Detecting Anti-Social Norms in Large-Scale Online Discussions

Yotam Shmargad, Associate Professor, School of Government & Public Policy, University of Arizona
Stephen A. Rains, Professor, Department of Communication, University of Arizona
Kevin Coe, Professor, Department of Communication, University of Utah
Kate Kenski, Professor, Department of Communication, University of Arizona
Steven Bethard, Associate Professor, School of Information, University of Arizona

Abstract

This chapter develops a theoretical framework for investigating social norms in online discussions and applies it to study anti-social commenting in two datasets ten years apart. Our thesis is that anti-social commenting is promoted through at least two different social processes. First, discussion contributors mimic one another, deploying anti-social comments after other contributors have done so. Second, contributors are responsive to ”votes” of approval that they have received for prior instances of anti-social commenting. We argue that these two processes map onto a distinction made in the literature on social norms between descriptive and injunctive norms and investigate both processes at the individual and collective levels. We compare human annotations of anti-social commenting with several automated classifiers and provide evidence that some classifiers are well-suited for understanding the norms associated with anti-social online commenting. Our framework can be applied to online discussions at scale and makes use of both the relational and temporal aspects of the digital trace data that are generated when people use the web and social media.

Keywords: Automated Classification, Content Analysis, Discussion Threads, Likes, News Comments, Reddit, Social Norms, Twitter, Up and Down Votes

Detecting Anti-Social Norms in Large-Scale Online Discussions

Online discussions are fundamentally social. Participants are engaged audiences of the messages that other contributors post and are concurrently exposed to the social approval and disapproval (as seen in "votes") that those (and their own) messages receive. These two sources of social information are deeply intertwined—messages receive votes based on their contents (Rains et al., 2017), while votes shape the content that future messages contain (Shmargad et al., 2022). Posts and votes differ, however, in the kind of information that they tend to communicate. Information within a post often reveals the *descriptive norms* of a discussion setting, or signals of what other people do, while votes reflect the *injunctive norms*, or signals of what people ought to do (Cialdini et al., 1990). Because people rely on descriptive and injunctive information to guide their behavior (Rimal & Real, 2005), online discussion data are uniquely suited for the study of social norm formation and evolution, compliance and deviance. Historical records of online discussions are often readily available, enabling a better understanding of online socialization processes at both the collective level (e.g., by tracking aggregate trends in posting and rating behaviors) and individual level (e.g., by analyzing a person's posting and rating behavior over time). Lapinski and Rimal (2005) label these two analytical levels *collective* and *perceived* social norms, respectively.

Using the lens of social norms can help to shed light on the various forms of anti-social commenting that are prevalent online, including incivility (Coe et al., 2014), trolling (Cheng et al., 2017), and online hate speech (ElSherief et al., 2018). These forms of commenting have been of increasing concern, with nearly a third of adults (and over half of those between the ages of 18-29) reporting they have been called an offensive name online (Vogels, 2021). This increase in public concern has been met with an increase in research into the topic, with much of it focusing on the automated detection of anti-social commenting (Tontodimamma et al., 2021). One project that offers free access to several automated classifiers is called Perspective which originates from Google's Jigsaw lab (Lees et al., 2022). We authors have built an additional classifier that identifies name-calling specifically (and it is available to others via the platform Hugging Face[1] (Ozler et al., 2020; Sadeque et al., 2019). Automated classifiers make it possible to detect text features such as name-calling at scale, and to discover variations in their deployment across time and individuals (e.g., Rains et al., 2021; Rains et al., 2023a; Rains et al., 2023b).

This chapter is about how the internet shapes social processes that foster the expression of incivility, hate, and other anti-social language on the one hand and how scientists study these processes on the other. We make a modest contribution to knowledge by analyzing anti-social commenting by both human-coded annotations and automated classification techniques in order to identify the extent that automated classifiers are suitable for the large-scale study of online (anti-)social norms. To do so, we replicated results from previous work that relied on human coded annotations (Rains et al., 2017; Shmargad et al., 2022), here using automated classifiers instead of human coders. Our findings are mixed, with some classifiers more reliably replicating prior work than others. In general, however, there was a meaningful overlap between using human annotations and automated classifiers. To show how our framework can be applied to contemporary online discussions, we applied automated classifiers to discussions on Reddit and Twitter during the January 6th Capitol riots, and we compared the norms surrounding anti-social commenting across these two platforms. Using these multiple approaches, several important points became clear. First, that anti-social commenting is promoted through at least two social processes, one in which people mimic others and another in which people who get rewarded for anti-social messages subsequently generate more of them. Second, that automated classification is a measurement advancement that can aid in the study of both the collective and perceived norms surrounding anti-social commenting.

Social Norms in Online Communities

Theorizing about deviance from societal normative boundaries goes at least as far back as Durkheim's (1893) work on the topic—for more informal examples, you can go back as far as *The Epic of Gilgamesh* or Shakespeare for relevant discussions. Durkheim argued that deviance is a necessary, even beneficial, part of society because it clarifies the norms (thereby encouraging compliance), strengthens bonds among those reacting to deviance, and can lead to positive social change by challenging people's existing views. Lofland (1969) clarifies the process by which deviance is socially constructed, becomes ingrained in a society's view of itself, and creates a dual process whereby deviants increasingly associate deviance with their own identity. Maratea and Kavanaugh (2012) update these classic sociological ideas to provide a modern understanding of online deviance, specifically. They point out that emerging information and communication technologies allow "scholars to study deviant subcultures that did not exist, or were too hidden to access, prior the advent of the internet" (p. 107). As a site of contemporary deviance, online discussion threads thus represent a promising venue for investigating anti-normative behavior.

While incivility and other forms of anti-social commenting are sometimes conceptualized as deviance from socially accepted manners of speech (Jamieson et al., 2017), one would be hard pressed to find societal benefits for some of the speech that can be found online. And yet, in addition to the (sometimes circular) arguments about one's rights to freedom of speech, unsavory online language cannot be uniformly treated as "bad" as it can serve positive, even necessary, societal and democratic functions (e.g., Edyvane, 2020; Rossini, 2022). Such considerations, ethical in nature, are central to recent debates about how digital platforms should be moderating, and whether and when they should be censoring, information that circulates on the web (Forestal, 2021). Proposed solutions, such as focusing only on the most extreme forms of commenting such as hate speech (Jiang et al., 2020) or paying special attention to the targets of such speech (Zampieri et al., 2019), showcase how politically fraught these considerations can become – particularly when automated methods are employed (Udupa et al., 2023). For example, Haimson et al. (2021) find that the people most likely to report that their posts were removed from online platforms were either ideologically conservative, transgender, or Black. The reasons provided for removal, however, varied substantially across these groups, with the latter two groups having more of their posts removed that either did not violate platform policies or fell under moderation gray areas.

Despite clarifying ethical nuances surrounding censorship of anti-social comments and the groups that engage in or are targeted by such language, discussions of content moderation often ignore the underlying social processes that culminate in specific communication patterns. The work that exists often treats anti-social commenting as either a contagious process (Song et al., 2022), one in which particular people are drawn into contentious discussion (Bor & Peterson, 2021), or both (Kim et al., 2021). While their work starts to unpack the mechanisms behind anti-social behavior, they still leave much to ponder. Is such language always contagious or are there social contexts that are more conducive to memetics? If specific people are primarily responsible for its spread, how does one become (or learn to respond to) such a person? These questions suggest a focus on social processes as critical forces affecting the expression of anti-social content. It is important to focus on both the relational drivers of behavior as well as a longer time window through which to witness the socialization of anti-social expressions take place. The very platforms often blamed for fueling anti-social behavior also provide such a relational and temporal view of human interaction and development.

Most traces of behavior generated by everyday online activity are intrinsically *relational* (Golder & Macy, 2014). For example, on Twitter alone, people share or *retweet* messages that others post, show approval of a post with a *favorite*, respond to a post with a *reply*, inform other users of their message with a *mention*, or *quote* a post by sharing it with additional commentary. On platforms such as Reddit and YouTube, people provide comments and videos, respectively, with *Up* and *Down votes* to signal their approval and disapproval. While platforms may provide various ways for people to register their feedback about other users' posts, we refer in this chapter to clicks of approval or disapproval as *votes*. Votes can be said to be relational because they have meaning not just for voters but also for receivers of the vote, for audience members who can view aggregated statistics of votes, and for platforms that use votes to filter posts in and out of people's content streams via algorithms (Burrell & Fourcade, 2021). From the perspective of a researcher studying online behavior, votes are understood differently depending on whether the sender or receiver of the vote is the focus. For example, retweeting a message can be interpreted as a signal of homophily, or similarity, with the person who posted the message (Barbera, 2015). Receiving retweets, on the other hand, can imply that a message has resonated (McDonnell et al., 2017) and was influential in getting people to pay attention to the poster (Shmargad, 2022).

In addition to reflecting the relational aspects of human behavior, digital trace data are also inherently *temporal* in nature. For example, timestamps typically accompany post data that are collected from social media platforms, so that it is possible to construct sequential timelines among comments. For the study of anti-social commenting, the temporal nature of digital trace data can be used to understand macroscopic trends (Rains et al., 2021) on the one hand and microscopic dynamics (Shmargad et al., 2022) on the other. Rains et al. (2021) used a dataset of Russian troll tweets (Linvill & Warren, 2020) to study their use of anti-social commenting across different periods of the 2016 U.S. presidential election cycle. Shmargad et al. (2022) used the temporal nature of digital trace data to understand the dynamics of anti-social commenting as a discussion thread evolves. They found that anti-sociality is more likely after prior anti-sociality by other commenters as well as votes of approval for one's own anti-social commenting. These two applications of temporality—analyzing macroscopic trends and microscopic dynamics—can aid researchers in the study of the collective and perceived norms (Lapinski & Rimal, 2005) surrounding anti-social commenting, respectively.

We build here on our prior work (Shmargad et al., 2022) to argue that the tone of online contributors' posts at a particular point in time depends in large part on two factors: the tone of other users' previous posts and the Up or Down votes that they have seen posts receiving. These factors reflect the descriptive and injunctive norms, respectively, that users perceive. Each of these norms, in turn, have two further components: They can be either *self-* or *other-focused* (see Figure 1). One's *self-focused descriptive norms* are their perceptions of contributions they themselves have made in the past, while their *self-focused injunctive norms* are their perceptions of how those contributions were rewarded or penalized. One's *other-focused descriptive norms* are perceptions of the contributions others have made, while *other-focused injunctive norms* are perceptions of how those contributions were rewarded or penalized (either by the focal user or by others).

<Figure 1>

Figure 1 depicts in dotted lines the *immediate signals* that will inform an individual's posting behavior at a particular point in time. These are not the only signals that will matter but represent what a first attempt at a test of our framework might look like. The total of the votes, or clicks of approval and disapproval, reflect the aggregated feedback of other users. Because votes provide the person posting with feedback about what is and is not appropriate (i.e., self-focused injunctive norms), we might expect that a person will continue to post in ways that provide positive feedback and stop posting in ways that yield negative feedback. However, this is not always the case; for example, Cheng et al. (2014) show that negative feedback can backfire and increase the likelihood of future posts that also receive negative feedback. As such, the role that votes play in discouraging or encouraging particular behaviors, such as anti-social commenting, is an empirical question that our framework can help to address.

