 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 Facebook. 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 Facebook.[2]

2.2.1 Advantages of our coding approach

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) suggest is inevitable with this type of task. Because all conceptualizations of incivility may perhaps be legitimately

[2] While we would have liked to include hate speech in our analysis, we find that it is exceedingly rare in public Facebook comments, most likely due to Facebook's automated detection systems, which delete comments that violate their Community Standards. In our experience, examples of hate speech, racial slurs, threats to democracy, advocating violence, and other such severe types of incivility are very rare on Facebook. 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. Furthermore, the detection of hate speech presents additional complexities, namely the fact that is is a legal definition that varies across countries.

contested, we argue the best solution is to use a classification method that divides its concepts 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," we can test if any disputes with a specific index element change overall results. 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 is enhanced with index methods the more categories one uses. 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.

3 Research design

3.1 Data

In order to better understand the dynamics of incivility on social media, we gathered data on the public 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. In this paper we focus our attention on the nearly 7.5 million comments that were published on the MCs' pages during this period. Using a combination of supervised machine learning and dictionary methods, we will predict whether each of these comments can be considered uncivil, and which type of incivility each of them represents. As discussed in the introduction, we selected Members of the U.S. Congress as our population of interest given the importance of these pages as a space for civic interactions and its effect on the quality of political representation. In future iterations of this project, we plan to also include other Facebook pages, such as media outlets or candidates running for office.

3.2 Measuring incivility

3.2.1 Building a training dataset

The first step in our analysis is the construction of a labeled dataset, where each comment has been coded by human annotators along each of our six dimensions of incivility, and which we can then use both for training purposes and to evaluate the performance of each of the two classification methods we used.

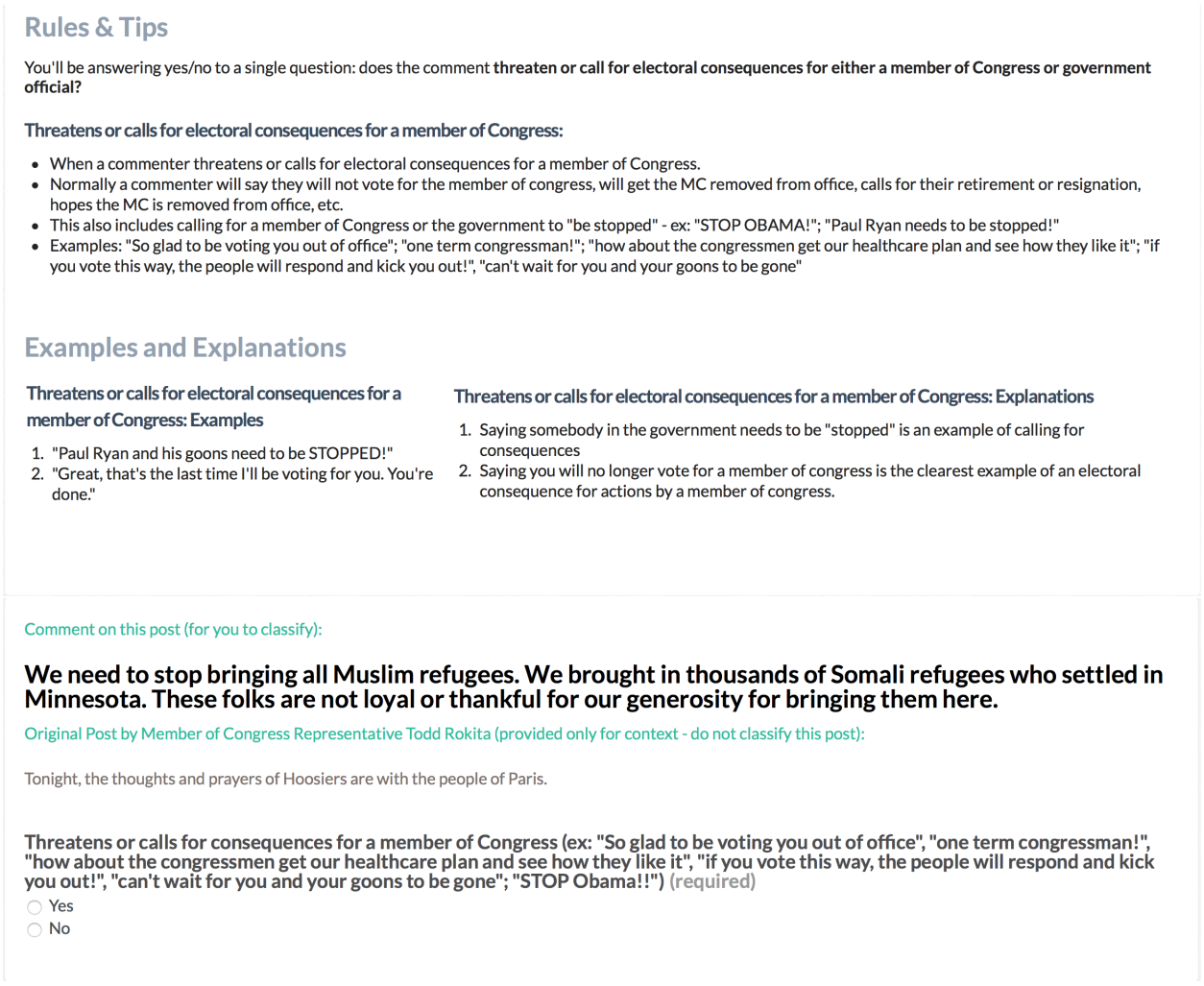
We coded a random sample of approximately 5,000 comments (stratified by MC) using Figure Eight (formerly called Crowdflower), a crowdsourcing classification platform (Benoit et al., 2016). As we described earlier, our codebook (shown in Figure 1) was developed both based on theory and inductively motivated 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 coding scheme development.

Our coding approach introduces 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 negative. 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 positive. 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 will detect which comments are negative and then we will classify them into each of our incivility categories.

(2) We built 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 categories were present, this approach breaks down the classification task

Figure 2: Preview of a Coder's Actual Classification Job for a single 1-item Text



into a series of six different tasks where only one question is being asked each time, as shown in Figure 2. 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 accuracy 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.

3.2.2 Automated text classification

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 explored a supervised learning classifier that would learn 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 tried with a dictionary method that manually tries to identify these words based on the authors' substantive knowledge and expanded using word embeddings. 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 1 reports the results of our supervised learning classifiers. While our first-level classifier (sentiment) achieves a 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 probably 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.

Table 1: 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	–	–	–
Contempt	60%	82%	**.679**	**.738**	**.707**	.757	.131	.223
Profanity	4%	94%	.100	.125	.111	**.578**	**.488**	**.529**
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**
Seditious language	2%	95%	.000	.000	–	**.111**	**.378**	**.172**

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.

Given its poor performance, we also explored an alternative method to supervised learning classification. In particular, we built our own dictionaries for each of these categories. To do so, we started with a set of seed words that based on theory we identified as being relevant 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 3 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 domain knowledge could be relevant to the latent concept being measured.

Table 1 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, given the difficulty of identifying some of these categories, we consider these results acceptable for our task, and in the rest of our analysis we will use the estimates from the dictionary methods (with the only exception of the contempt category, where we still use the classifier given that it performed better).