The extent that language in other people's posts will shape the language that is chosen by a subsequent user depends on various factors. Goldberg and Stein (2018) argue that the spread of information relies on the mental frames of the receiver, a process they call *associative diffusion*. In-group status (Rimal & Real, 2005), social tie strength (Bakshy et al., 2012), online anonymity (Kim et al., 2019), identity performance (Freelon et al., 2020), political influence (Shmargad, 2022), elite status (Rains et al., 2023b), and many other factors will likely shape how contagious a person's language might be. The content in another commenter's posts can be viewed as contributing to a descriptive norm because it represents what another user is doing. A subset of posts may also include injunctive information, however, and the large-scale extraction of such information is a worthwhile direction for future research. Moreover, the full set of posts that constitute the descriptive norms guiding a person's posting behavior at a specific point in time could be large and varied, and identifying the bounds of this evolving set is also an important research direction.

While votes provide people with quantitative measures of feedback, the text contained within a post is "unstructured data" and meaning must be extracted from posts by discussion participants and behavioral researchers alike. Automated ways for extracting meaning from text have increased both in number and ease of use, with the most recent advancements evident in the large language models (LLMs) that may represent the future of data annotation (Ding et al., 2022). The primary advantage of automated classification techniques is that one can process large amounts of text quickly and cheaply—which, for the study of socialization, implies that a broader set of people, discussions, and contexts can be studied. However, to be influenced by the behavior of another requires that discussant participants are able to extract information from another person's post (e.g., the presence of anti-social commenting) and use that information in their own posting decisions. For example, if a person does not pick up on a particular name-call (e.g., the capitalization of the R in democRat; see Sadeque et al., 2019), they may not take it as an insult and may thus not return in kind. To test the extent that automated techniques pick up on human interpretations, we compare human coded annotations and automated classification techniques to study the collective and perceived norms surrounding anti-social commenting. This allows us to both evaluate the framework in Figure 1 and to validate the use of automated classification techniques in the detection of anti-social normative behavior.

Comparing Human Annotation to Automated Classification of Anti-Social Commenting

The data that we discuss in this section are similar to those analyzed in Coe et al. (2014), Rains et al. (2017), and Shmargad et al. (2022), and include all of the online comments made on news articles published in the online website for the *Arizona Daily Star* (*ADS*) over a 3-week period in October and November of 2011. These news articles, along with their comments, were printed as PDF files and the text of the comments was then manually annotated by human coders for five different measures of incivility (*name-calling*, *aspersion*, *accusations of lying*, *vulgarity*, and *pejorative speech*). The annotation process was designed to increase inter-coder reliability. It began by having trained coders independently annotate the same set of comments, then discuss disagreements in their annotations, and finally to update a codebook that further clarified how annotations were to be made. When the coders reached sufficient intercoder reliability in their annotations, they then independently coded the actual comments on the *ADS* articles. Further details about the methodology, including intercoder reliability scores for the different measures of incivility and specific examples of each measure, can be found in Coe et al. (2014).

An independent research effort (Sadeque et al., 2019) developed an automated classifier using the coded data for name-calling, specifically, the application of which required the extraction of text from the comments in the original *ADS* discussion files. We use those data, which include the text of the comments in addition to information in the original dataset (e.g., comment numbers, counts of the down and up votes that comments received, etc.). We thus had the human annotations of incivility alongside the comment text, the latter of which was prepared for automated classification. We applied our own classifier for name-calling to the comment text, which was built by Ozler et al. (2020) and trained on several annotated datasets in addition to the *ADS* comments. We also processed the text in the comments using Google's Perspective API (Lees et al., 2022) to obtain scores for several additional attributes, including *toxicity*, *severe toxicity*, *identity attack*, *insult*, *profanity*, and *threat*. The Perspective API also provided text attributes that were trained on a set of comments from *New York Times* articles, and we used these as well given their similarity to our data.[2] These attributes include *attack on author*, *attack on commenter*, and *inflammatory language*. The various classifiers provided different measures of anti-social commenting, each of which was a possible candidate for the detection of normative behavior (i.e., for evidence of mimetics and/or response to social votes). Automated classifiers each provided scores between 0 and 1, and we converted these scores into binary variables using a threshold of .5.

To test the framework from the previous section (Figure 1), we adopted a data preparation strategy first outlined in Shmargad et al. (2022) to transform the comment data into triplets such that: (1) the three comments in a triplet were on the same news article, (2) the second comment in a triplet followed immediately after the first, (3) the first and second comments were authored by different contributors, and (4) the first and third comments were authored by the same contributor. This empirical strategy let us test how three factors highlighted in Figure 1 above—a person's prior comment, the votes that person's comment received, and another person's comment—contribute to the nature of a person's subsequent comment. Figure 2 depicts the empirical strategy that we adopted as a test of the framework in Figure 1. The votes that Person i received on their initial comment captures self-focused injunctive norms, while the presence of anti-social commenting in Person i's and j's comments capture self- and other-focused descriptive norms, respectively. Because this empirical strategy allowed us to investigate the (possibly interactive) effects of descriptive and injunctive norms, it could also be construed as a partial test of Rimal and Real's (2005) theory of normative social behavior (TNSB) according to Shmargad et al. (2022). (We leave the study of other-focused injunctive norms for future research.)

<Figure 2>

Collective descriptive norms in the Arizona Daily Star

Shmargad et al. (2022) used a single, collapsed measure of anti-social commenting, which was set to 1 for comments that included *name-calling*, *aspersion*, *accusations of lying*, *vulgarity*, or *pejorative speech*, and 0 for comments that lacked those message features. The classifier in Ozler et al. (2020) was trained on the *name-calling* annotations only, so that in addition to Shmargad et al.'s (2022) collapsed measure we also defined two new measures—”*name-calling*” tracked the presence of name-calling, specifically, while “*not name-calling*” tracked the presence of any of the other incivility measures. Table 1 includes summary statistics for these three measures as well as those obtained from the Perspective API and *New York Times* attributes. These statistics capture the *collective* descriptive norms surrounding anti-social commenting in this community. We include statistics for all comments, as well as for the subset of comments that started a triplet. The latter will serve as a useful comparison in the next section, where only comments that were included as parts of a triplet were classified for the presence of anti-social commenting. Table 2 includes definitions and examples of the various forms of anti-social commenting that we classified using automated methods.

<Table 1>

<Table 2>

Several observations in Table 1 are worth discussing. First, the rate of anti-sociality in comments that started a triplet did not deviate much from the rate across all comments. Since the focus later will be on triplets to provide a test of our theoretical framework, it is reassuring that these triplets did not begin in an atypically pro- or anti-social manner.[3] Second, the name-calling classifier successfully recovered the rate of name-calling that was coded by human coders (i.e., the means were identical or close). This result highlights the utility of classifiers that are trained on properly annotated data as a means of evaluating name-calling behavior at scale. Third, the scores obtained from the Perspective API measures were low for these data, with *severe toxicity* not appearing in any of the comments. This finding raises questions about the sensitivity of the Perspective measures. We thus also constructed a collapsed measure across the six Perspective API measures, which is 1 if either *toxicity, severe toxicity, identity attack, insult, profanity,* or *threat* were present, and 0 otherwise. Using the collapsed Perspective API measure, only 6% of the comments included at least one of these forms of anti-social commenting, compared to 20% that included at least one of the incivility measures coded by Coe et al. (2014). Although the types of behaviors captured in the Perspective API appear broader, they are less sensitive to common forms of incivility. Finally, the rates of the *New York Times* attributes were higher than measures obtained with the Perspective API, with a third of the comments containing instances of *inflammatory language*. Figure 3 presents the correlation matrix of these measures of anti-social commenting.

<Figure 3>

Collective injunctive norms in the Arizona Daily Star

As previously mentioned, in addition to measures of incivility the *ADS* dataset includes counts of the down and up votes that comments received from other readers. Figure 4 presents a plot of the relationship between the down and up votes that comments received. We standardize the number of down and up votes for each article by subtracting the article-specific mean and dividing by the article-specific standard deviation. This controls for correlations across down and up votes that result from some articles simply drawing more attention. We then transform the standardized down and up vote measures by taking a logarithm to remove skew. As we can clearly see in Figure 4, down and up votes were highly correlated even after controlling for the specific news article, suggesting that comments frequently divided the community (i.e., a given comment received a proportional number of down and up votes). These rankings thus reflect in- and out-group dynamics in a broad sense, some of which can be explained by partisanship (Rains et al., 2017). Papakyriakopoulos et al. (2023) show that discussions on forums with both up and down vote capabilities are less civic-natured than forums that only allow up votes (though, notably, more civic-natured than forums that allowed neither up nor down votes).

<Figure 4>

The next set of analyses relate the presence of anti-social commenting to the number of down and up votes that comments received, which reflect *collective* injunctive norms around anti-social commenting in this community. Table 3 reports results for the various measures of anti-social commenting. Each coefficient was obtained with a multilevel model that included random effects for the news article and for the contributor who posted the comment. The random effects were included to remove variation from specific articles that drew more anti-social commenting or from people who were more likely to deploy anti-social commenting. We also included an additional control variable, the numerical order of the comment in the set for that article. As reported in Rains et al. (2017), comments that included incivility were more likely to draw both down votes and up votes. Interestingly, *name-calling* was primarily responsible for the increases in up votes while the other incivility measures were responsible for increases in down votes. Among the Perspective API measures, only *insult* showed a positive relationship with up votes. *Attacks on commenters* were associated with fewer up votes, while *inflammatory language* was associated with more down votes. Figure 5 portrays these effects.

<Table 3>

<Figure 5>

Perceived descriptive and injunctive norms in the Arizona Daily Star

The final set of analyses in this section replicated Shmargad et al.'s (2022) on the effects of *perceived* descriptive and injunctive norms on the spread of incivility (Figure 2). The purpose of these analyses was to examine how community responses to incivility influence the degree to which it is perpetuated in online discussion. We used multilevel models with random effects for article and commenter and included control variables for the numerical order of the comment as well as the "gap" (i.e., number of comments) between the first and last comment in a triplet. Note that the triplets, as constructed, allow for several comments to occur between the first and last comment, so far as none of these comments were contributed by the author of the first and last comment. We did not run the model for the six Perspective API measures separately (*toxicity, severe toxicity, identity attack, insult, profanity,* or *threat*) as they were not prevalent enough. Instead, we report results for a collapsed measure that tracked if any of these six Perspective API attributes were present (0 no, 1 yes). We do not report numerical estimates here, but instead provide images of the marginal effects in Figures 6 and 7 below.

Figure 6 provides estimates for the human coded measures of incivility as well as for the automated classifier for *name-calling*. The right panels depict the effects of down votes while the left panels depict the effects of up votes. The first row in Figure 6 replicates the results from Shmargad et al. (2022): The effect of incivility in another user's comment depends, in part, on the number of up votes that a user's initial comment received. When the user was initially uncivil and received no up votes, the presence of incivility in another user's comment decreased the likelihood of incivility in the initial user's subsequent comment. However, as the number of up votes a user received increased, the effect of another commenter's incivility also increased. This implies that incivility was met with incivility when it was initially rewarded. Down votes, on the other hand, did not influence the effect of another commenter's incivility.