Figure 3: 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
```

4 Results

4.1 Who is the target of uncivil comments?

We start our descriptive analysis of incivility on Facebook pages of U.S. politicians by trying to determine the variables that predict the overall level of negative and uncivil language. Here we define as "uncivil" any comment that is predicted to fall in at least one of the six categories of incivility we use. Figure 4 breaks down these two categories across party and gender, with the vertical axis displaying the proportion of comments in each category. 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 5 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.

Figure 6 disaggregates incivility into the six dimensions we consider, and across the gender

Figure 4: Prevalence of Negative and Uncivil Comments, by Party and Gender

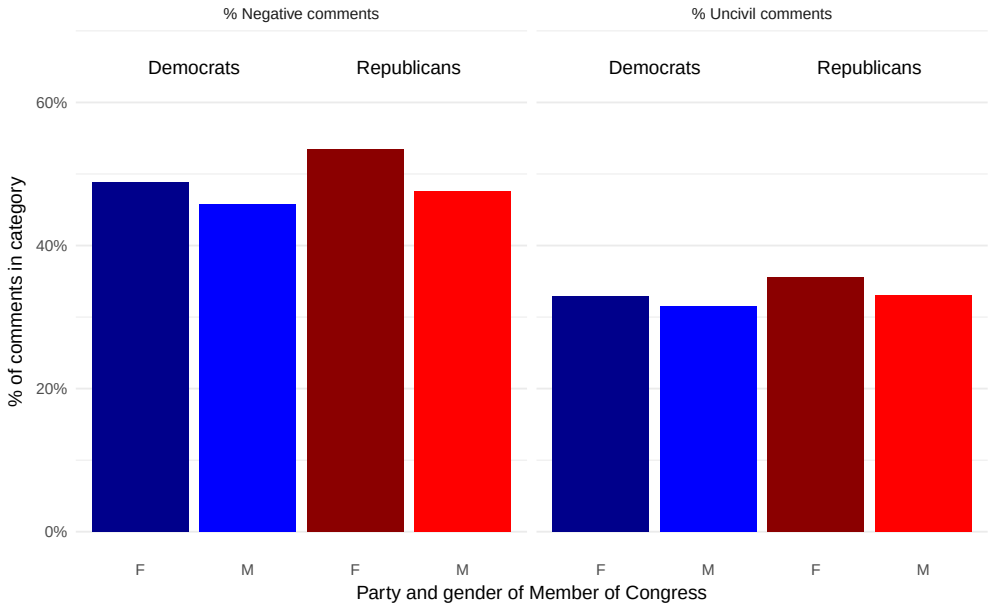
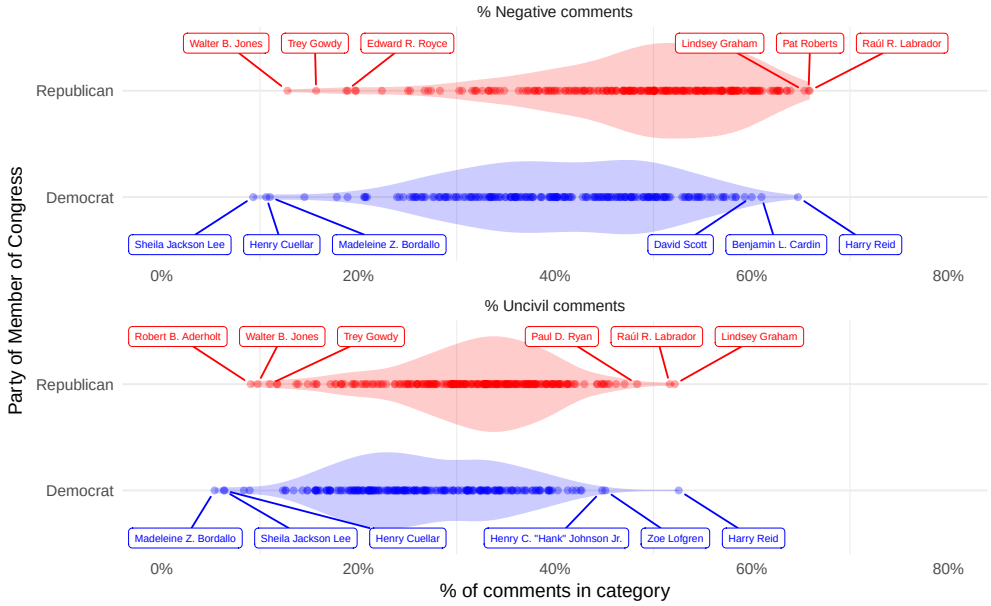


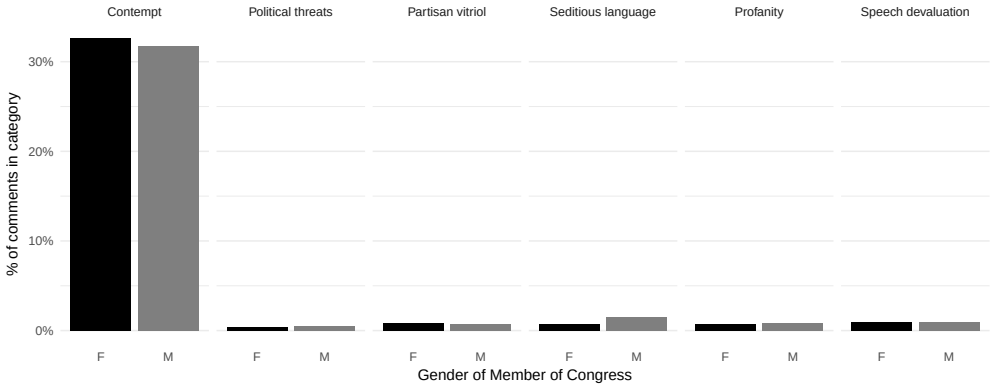
Figure 5: Distribution of Prevalence of Incivility, by Party



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

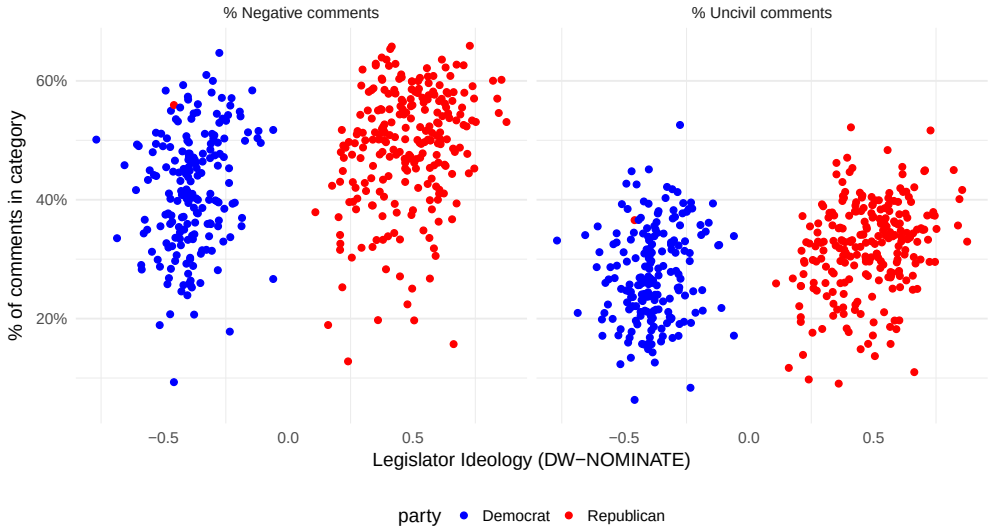
devaluation appear to be the most frequent, although without large differences across gender groups.

Figure 6: Prevalence of Uncivil Comments, by Gender and Type of Incivility



As a final type of bivariate analysis, we examine how political ideology – measured using 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 7, do not reveal any clear pattern regarding this relationship.

Figure 7: Prevalence of Uncivil Comments, by Ideology



Of course, one key limitation of these bivariate analysis is that our results could be driven by confounding variables (e.g. gender and party ID are highly correlated, since most female leg-

islators are Democrats). To better understand who is the target of incivility, we also estimate multivariate linear regressions at the legislator level (using weights based on comment count), where 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 2: WLS regression of % posts in category

	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

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

We show our results in Table 2. We find three clear patterns. First, Senators 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 "received" 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.

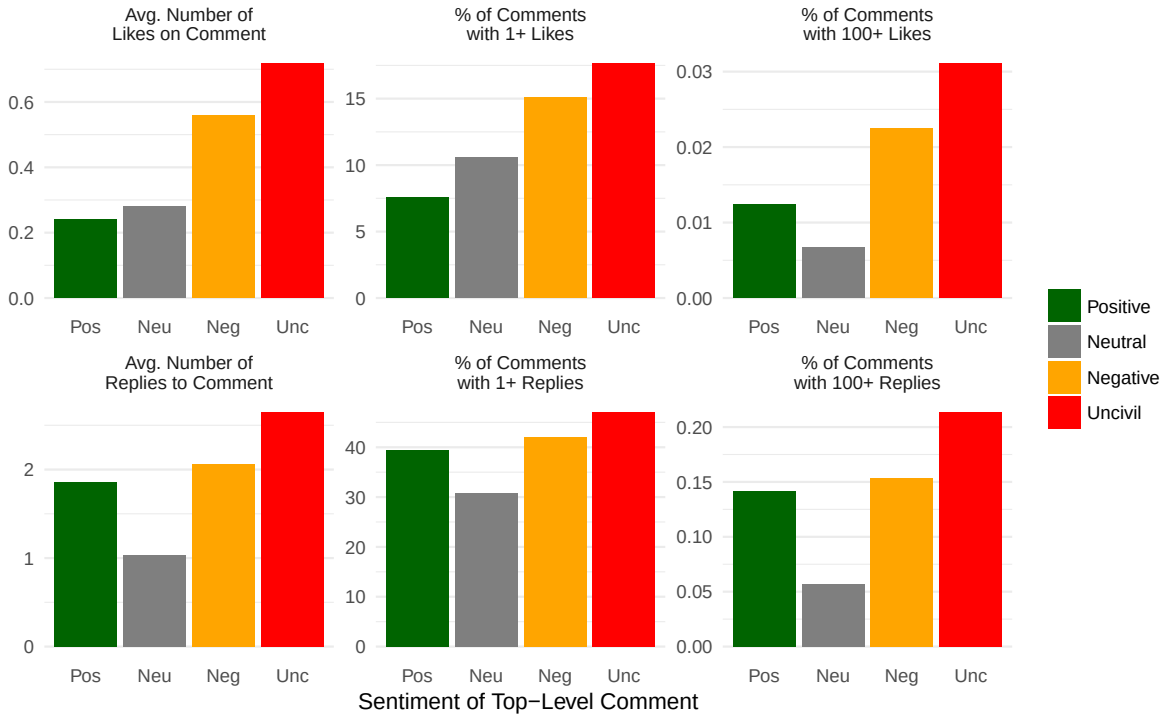
4.2 Uncivil comments receive higher engagement

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 in 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 exposure.

To test this possibility, we collected data on the number of likes and replies that each comment received. As we show in Figure 9, 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. Figure 8 displays three different metrics of user engagement depending on the characteristics of each comment (positive, neutral, negative+civil, and negative+uncivil). 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).

Note that the estimated level of engagement with uncivil comments is likely to be an underestimate, since not all Facebook users response to a comment via the reply button. Some people comment in reply to another comment, but write their message in the general thread, usually writing the name of the person or tagging them. Figure 9 below illustrates this point. These sort of

Figure 8: Uncivil comments receive higher engagement



general-thread responses anecdotally appear to be fairly common from our observations of Facebook comment threads.

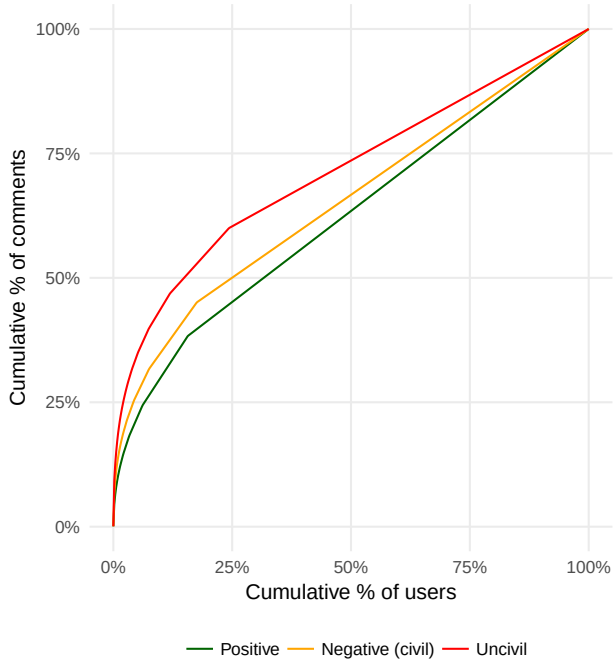
Figure 9: Example of different types of Facebook comment responses

4.3 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 IDs 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 10 displays our attempt at answering this question. Here, each curve shows the cumulative distribution of the proportion of users sending a proportion of messages within each category – positive, negative + civil, and negative + 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, the degree to which this type of messages is concentrated within this minority is larger than the equivalent number for positive and negative (but civil) comments.

Figure 10: Lorenz curve showing amount of users responsible for types of speech

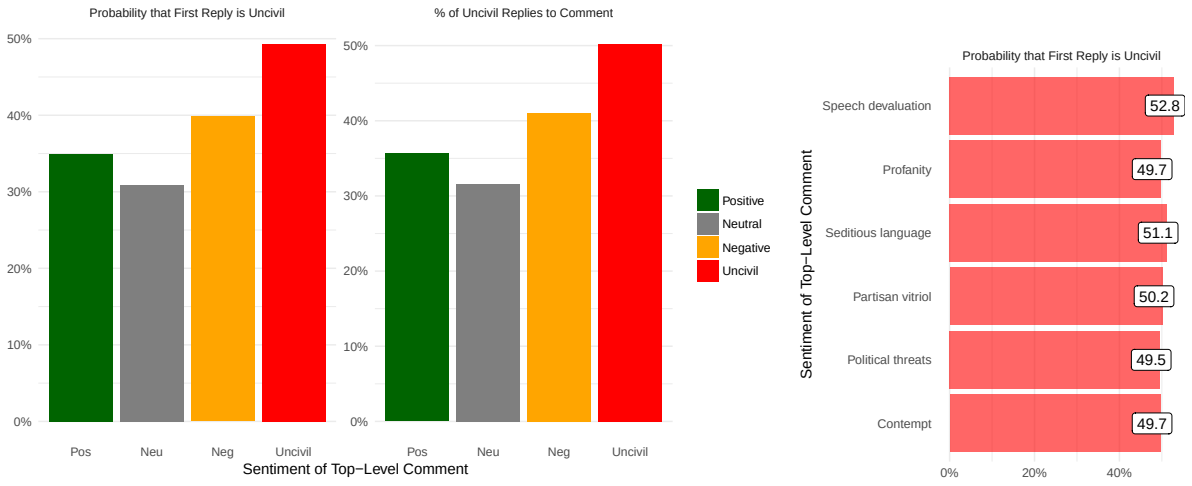


4.4 The cycle of incivility

The final step in our descriptive analysis consists on exploring patterns of diffusion of incivility. Our main interest here is to understand whether "incivility begets incivility," that is, if just a few hateful messages are enough to lower the quality of political deliberation that could take place within Facebook pages. To explore this hypothesis, 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 11 displays the results of our analysis. We find that uncivil comments are between 25% and 4% more likely to receive uncivil replies, either as the first response to the comment or as the overall volume of uncivil replies in the entire thread. 