<Figures 6 & 7>

When we break up the collapsed human-annotated incivility measure into two categories, *name-calling* and all other measures (*not name-calling*), the effect of proximate incivility increased as the number of up votes increased for other incivility measures but not for *name-calling* itself. When the same analyses were conducted using the automated *name-calling* classifier (Ozler et al., 2020) rather than the human-annotated measures, results matched. The results suggest that the effect of other users' name-calling on one's own name-calling was not shaped by how many up votes the initial name-calling received. This suggests that different forms of incivility may elicit different normative responses, with name-calling showing less sensitivity to mimetics and social rewards. The results (as also seen in Shmargad et al., 2022) replicate with the collapsed Perspective API measures (the presence of any *toxicity, severe toxicity, identity attack, insult, profanity,* or *threat*), but not with any of the *New York Times* attributes. Interestingly, for both the collapsed Perspective API measure and *inflammatory language*, the effect of proximate anti-sociality was positive when a user's initial anti-sociality received no down votes—an effect that went away as down votes increased. The Perspective API thus picks up on forms of anti-social commenting that are sensitive to both descriptive and injunctive norms.

Comparing Reddit and Twitter Discussions of the January 6th Capitol Riots

In an effort to apply the norms-based framework we developed – the potential effects of others' antisocial commenting as well as social approval votes – to a more contemporary online context, we employed a novel dataset consisting of discussions surrounding the insurrection that followed the 2020 U.S. presidential election. The collective that raided the United States Capitol building on January 6th, 2021 was, in part, a product of online socialization processes (Ng et al., 2022). Given this, we looked for norms surrounding anti-social commenting as the events of January 6th unfolded. Data from both Reddit and Twitter highlight both the broad applicability of our framework as well as the nuanced understanding of social norms that it can provide. These two platforms differ in the way that social interactions are structured, with Reddit organizing discussions in topic-based forums and Twitter employing a network graph of follower relations. Moderation also works differently across these two platforms, with Reddit relying on community members and Twitter on algorithmic solutions. Norms may be more influential on platforms like Twitter, where out-of-community interactions are more likely and moderation is less specific to one's own community. The analysis of the Capitol riots provides an opportunity to examine anti-social commenting norms ten years after the *ADS* dataset was constructed, during an especially contentious time, and across these different platforms.

Reddit data are organized into ”submissions” and ”comments.“ Submissions are prompts and comments are responses either to prompts or to other comments. We first collected all of the submissions that mentioned the word ”Capitol” between 11am on January 6th and 11am on January 7th EST (i.e., Washington, D.C. time, where the insurrection took place). We then filtered down the set of submissions to include only those that had between 100-500 comments. This set reflects 3% of submissions and 9% of comments, respectively. This was done to constrain the amount of comment data we analyzed for anti-sociality on the one hand, but also to ensure that (1) there were enough comments per submission to model submission-specific random effects, and (2) all of the comments for each submission were obtained, as the Reddit API (which was used for data collection) only allows for 500 comments per submission to be collected. We then collected all of the comments posted on the submissions in our filtered set, organized the comments into triplets similar to those discussed in the previous section, and applied several automated classifiers of anti-social commenting (i.e., name-calling, toxicity, severe toxicity, identity attack, insult, threat, attack on author, attack on commenter, and inflammatory) to each of the comments in these triplets.

Data collection for Twitter proceeded in much the same way. Since tweet volume is much larger than that of Reddit submissions, we sampled one ten-second interval per minute for the same 24-hour period and collected all of the tweets in these intervals that included the word ”Capitol.“ We removed retweets and replies that matched our query (i.e., we constrained the data to original tweets) and filtered down the tweets to those that had between 100-500 replies, in order to be consistent with the Reddit data collection. The filtered set of original tweets and replies captures .3% and 17% of the respective totals. We then collected the replies to the tweets in our filtered set, created triplets, and processed the tweets in these triplets using the automated classifiers. One additional detail is that discussions on Reddit and Twitter are organized in tree-like threads rather than single comment streams as in the *ADS* dataset. Our triplets thus reflect any three sequential replies in which the initial and final comments were made by the same user. Unlike the triplets constructed from the *ADS* dataset, we did not allow for a “gap” of multiple comments separating the first and third comment, as these are not well-specified in tree-like threads because each comment can split into separate sub-threads. In all, we analyzed 12,594 triplets on Reddit and 6,303 triplets on Twitter. Table 4 presents summary statistics for the three comments in each triplet.

Collective descriptive norms on Reddit and Twitter

These first set of analyses we report reveal the *collective* descriptive norms surrounding anti-social commenting on Reddit and Twitter. *Name-calling* was prevalent in these data, with 21% and 19% of the initial comments including name-calling on Reddit and Twitter, respectively (compared to just 13% in the *ADS* comments). Scores on the Perspective API measures were substantial, with 41% and 35% of comments on Reddit and Twitter including at least one of the six measures, compared to just 6% of the comments in the *ADS*. While these differences between the 2011 ADS dataset and 2021 Capitol dataset are notable, recent work acknowledges the limitations of the Perspective API for making comparisons over time (Pozzobon et al., 2023). The *New York Times* attributes' scores were not particularly high—in fact, *attacks on author* were less common than in the *ADS* analyses, at 5% and 2% for Reddit and Twitter, respectively (compared to 10% in the *ADS*). Anti-social comments were typically more prevalent on Reddit than on Twitter. One exception was *attacks on other commenters*, which occurred in a staggering 45% of initial comments on Twitter, compared to 24% of those on Reddit. Understanding the reasons for these cross-platform differences is beyond the scope of this chapter but is a worthwhile direction for future research.

<Table 4>

Figure 8 shows correlations for these different forms of anti-social commenting. The bottom panels report correlations for Reddit comments and Twitter replies separately. The upper left panel includes both platforms together. Finally, the upper right panel includes differences in the correlations across the two platforms. The classifier for *name-calling* was consistently correlated with the Perspective API measures, aside from *threat*. This was not the case in the *ADS* comments, which featured relatively low rates of the Perspective API measures, in contrast to which, comments on Reddit and Twitter frequently included *several* forms of anti-social commenting. Finally, there tended to be stronger correlations among the Perspective API measures on Reddit (more positive) and higher correlations between the *New York Times* attributes and Perspective API measures on Twitter (more negative). The correlations among the

New York Times Attributes were split, with Reddit showing a stronger correlation between *attacks on author* and *attacks on commenter*, and Twitter showing a stronger correlation between *attack on commenter* and *inflammatory language*. Multiple forms and variations of anti-social commenting were often used in harmony, with slight differences in co-occurrence rates between the two platforms.

<Figure 8>

Collective injunctive norms on Reddit and Twitter

Next, we analyze the *collective* injunctive norms across the two platforms by modeling the effects of anti-sociality on how many votes comments received. On Twitter, comments can be favorited (or not rated at all) while on Reddit comments can receive down or up votes (or neither). However, the Reddit API only provides a single "score," which captures the difference in the number of up and down votes and can thus be negative. As such, we use a linear model specification with fixed effects for the submission (on Reddit) or conversation ID (on Twitter), as well as for the specific commenter. We do not include a control variable for the order of the comment, as this is not clearly defined for a tree-like thread structure. We report the results of these analyses in Table 5 and depict them in Figure 9. The name-calling classifier annotated only about a third of the Reddit comments because many of them exceeded the default input length limitations of the tool.[4] This introduces bias in our analysis towards shorter comments, but also makes the comparison to Twitter more apt. All of the Twitter tweets were coded properly as they tend to be shorter than Reddit comments. Several measures of anti-sociality were rewarded (i.e., received relatively more up votes than down votes) on Reddit, including *toxicity*, *severe toxicity*, *insult,* and *inflammatory language*. In contrast, *toxicity* and *insult* were associated with fewer votes (i.e., 'favorites') on Twitter. *Attacks on other commenters* were consistently associated with lower social rewards on both platforms.

<Table 5>

<Figure 9>

Perceived descriptive and injunctive norms on Reddit and Twitter

We close this section with a set of analyses investigating the role of *perceived* descriptive and injunctive norms on the spread of anti-social commenting across the two platforms. While the previous analyses of collective norms focused on aggregate rates of anti-social comments and their associated social rewards, an analysis of perceived norms investigates instead the effects on individuals' postings due to their exposure to anti-social commenting and associated rewards within a conversational thread. We modeled the outcomes, or dependent variables, using several of the anti-social features that were used in the previous analyses, above, but as they appeared in the final triplet comment. We tested whether anti-social comments resulted from a statistical interaction between (1) anti-sociality in the triplet's initial comment, (2) anti-sociality in the proximate comment, (3) the number of votes of approval that the initial comment received, and (4) whether the comments were on Reddit or Twitter. A multilevel modeling procedure specified a random effects variable representing the submission number (on Reddit) or conversation ID (on Twitter), as well as a random effects variable representing each contributor. The full numerical results are available from the first author upon request, but the marginal effects depicted in Figure 10 reflect definitive patterns.

<Figure 10>

Across a range of anti-social features, proximate anti-sociality on Reddit was associated with a greater likelihood of anti-sociality in the final comment. However, we found no effects of votes on Reddit, and the rate of anti-sociality in a Redditor's final comment did not differ due to anti-sociality in that user's initial comment. On Twitter, on the other hand, the rate of anti-sociality in the final comment was regularly associated with anti-sociality in prior posts as well as votes that a triplet's initial comment received for anti-sociality. This result suggests that the theory of normative social behavior (Rimal & Real, 2005) applies to Twitter's dynamics more than Reddit's. The theory predicts an interaction between descriptive and injunctive norms (in this case, between proximate anti-sociality and votes for initial anti-sociality). *Insult* on Twitter, in particular, appears partially caused by such an interaction effect and, to a lesser extent, so does *severe toxicity*, *identity attacks*, *profanity*, and *threat*. An unexpected finding was that, on Twitter, votes for initial anti-sociality were associated with *lower* rates of subsequent *name-calling* and *inflammatory language*, suggesting a possible satiation effect whereby rewards for anti-sociality filled a need that no longer must be met, a threshold effect, providing an interesting direction for future research.

To summarize, we found notable differences between Reddit and Twitter in the constitution of norms surrounding anti-social commenting. At the collective level, discussions on Reddit tended to feature higher rates of anti-social commenting than Twitter, except for *attacks on commenters* which were greater on Twitter. These aggregate descriptive norms are useful for understanding the kinds of language that users are exposed to on the two platforms, albeit at a very abstract level. Rewards for anti-social commenting were more common on Reddit than Twitter, suggesting that (collective) injunctive norms are more favorable to anti-sociality on Reddit. However, when shifting from collective to perceived norms, a slightly different picture emerges. In particular, while votes for anti-sociality were more common on Reddit, they may be more *influential* on Twitter. Being rewarded for anti-social comments on Twitter increased a contributor's likelihood of repeating anti-social behavior, while the same rewards did not produce additional anti-social messaging on Reddit. Injunctive norms surrounding anti-sociality are thus more incendiary on Twitter than Reddit, possibly due to the aforementioned differences in their interaction structure or moderation practices. The platforms do not appear to differ, however, in the impact of descriptive norms—on both platforms, being exposed to anti-social language is associated with higher rates of anti-social language use, suggesting that mimetics perpetuate anti-sociality across both platforms.

Discussion

If online anti-social language use is based in normative considerations, then the combination of online discussion thread data and automated text classification techniques are together the equivalent of a microscope and telescope (i.e., macroscope) for the study of social norm formation and evolution. A microscope because individual-level behavior can be tracked over time to study the formation and evolution of perceived norms, and a telescope because aggregate trends in anti-social commenting can be tracked to understand the formation and evolution of collective norms at scale. By providing behavioral researchers with a rich set of *relational* artifacts as well as a long and granular *temporal* frame, online discussion data can be used to track when and why people and collectives conform to their social surroundings. This study validates and applies text-based classification methods to uncover normative dynamics that underlie the use and spread of anti-social language online. In addition to distinguishing between collective and perceived norms at the analytical level (Lapinski & Rimal 2005), we also separate descriptive and injunctive norms (Cialdini et al., 1990) at the measurement level. We argue that the inter-relations of series of texts among online comments can capture descriptive norms surrounding anti-sociality, while the allocation of social votes reflects the injunctive norms, and that both can play a role in fueling anti-social comments.