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 seditious language and speech devaluation are the two types of incivility that appear to be more "contagious."

Figure 11: Uncivil comments receive higher engagement



5 Discussion and Conclusion

The overarching goal of this project is to measure the prevalence of incivility on social media, to understand how it spreads, and what impact it may have on citizens' willingness to engage in meaningful political discussions and th overall quality of democratic politics. As a preliminary step in this direction, here we focused on developing a more nuanced measure of incivility on social media, and on offering 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, which in turn creates incentives for people to post more content of this type.

The next steps for this project are increasing the accuracy of the machine learning classifier by increasing the number of positive labels of other types of incivility, not only comments that attack or insult an individual or group. We also plan to use topic models in combination with matching techniques to understand whether incivility focuses on content related to specific policy issues, and to try to offer causally identified evidence for the cycle of incivility after controlling for topic.

We also acknowledge limitations in our approach, which we plan to address in future iterations. First, sometimes it is unclear whether a message is uncivil or if it advocates for a valid policy position. Is advocating for a policy that bars muslim refugees from entering the US anti-muslim and thus uncivil, or is it a legitimate policy position based on concern for national security? The task of classifying these types of issues will always carry degrees of subjectivity, and consequently, lack of absolute confidence in classification. Because of these problems, we limited the scope of this project to measuring common incivility, but over time we will also develop solutions to measure severe incivility.

And second, it is also important to note that some aspects of incivility may have positive effects from a normative perspective. For example, Brooks and Geer (2007) 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. Similarly, in our case it is possible that some people counter-intuitively become more interested in politics if they reached a page by a politicians precisely because they saw there some uncivil controversy taking place.

References

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

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

Blom, R., Carpenter, S., Bowe, B. J., and Lange, R. (2014). Frequent contributors within u.s. newspaper comment forums: An examination of their civility and information value. *American Behavioral 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 Geer, J. G. (2007). Beyond negativity: The effects of incivility on the electorate. *American Journal of Political Science*, 51(1):1–16.

Burnap, P. and Williams, M. L. (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 Guestrin, C. (2016). Xgboost: A scalable tree boosting system.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (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, pages 512–515.

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

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

Jamieson, K. (1997). Civility in the house of representatives. *APPC Report 10'*, https://www.annenbergpublicpolicycenter.org/civility-in-the-house-of-representatives/.

Jamieson, K. and Falk, E. (1998). Civility in the house of representatives: An update. *APPC Report 20'*, https://www.annenbergpublicpolicycenter.org/Downloads/Civility/Old

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

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

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

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

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

Oz, M., Zheng, P., and Chen, G. M. (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., Winter, S., and Krämer, N. C. (2016). Dangerous minds? effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior*, 58:461 – 470.

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

Sobieraj, S. and Berry, J. (2011). From incivility to outrage: Political discourse in blogs, talk radio,

and cable news. *Political Communication*, 28(1):19–41.

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

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

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