This chapter offers several contributions to understanding the social processes that underlie anti-social commenting. First and foremost, it maps the widely-used social scientific constructs, *descriptive* and *injunctive norms*, onto features of digital trace data that are increasingly useful for contemporary understanding of human behavior. In particular, we argue that descriptive norms can be measured using the text contained within a comment, while injunctive norms can be assessed using data about the social "votes" that comments receive. Second, we demonstrated how automated classification techniques can capture the presence of anti-social commenting within a comment, yielding similar results as human-coded data in many cases (and especially for the measures obtained with the Perspective API). Using human-coded data, Shmargad et al. (2022) show that anti-social commenting is sensitive to descriptive and injunctive norms, and we show here, using comments from the online *Arizona Daily Star* that the Perspective API (Lees et al., 2022) can be used to replicate these findings. Finally, we analyzed the collective and perceived norms around anti-social commenting as the January 6th capitol riots unfolded and showed how these differed across the social media platforms Reddit and Twitter. While anti-social language was more likely to be rewarded on Reddit than Twitter, rewards were more influential for the spread of anti-social language on Twitter. The presence and influence of anti-social norms can thus vary widely across sociotechnical contexts.

The broad theoretical approach we introduce, which delineates self- and other-focused signals of descriptive and injunctive information available in online discussion data, can be used in a variety of ways that go beyond the analyses that we present. One direction worth pursuing is the adoption of a longer time window of individual-level behavior (e.g., Rains et al., 2021) and the social contexts from which it emerges. In contrast, the temporal dimension we study here is short and focuses more on the dynamics of comments in a single thread. A longer view of how an individual's contributions change over time could yield new knowledge about how anti-social personalities develop (Bor & Peterson, 2021) and, more importantly, knowledge about when and why people evolve *out of* anti-social patterns. This will help to inform moderation policies that are less myopic than those that focus on the removal of specific comments, and to provide more sustainable solutions for how we might design public spaces that yield the kinds of discussions (and deviance) that achieve a balance between individual and collective ambitions, thereby more accommodating to the inherent social processes the underlie online posting and commenting.

New applications in automated text analysis is likely to propel even more advances. Text similarity techniques such as TF-IDF (Bail, 2016), could be used to measure the extent that one's posting behavior conforms to or deviates from their previous posts, or from other people's posts on the same discussion thread. Large language models, such as those created by OpenAI, can be used to construct more nuanced annotations and interpretations of text that are then applied at scale.[5] For example, a large language model can be prompted to detect whether or not two sequential comments are in agreement or disagreement, which can be an important moderator in addition to social votes for whether descriptive norms are propagated. This could provide more contextual information that can be used to investigate both moderators and catalysts of anti-social language spread. The validation of such language models for large-scale annotation is an important direction for future research.

Despite the aforementioned advantages of these new data sources and text classification techniques, there of course remain many barriers to their successful application in social science research. The findings we report here suggest that automated classifiers, and especially those obtained with the Perspective API, can be used to generate similar behavioral findings as human annotations *on average*. There will, however, inevitably remain variation in how people interpret the contents of online messages, and incivility in particular (Kenski et al., 2020). One risk of the kinds of automated techniques we deploy is that, by providing central tendencies in text interpretation, they remove variation in human interpretation, especially infrequent or extreme content, that may very well be informative sources or sites of social deviance. Finally, we note that people do not live their entire lives online and there is enduring variation across (as well as within) people in their engagement with online platforms. Our framework should thus be viewed as a starting point for incorporating online discussion data into the study of anti-sociality, and we encourage the concurrent investigation of multiple platforms, offline activities, and other measures obtained through more traditional instruments such as interviews and surveys.

EndNotes

[1] https://huggingface.co/civility-lab. Accessed November 22, 2023.

[2] More information on the attributes from the Perspective API, including those trained on the New York Times comments, can be found here: https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages. Accessed November 22, 2023.

[3] That the rate of anti-sociality in comments starting a triplet does not differ substantially from that of comments in general should not be viewed as evidence for a lack of social influence. This is because early comments will not necessarily start a triplet, as the triplets we constructed to have the particular structure that allows for a test of our framework (that is, they originate and end with the same commenter with a different commenter sandwiched between them). For example, if the identities of the contributors of a string of comments are 1234546, then there will only be one applicable triplet (454) and it will not be found at the beginning of the set.

[4] Upon further investigation, we discovered that the default input length limitations can be circumvented with additional code. However, a complete reanalysis of the data was prohibitively time-consuming and is left for future work.

[5] More information about OpenAI's API can be found here: https://platform.openai.com/overview. Accessed November 22, 2023.

References

Barbera, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76-91. https://doi.org/10.1093/pan/mpu011

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. A. (2012). The role of social networks in information diffusion. *Proceedings of the 21st international conference on world wide web* (pp. 519-528). ACM. https://doi.org/10.1145/2187836.2187907

Bail, C. A. (2016). Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences*, *113*(42), 11823-11828. https://doi.org/10.1073/pnas.1607151113

Bor, A., & Peterson, M. B. (2021). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*, *116*(1), 1-18. https://doi.org/10.1017/S0003055421000885

Burrell, J., & Fourcade, M. (2021). The society of algorithms. *Annual Review of Sociology*, *47*, 213-237. https://doi.org/10.1146/annurev-soc-090820-020800

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2014). How community feedback shapes user behavior. *Proceedings of the Eighth international AAAI conference on weblogs and social media* (ICWSM), *8*(1), 41-50. Association for the Advancement of Artificial Intelligence. https://doi.org/10.48550/arXiv.1405.1429

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Antisocial behavior in online discussion communities. *Proceedings of the Eleventh international AAAI conference on web and social media* (ICWSM), *9*(1), 61-70. Association for the Advancement of Artificial Intelligence. https://doi.org/10.48550/arXiv.1504.00680

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015-1026. https://doi.org/10.1037/0022-3514.58.6.1015

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658-679. https://doi.org/10.1111/jcom.12104

Ding, B., Qin, C., Liu, L., Bing, L., Joty, S., & Li, B. (2022). Is GPT-3 a good data annotator? *ArXiv*. https://doi.org/10.48550/arXiv.2212.10450

Durkheim, E. (1893). *The division of labor in society*. The Free Press.

Edyvane, D. (2020). Incivility as dissent. *Political Studies*, *68*(1), 93-109. https://doi.org/10.1177/0032321719831983

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. *Proceedings of the Twelfth international AAAI conference on web and social media* (ICWSM), *12*(1). Association for the Advancement of Artificial Intelligence. https://doi.org/10.48550/arXiv.1804.04257

Forestal, J. (2021). Beyond gatekeeping: Propaganda, democracy, and the organization of digital publics. *Journal of Politics*, *83*(1), 306-320. https://doi.org/10.1086/709300

Freelon, D., Bossetta, M., Wells, C., Lukito, J. Xia, Y., & Adams, K. (2020). Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Social Science Computer Review*, *40*(3), 560-578. https://doi.org/10.1177/0894439320914853

Golder, S. A., Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, *40*, 129-152. https://doi.org/10.1146/annurev-soc-071913-043145

Goldberg, A., Stein, S. K. (2018). Beyond social contagion: Associative diffusion and the emergence of cultural variation. *American Sociological Review*, *83*(5), 897-932. https://doi.org/10.1177/0003122418797576

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on human-computer interaction* (CSCW), *5*, 1-35. https://doi.org/10.1145/3479610

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, *2*(4), 1-4. https://doi.org/10.1016/j.patter.2021.100241

Jamieson, K. H., Volinsky, A., Weitz, I., & Kenski, K. (2017). The political uses and abuses of civility and incivility. *The Oxford handbook of political communication* (pp. 205-218). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793471.013.79_update_001

Jiang, S., Robertson, R. E., & Wilson, C. (2020). Reasoning about political bias in content moderation. *Proceedings of the 34th AAAI conference on artificial intelligence, 34*(9), 13669-13672). Association for the Advancement of Artificial Intelligence. https://doi.org/10.1609/aaai.v34i09.7117

Kenski, K., Coe, K., Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research*, *47*(6), 795-814. https://doi.org/10.1177/0093650217699933

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication, 71(6)*, 922-946. https://doi.org/10.1093/joc/jqab034

Kim, K. K., Lee, A. R., & Lee, U. (2019). Impact of anonymity on roles of personal and group identities in online communities. *Information & Management*, *56*, 109-121. https://doi.org/10.1016/j.im.2018.07.005

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *ArXiv*. https://doi.org/10.48550/arXiv.2302.02083

Lapinski, M. K., & Rimal, R. N. (2005). An explication of social norms. *Communication Theory*, *15*(2), 127-147. https://doi.org/10.1111/j.1468-2885.2005.tb00329.x

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., & Vasserman, L. (2022). A new generation of perspective API: Efficient multilingual character-level transformers. *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (KDD 2022), 3197-3207. https://doi.org/10.1145/3534678.3539147

Lofland, J. (1969). *Deviance and identity*. Prentice-Hall.

Maratea, R. J., & Kavanaugh, P. R. (2012). Deviant identity in online contexts: New directions in the study of a classic concept. *Sociology Compass*, *6*(2), 102-112. https://doi.org/10.1111/j.1751-9020.2011.00438.x

McDonnell, T. E., Bail, C. A., & Tavory, I. (2017). A theory of resonance. *Sociological Theory*, *35*(1), 1-14. https://doi.org/10.1177/0735275117692837

Ng, L. H. X., Cruickshank, I. J., & Carley, K. M. (2022). Cross-platform information spread during the January 6th capitol riots. *Social Network Analysis and Mining*, *12*(1), 133. https://doi.org/10.1007/s13278-022-00937-1

Ozler, K. B., Kenski, K., Rains, S. A., Shmargad, Y., Coe, K., & Bethard, S. (2020). Fine-tuning for multi-domain and multi-label uncivil language detection. *Proceedings of the fourth workshop on online abuse and harms*, 28-33. https://doi.org/10.18653/v1/2020.alw-1.4

Papakyriakopoulos, O., Engelmann, S., & Winecoff, A. (2023). Upvotes? Downvotes? No votes? Understanding the relationship between reaction mechanisms and political discourse on Reddit. *Proceedings of the 2023 CHI conference on human factors in computing systems*, 549, 1-28. ACM. https://doi.org/10.1145/3544548.3580644

Pozzobon, L., Ermis, B., Lewis, P., & Hooker, S. (2023). On the Challenges of Using Black-Box APIs for Toxicity Evaluation in Research. *ArXiv. https://doi.org/10.48550/arXiv.2304.12397*

Rains, S. A., Kenski, K., Coe, K., & Harwood, J. (2017). Incivility and political identity on the internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication*, *22*(4), 163-178. https://doi.org/10.1111/jcc4.12191

Rains, S. A., Shmargad, Y., Coe, K., Kenski, K., & Bethard, S. (2021). Assessing the Russian troll efforts to sow discord on Twitter during the 2016 US election. *Human Communication Research*, *47*(4), 477-486. https://doi.org/10.1093/hcr/hqab009

Rains, S. A., Harwood, J., Shmargad, Y., Kenski, K., Coe, K., & Bethard, S. (2023a). Engagement with partisan Russian troll tweets during the 2016 US presidential election: a social identity perspective. *Journal of Communication*, *73*(1), 38-48. https://doi.org/10.1093/joc/jqac03

Rains, S. A., Kenski, K., Dajches, L., Duncan, K., Yan, K., Shin, Y., Barbati, J. L., Bethard, S., Coe, K., & Shmargad, Y. (2023b). Engagement with incivility in tweets from and directed at local elected officials. *Communication and Democracy*, *57*(1), 143-152. https://doi.org/10.1080/27671127.2023.2195467

Rimal, R. N., & Real, K. (2005). How behaviors are influenced by perceived norms: A test of the theory of normative social behavior. *Communication Research*, *32*(3), 389-414. https://doi.org/10.1177/0093650205275385

Rossini, P. (2022). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, *49(*3), 399-425. https://doi.org/10.1177/0093650220921314

Sadeque, F., Rains, S. A., Shmargad, Y., Kenski K., Coe, K., & Bethard, S. (2019). Incivility detection in online comments. *Proceedings of the eighth joint conference on lexical and computational semantics* (SEM 2019) (pp. 283-291). Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-1031

Shmargad, Y. (2022). Twitter influencers in the 2016 US Congressional races. *Journal of Political Marketing*, *22*(1), 23-40. https://doi.org/10.1080/15377857.2018.1513385

Shmargad, Y., Coe, K., Kenski, K., & Rains, S. A. (2022). Social norms and the dynamics of online incivility. *Social Science Computer Review*, *40*(3), 717-735. https://doi.org/10.1177/0894439320985527

Song, Y., Lin, Q., Kwon, K. H., Choy, C. H. Y., & Xu, R. (2022). Contagion of offensive speech online: An interactional analysis of political swearing. *Computers in Human Behavior*, *127*, 107046. https://doi.org/10.1016/j.chb.2021.107046

Tontodimamma, A. Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*, 157-179. https://doi.org/10.1007/s11192-020-03737-6

Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence. Big Data & Society 10(1). https://doi.org/10.1177/20539517231172424

Vogels, E. A. (2021). The state of online harassment. *Pew Research Center*. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 1415-1420). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1144

Zerubavel, E. (1999). *Social mindscapes: An invitation to cognitive sociology*. Harvard University Press.

Table 1

Summary Statistics of Anti-Social Commenting in the Arizona Daily Star

	All Comments			First Comment in Triplet		
	N	*M*	*SD*	*N*	*M*	*SD*
Annotated Incivility	6,165	0.20	0.40	2,672	0.19	0.39
Name-calling		0.14	0.35		0.12	0.33
Not Name-calling		0.09	0.28		0.08	0.28
Automated Name-calling	6,121	0.14	0.35	2,620	0.13	0.33
Perspective API	5,998			2,534		
Toxicity		0.05	0.21		0.04	0.20
Severe Toxicity		0.00	0.00		0.00	0.00
Identity Attack		0.01	0.08		0.01	0.07
Insult		0.05	0.22		0.04	0.20
Profanity		0.00	0.06		0.00	0.07
Threat		0.01	0.08		0.00	0.06
Any Perspective		0.06	0.24		0.05	0.22
New York Times	5,998			2,534		
Attack on Author		0.10	0.30		0.10	0.30
Attack on Commenter		0.22	0.41		0.25	0.43
Inflammatory		0.33	0.47		0.30	0.46

Table 2
Definitions and Examples of Anti-Social Commenting in the Arizona Daily Star

Form of Anti-Sociality	Definition	Example
Name-calling	Mean-spirited or disparaging words directed at a person or group of people	“You ARE BREAKING THE LAW, you dopes.”
Perspective API		
Toxicity	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion	“Useful idiots!
Severe Toxicity	A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words	“Arizona voters are stupid and they get what they deserve by electing these scum sucking sleaze balls.” (Note: this had the highest severe toxicity score, but at .45 did not meet the .5 threshold to be classified as severe toxicity)
Identity Attack	Negative or hateful comments targeting someone because of their identity	“Mexico sending more INVADERS to our nation.”
Insult	Insulting, inflammatory, or negative comment towards a person or a group of people.	“You are truly an IDIOT.”
Profanity	Swear words, curse words, or other obscene or profane language	“Just build the damn mine already!”
Threat	Describes an intention to inflict pain, injury, or violence against an individual or group	“Every person on their death-bed should die in pain.”
New York Times		
Attack on Author	Attack on the author of an article or post.	“Leave it to Tony to turn a ‘news’ story into a one sided bleeding heart opinion piece.”
Attack on Commenter	Attack on fellow commenter.	“All of what you said just shows that you haven't grown up yet.”
Inflammatory	Intending to provoke or inflame.	“Want a chance at a job, get rid of Obama, Pelosi, and the rest. Want a government handout like the OWS clowns, then vote for Obama. Simple enough.

Note: Definitions for the Perspective API and New York Times features were obtained from https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages. For examples of the *annotated* measures of incivility (i.e., Name-Calling, Aspersion, Accusations of Lying, Vulgarity, and Pejorative for Speech), see Table 1 in Coe et al. (2014)

Table 3

Effects of Anti-Social Commenting on Downvotes and Upvotes

	N	Downvotes: Coeff	(S.E.)	Upvotes: Coeff	(S.E.)
Annotated Incivility	5,665	0.65*	(0.26)	1.20**	(0.43)
Name-calling		0.34	(0.30)	1.16*	(0.51)
Not Name-calling		0.79*	(0.37)	0.84	(0.55)
Automated Name-calling	5,621	0.55	(0.30)	0.44	(0.50)
Perspective API	5,506				
Toxicity		0.90	(0.56)	1.53	(1.03)
Severe Toxicity		0.00	(0.00)	0.00	(0.00)
Identity Attack		1.15	(2.78)	0.48	(2.55)
Insult		0.37	(0.51)	1.74*	(0.95)
Profanity		0.76	(2.29)	1.80	(2.91)
Threat		2.21	(2.04)	-2.17	(2.38)
Any Perspective API	5,506	0.32	(0.45)	1.62*	(0.85)
New York Times	5,506				
Attack on Author		0.59	(0.36)	1.13	(0.59)
Attack on Commenter		0.19	(0.25)	-1.26**	(0.40)
Inflammatory		0.96***	(0.24)	0.69	(0.39)

Note: * $p < .10$, ** $p < .05$, *** $p < .001$

Table 4

Means and Standard Deviations of Anti-Social Commenting on Reddit and Twitter

	Initial Comment		Proximate Comment		Subsequent Comment	
	Reddit	Twitter	Reddit	Twitter	Reddit	Twitter
Name-calling	.21 (.40)	.19 (.39)	.20 (.40)	.18 (.39)	.18 (.38)	.17 (.38)
Perspective API						
Toxicity	.28 (.45)	.23 (.42)	.26 (.44)	.24 (.43)	.24 (.42)	.22 (.42)
Severe Toxicity	.15 (.36)	.10 (.31)	.13 (.34)	.11 (.31)	.12 (.32)	.10 (.30)
Identity Attack	.12 (.32)	.11 (.32)	.10 (.30)	.11 (.31)	.09 (.28)	.10 (.30)
Insult	.27 (.45)	.23 (.42)	.25 (.43)	.23 (.42)	.23 (.42)	.22 (.42)
Profanity	.19 (.39)	.13 (.33)	.17 (.37)	.14 (.34)	.16 (.37)	.13 (.34)
Threat	.19 (.39)	.14 (.35)	.17 (.37)	.13 (.33)	.14 (.35)	.11 (.31)
All Perspective	.41 (.49)	.35 (.48)	.38 (.49)	.35 (.48)	.35 (.48)	.32 (.46)
New York Times						
Attack on Auth	.05 (.21)	.02 (.15)	.05 (.23)	.03 (.16)	.05 (.23)	.03 (.16)
Attack on Comm	.24 (.43)	.45 (.50)	.29 (.45)	.46 (.50)	.29 (.45)	.45 (.50)
Inflammatory	.41 (.49)	.33 (.47)	.37 (.48)	.33 (.47)	.34 (.47)	.30 (.46)

Table 5

Effects of Anti-Social Commenting on Votes in the Initial Comments

	Reddit			Twitter		
	N	Coeff.	(S.E.)	N	Coeff.	(S.E.)
Name-calling	2,982	4.91	(3.37)	4,939	-0.27	(1.42)
Perspective API	9,300			4,939		
Toxicity		6.73**	(1.89)		-1.45*	(0.76)
Severe Toxicity		9.32***	(2.18)		-0.56	(0.58)
Identity Attack		3.21	(2.87)		-1.30	(0.84)
Insult		8.54***	(1.99)		-1.71**	(0.70)
Threat		0.13	(1.84)		1.65	(1.53)
Any Perspective API		6.16***	(1.72)		-1.00	(0.93)
New York Times	9,300			4,939		
Attack on Author		-1.62	(2.14)		-0.02	(1.48)
Attack on Commenter		-3.27*	(1.71)		-2.82**	(1.17)
Inflammatory		6.50*	(2.83)		-1.04	(1.12)

Note: * *p* < .10, ** *p* < .05, *** *p* < .001

Figure 1

A Framework for Using Online Discussion Data to Study Socialization Processes

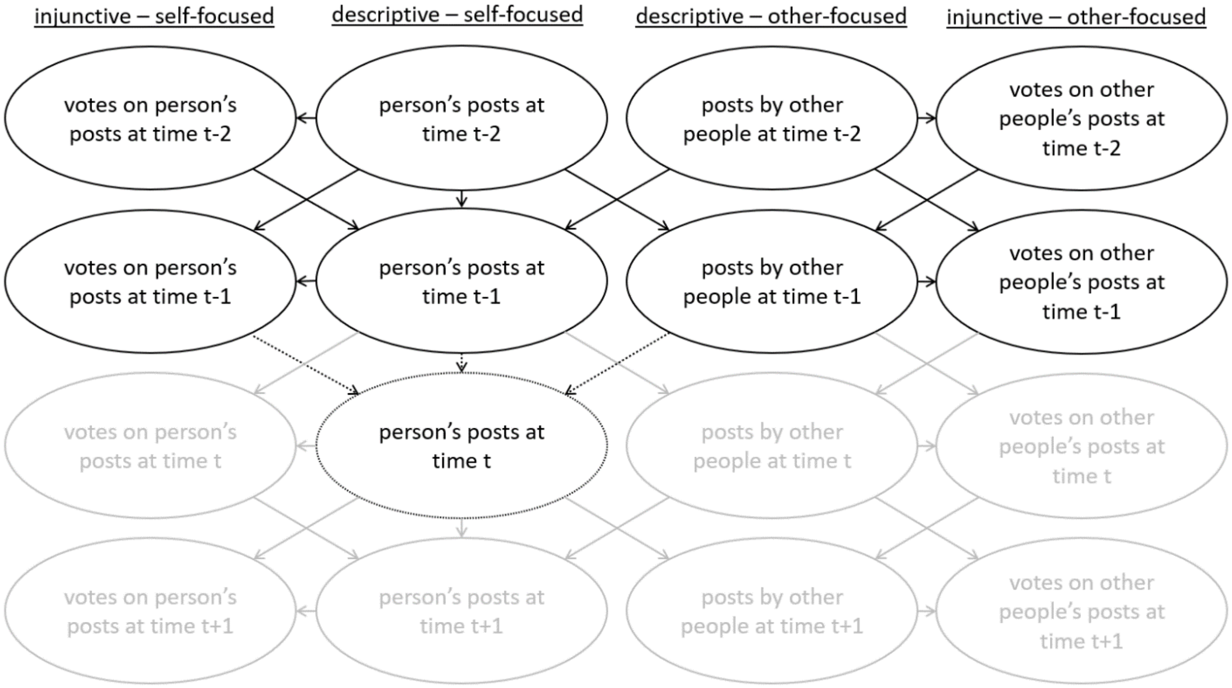

injunctive – self-focused
descriptive – self-focused
descriptive – other-focused
injunctive – other-focused
votes on person's posts at time t-2
person's posts at time t-2
posts by other people at time t-2
votes on other people's posts at time t-2
votes on person's posts at time t-1
person's posts at time t-1
posts by other people at time t-1
votes on other people's posts at time t-1
votes on person's posts at time t
person's posts at time t
posts by other people at time t
votes on other people's posts at time t
votes on person's posts at time t+1
person's posts at time t+1
posts by other people at time t+1
votes on other people's posts at time t+1

Figure 2

Organizing Comments into Triplets

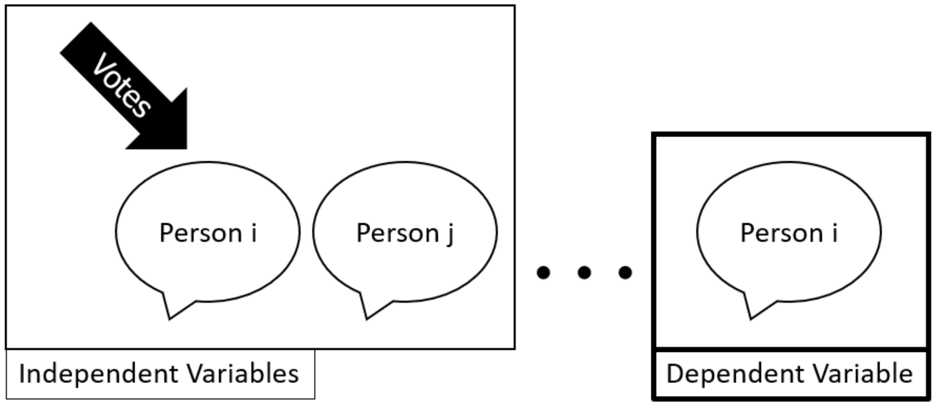
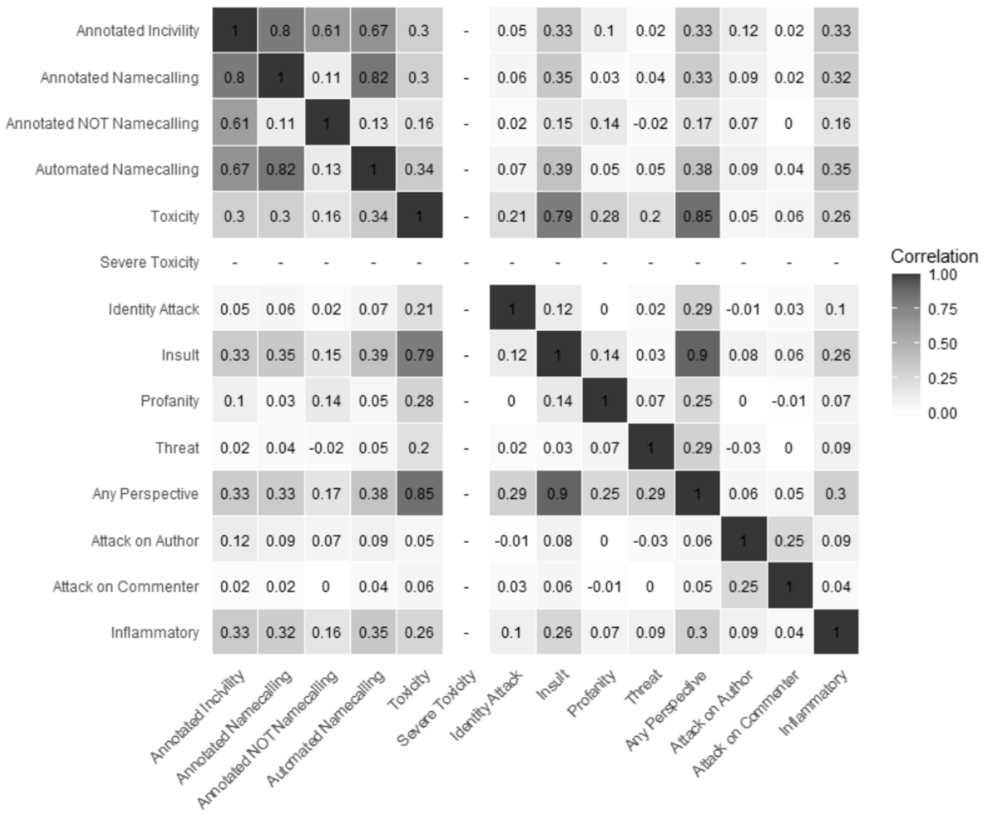


Figure 3

Correlation Matrix for Measures of Anti-Sociality in Online News Comments



	Annotated Incivility	Annotated Namecalling	Annotated NOT Namecalling	Automated Namecalling	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat	Any Perspective	Attack on Author	Attack on Commenter	Inflammatory
Annotated Incivility	1	0.8	0.61	0.67	0.3	-	0.05	0.33	0.1	0.02	0.33	0.12	0.02	0.33
Annotated Namecalling	0.8	1	0.11	0.82	0.3	-	0.06	0.35	0.03	0.04	0.33	0.09	0.02	0.32
Annotated NOT Namecalling	0.61	0.11	1	0.13	0.16	-	0.02	0.15	0.14	-0.02	0.17	0.07	0	0.16
Automated Namecalling	0.67	0.82	0.13	1	0.34	-	0.07	0.39	0.05	0.05	0.38	0.09	0.04	0.35
Toxicity	0.3	0.3	0.16	0.34	1	-	0.21	0.79	0.28	0.2	0.85	0.05	0.06	0.26
Severe Toxicity	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Identity Attack	0.05	0.06	0.02	0.07	0.21	-	1	0.12	0	0.02	0.29	-0.01	0.03	0.1
Insult	0.33	0.35	0.15	0.39	0.79	-	0.12	1	0.14	0.03	0.9	0.08	0.06	0.26
Profanity	0.1	0.03	0.14	0.05	0.28	-	0	0.14	1	0.07	0.25	0	-0.01	0.07
Threat	0.02	0.04	-0.02	0.05	0.2	-	0.02	0.03	0.07	1	0.29	-0.03	0	0.09
Any Perspective	0.33	0.33	0.17	0.38	0.85	-	0.29	0.9	0.25	0.29	1	0.06	0.05	0.3
Attack on Author	0.12	0.09	0.07	0.09	0.05	-	-0.01	0.08	0	-0.03	0.06	1	0.25	0.09
Attack on Commenter	0.02	0.02	0	0.04	0.06	-	0.03	0.06	-0.01	0	0.05	0.25	1	0.04
Inflammatory	0.33	0.32	0.16	0.35	0.26	-	0.1	0.26	0.07	0.09	0.3	0.09	0.04	1

Figure 4

Log-Log Plot of Down and Up Votes after Article Standardization

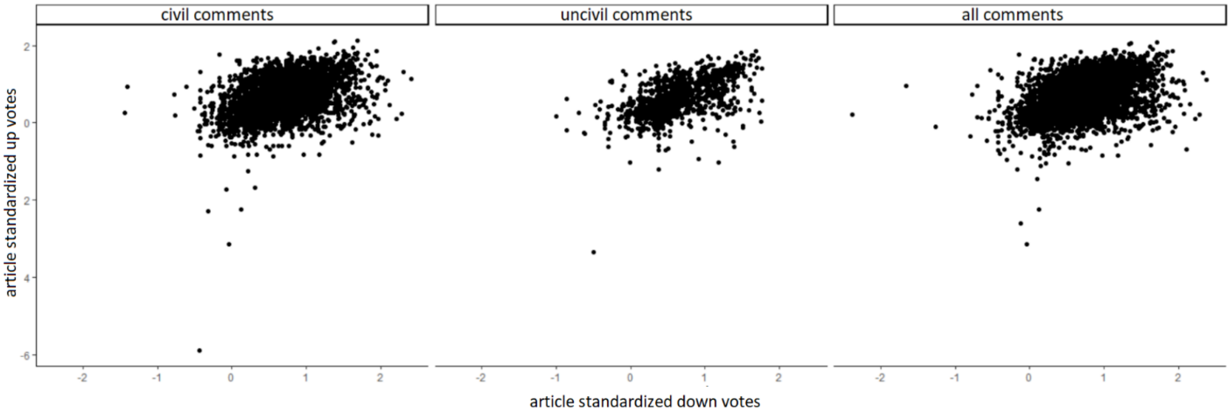


Figure 5

Effects of Anti-Social Commenting on Votes with 95% Confidence Intervals

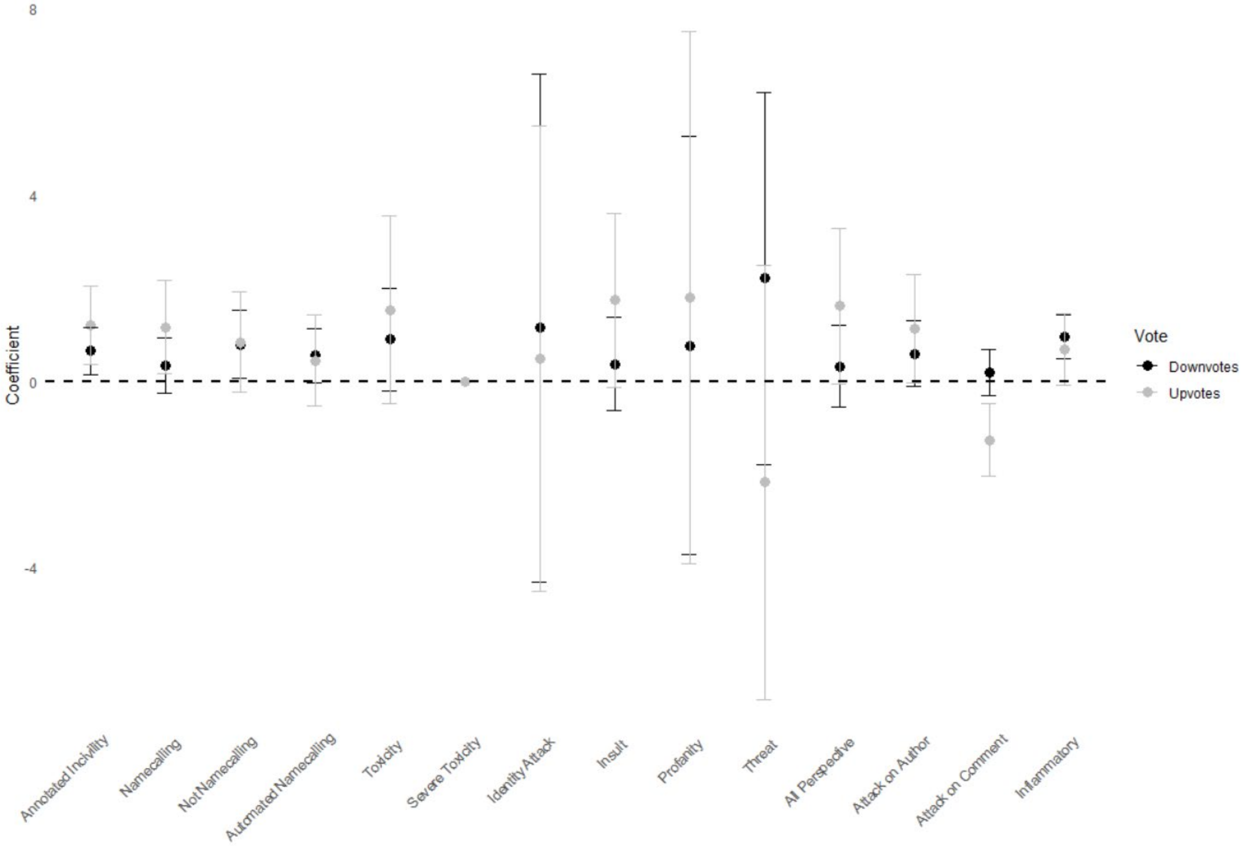
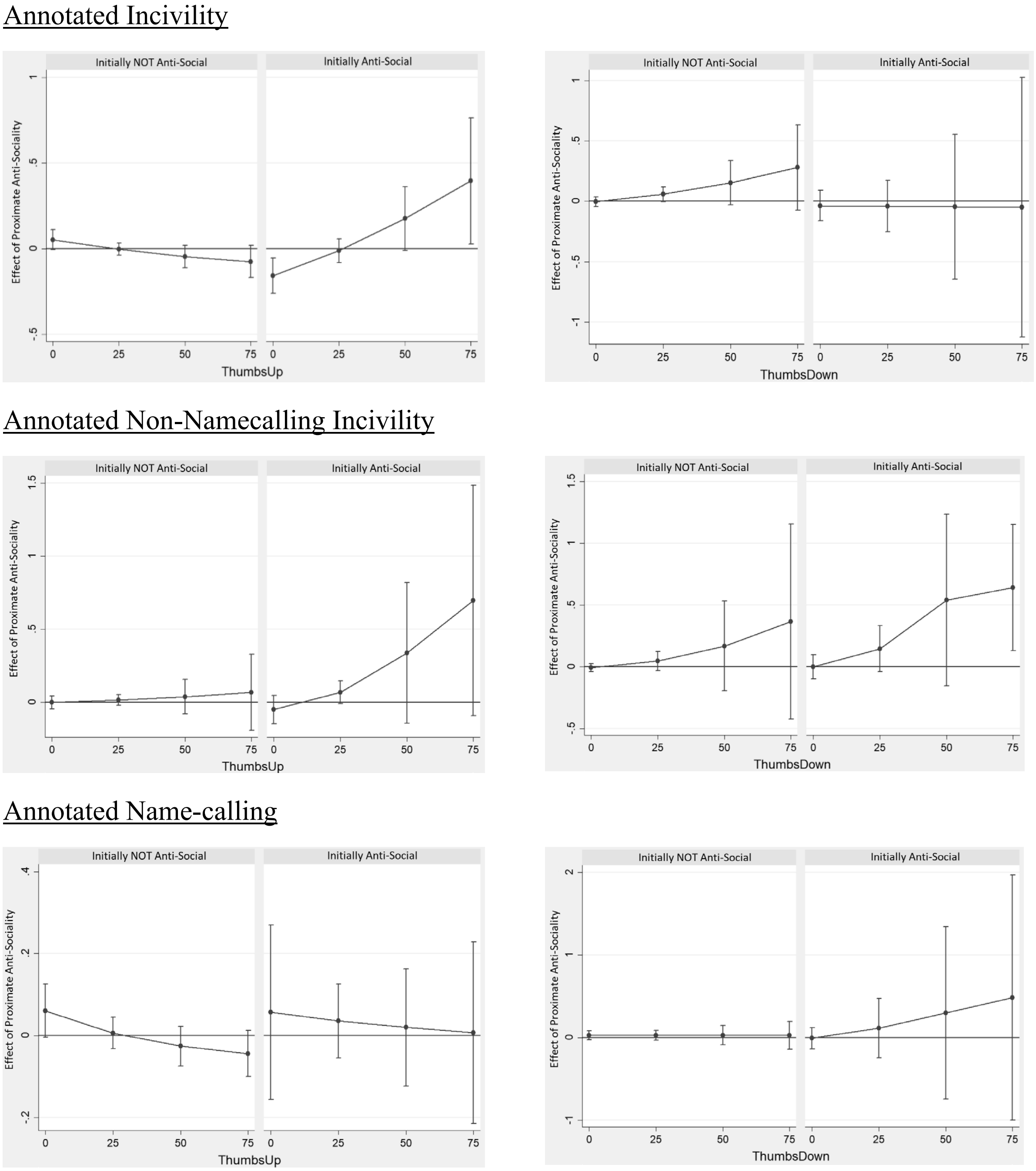


Figure 6

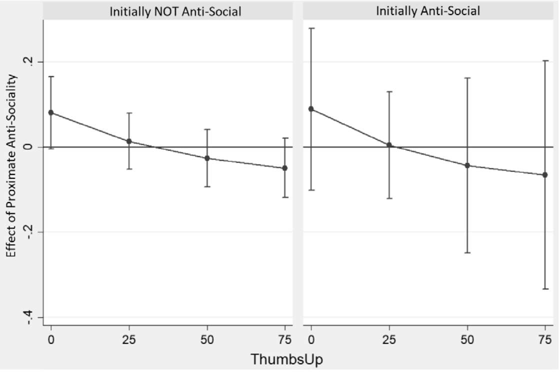
Testing TNSB with Human Annotation and Automated Classification of Incivility


Annotated Incivility

Annotated Non-Namecalling Incivility

Annotated Name-calling

Automated Name-calling


Initially NOT Anti-Social
Initially Anti-Social
Effect of Proximate Anti-Sociality
.2
0
-.2
-.4
0
25
50
75
0
25
50
75
ThumbsUp

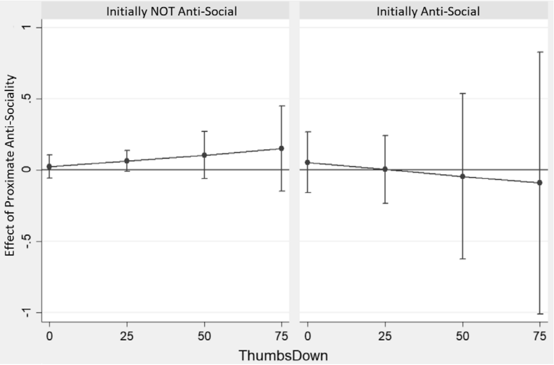
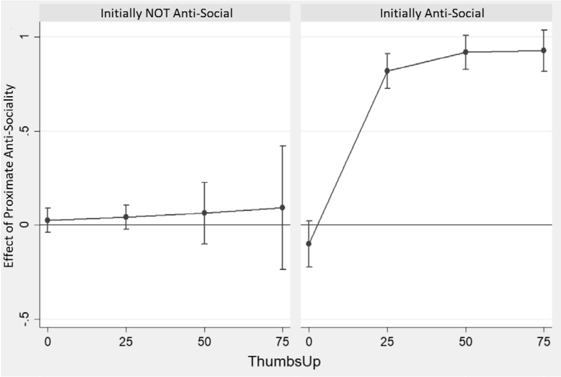
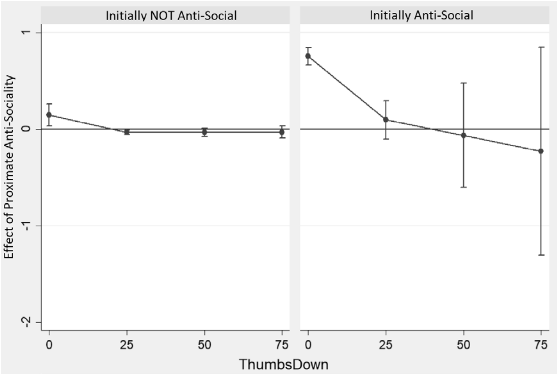

Initially NOT Anti-Social
Initially Anti-Social
Effect of Proximate Anti-Sociality
1
.5
0
-.5
-1
0
25
50
75
0
25
50
75
ThumbsDown

Figure 7

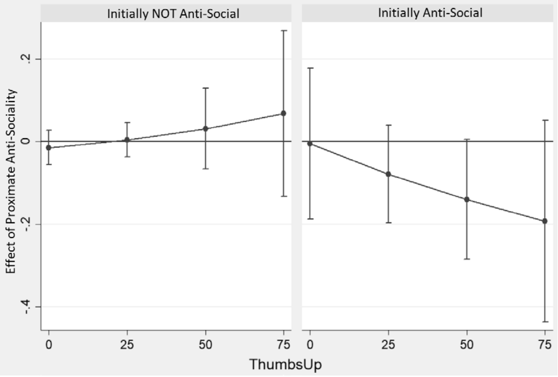
Testing TNSB with Google's Perspective API and New York Times Attributes

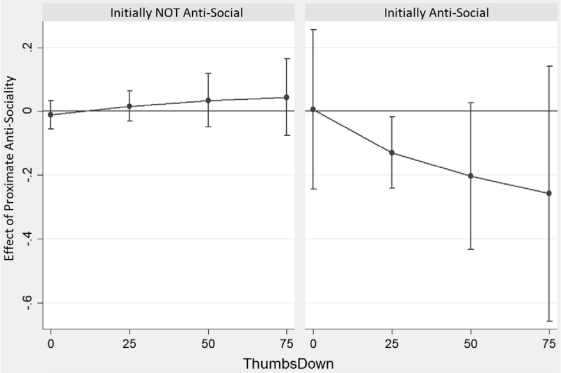
All Perspective (Toxicity, Severe Toxicity, Identity Attach, Insult, Profanity, or Threat)



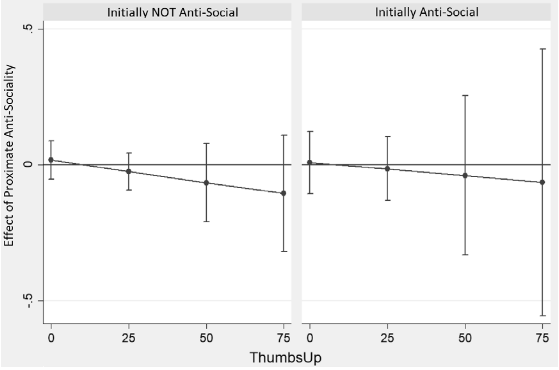


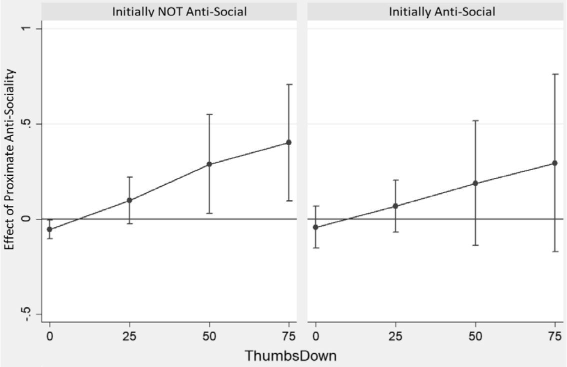
New York Times (Attack on Author)



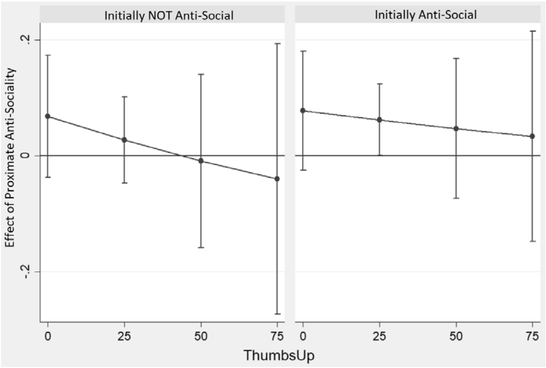


New York Times (Attack on Commenter)





New York Times (Inflammatory)


Initially NOT Anti-Social
Initially Anti-Social
Effect of Proximate Anti-Sociality
.2
0
-.2
0
25
50
75
0
25
50
75
ThumbsUp

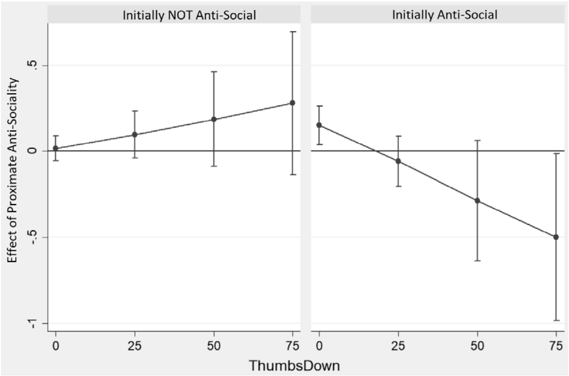
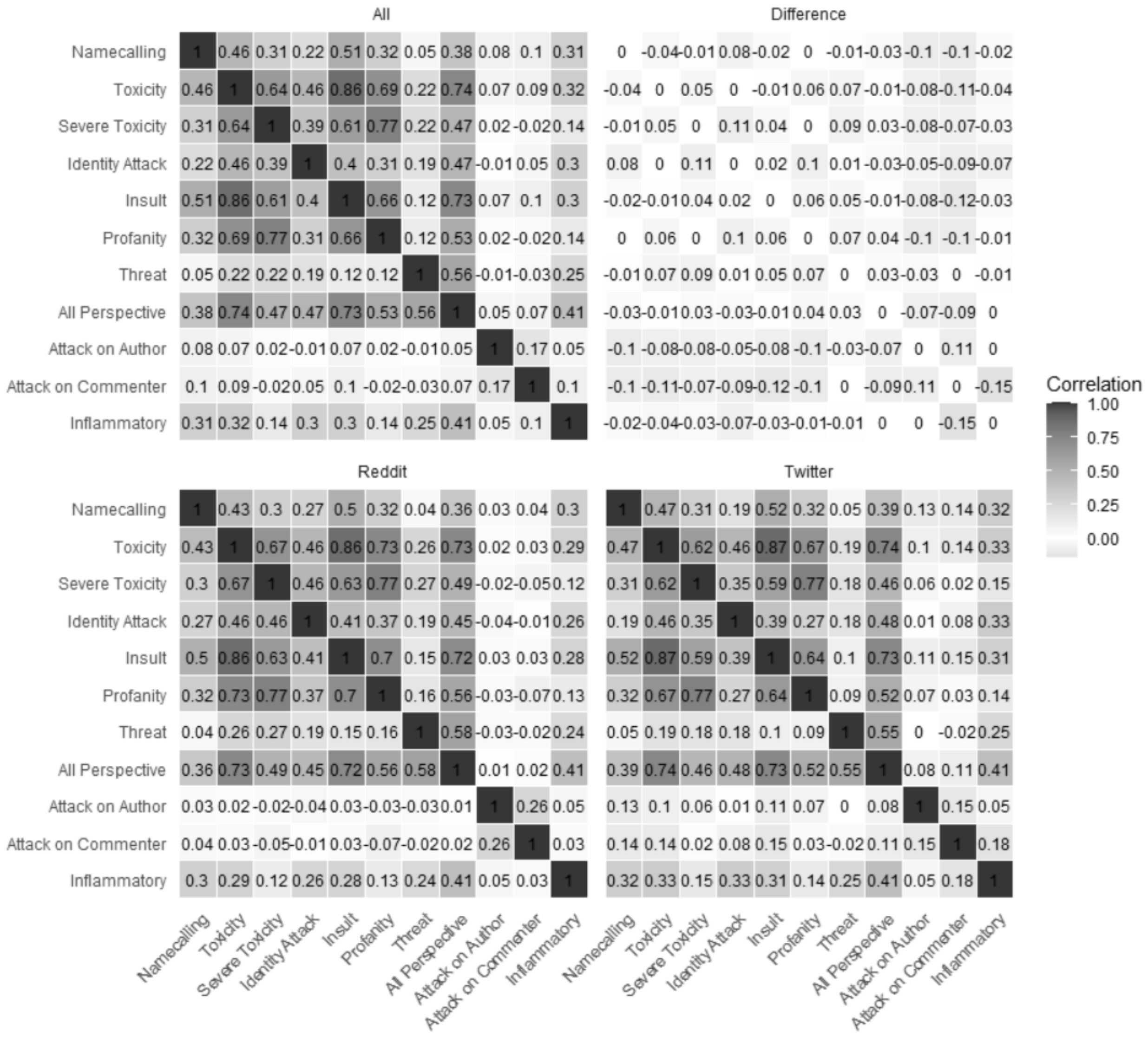

Initially NOT Anti-Social
Initially Anti-Social
Effect of Proximate Anti-Sociality
.5
0
-.5
-1
0
25
50
75
0
25
50
75
ThumbsDown

Figure 8

Correlation Matrices for Initial Comments on Reddit and Twitter



All

	Namecalling	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat	All Perspective	Attack on Author	Attack on Commenter	Inflammatory
Namecalling	1	0.46	0.31	0.22	0.51	0.32	0.05	0.38	0.08	0.1	0.31
Toxicity	0.46	1	0.64	0.46	0.86	0.69	0.22	0.74	0.07	0.09	0.32
Severe Toxicity	0.31	0.64	1	0.39	0.61	0.77	0.22	0.47	0.02	-0.02	0.14
Identity Attack	0.22	0.46	0.39	1	0.4	0.31	0.19	0.47	-0.01	0.05	0.3
Insult	0.51	0.86	0.61	0.4	1	0.66	0.12	0.73	0.07	0.1	0.3
Profanity	0.32	0.69	0.77	0.31	0.66	1	0.12	0.53	0.02	-0.02	0.14
Threat	0.05	0.22	0.22	0.19	0.12	0.12	1	0.56	-0.01	-0.03	0.25
All Perspective	0.38	0.74	0.47	0.47	0.73	0.53	0.56	1	0.05	0.07	0.41
Attack on Author	0.08	0.07	0.02	-0.01	0.07	0.02	-0.01	0.05	1	0.17	0.05
Attack on Commenter	0.1	0.09	-0.02	0.05	0.1	-0.02	-0.03	0.07	0.17	1	0.1
Inflammatory	0.31	0.32	0.14	0.3	0.3	0.14	0.25	0.41	0.05	0.1	1

Difference

	Namecalling	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat	All Perspective	Attack on Author	Attack on Commenter	Inflammatory
Namecalling	0	-0.04	-0.01	0.08	-0.02	0	-0.01	-0.03	-0.1	-0.1	-0.02
Toxicity	-0.04	0	0.05	0	-0.01	0.06	0.07	-0.01	-0.08	-0.11	-0.04
Severe Toxicity	-0.01	0.05	0	0.11	0.04	0	0.09	0.03	-0.08	-0.07	-0.03
Identity Attack	0.08	0	0.11	0	0.02	0.1	0.01	-0.03	-0.05	-0.09	-0.07
Insult	-0.02	-0.01	0.04	0.02	0	0.06	0.05	-0.01	-0.08	-0.12	-0.03
Profanity	0	0.06	0	0.1	0.06	0	0.07	0.04	-0.1	-0.1	-0.01
Threat	-0.01	0.07	0.09	0.01	0.05	0.07	0	0.03	-0.03	0	-0.01
All Perspective	-0.03	-0.01	0.03	-0.03	-0.01	0.04	0.03	0	-0.07	-0.09	0
Attack on Author	-0.1	-0.08	-0.08	-0.05	-0.08	-0.1	-0.03	-0.07	0	0.11	0
Attack on Commenter	-0.1	-0.11	-0.07	-0.09	-0.12	-0.1	0	-0.09	0.11	0	-0.15
Inflammatory	-0.02	-0.04	-0.03	-0.07	-0.03	-0.01	-0.01	0	0	-0.15	0

Reddit

	Namecalling	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat	All Perspective	Attack on Author	Attack on Commenter	Inflammatory
Namecalling	1	0.43	0.3	0.27	0.5	0.32	0.04	0.36	0.03	0.04	0.3
Toxicity	0.43	1	0.67	0.46	0.86	0.73	0.26	0.73	0.02	0.03	0.29
Severe Toxicity	0.3	0.67	1	0.46	0.63	0.77	0.27	0.49	-0.02	-0.05	0.12
Identity Attack	0.27	0.46	0.46	1	0.41	0.37	0.19	0.45	-0.04	-0.01	0.26
Insult	0.5	0.86	0.63	0.41	1	0.7	0.15	0.72	0.03	0.03	0.28
Profanity	0.32	0.73	0.77	0.37	0.7	1	0.16	0.56	-0.03	-0.07	0.13
Threat	0.04	0.26	0.27	0.19	0.15	0.16	1	0.58	-0.03	-0.02	0.24
All Perspective	0.36	0.73	0.49	0.45	0.72	0.56	0.58	1	0.01	0.02	0.41
Attack on Author	0.03	0.02	-0.02	-0.04	0.03	-0.03	-0.03	0.01	1	0.26	0.05
Attack on Commenter	0.04	0.03	-0.05	-0.01	0.03	-0.07	-0.02	0.02	0.26	1	0.03
Inflammatory	0.3	0.29	0.12	0.26	0.28	0.13	0.24	0.41	0.05	0.03	1

Twitter

	Namecalling	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat	All Perspective	Attack on Author	Attack on Commenter	Inflammatory
Namecalling	1	0.47	0.31	0.19	0.52	0.32	0.05	0.39	0.13	0.14	0.32
Toxicity	0.47	1	0.62	0.46	0.87	0.67	0.19	0.74	0.1	0.14	0.33
Severe Toxicity	0.31	0.62	1	0.35	0.59	0.77	0.18	0.46	0.06	0.02	0.15
Identity Attack	0.19	0.46	0.35	1	0.39	0.27	0.18	0.48	0.01	0.08	0.33
Insult	0.52	0.87	0.59	0.39	1	0.64	0.1	0.73	0.11	0.15	0.31
Profanity	0.32	0.67	0.77	0.27	0.64	1	0.09	0.52	0.07	0.03	0.14
Threat	0.05	0.19	0.18	0.18	0.1	0.09	1	0.55	0	-0.02	0.25
All Perspective	0.39	0.74	0.46	0.48	0.73	0.52	0.55	1	0.08	0.11	0.41
Attack on Author	0.13	0.1	0.06	0.01	0.11	0.07	0	0.08	1	0.15	0.05
Attack on Commenter	0.14	0.14	0.02	0.08	0.15	0.03	-0.02	0.11	0.15	1	0.18
Inflammatory	0.32	0.33	0.15	0.33	0.31	0.14	0.25	0.41	0.05	0.18	1

Figure 9

Effect of Anti-Social Commenting on Votes across Reddit and Twitter

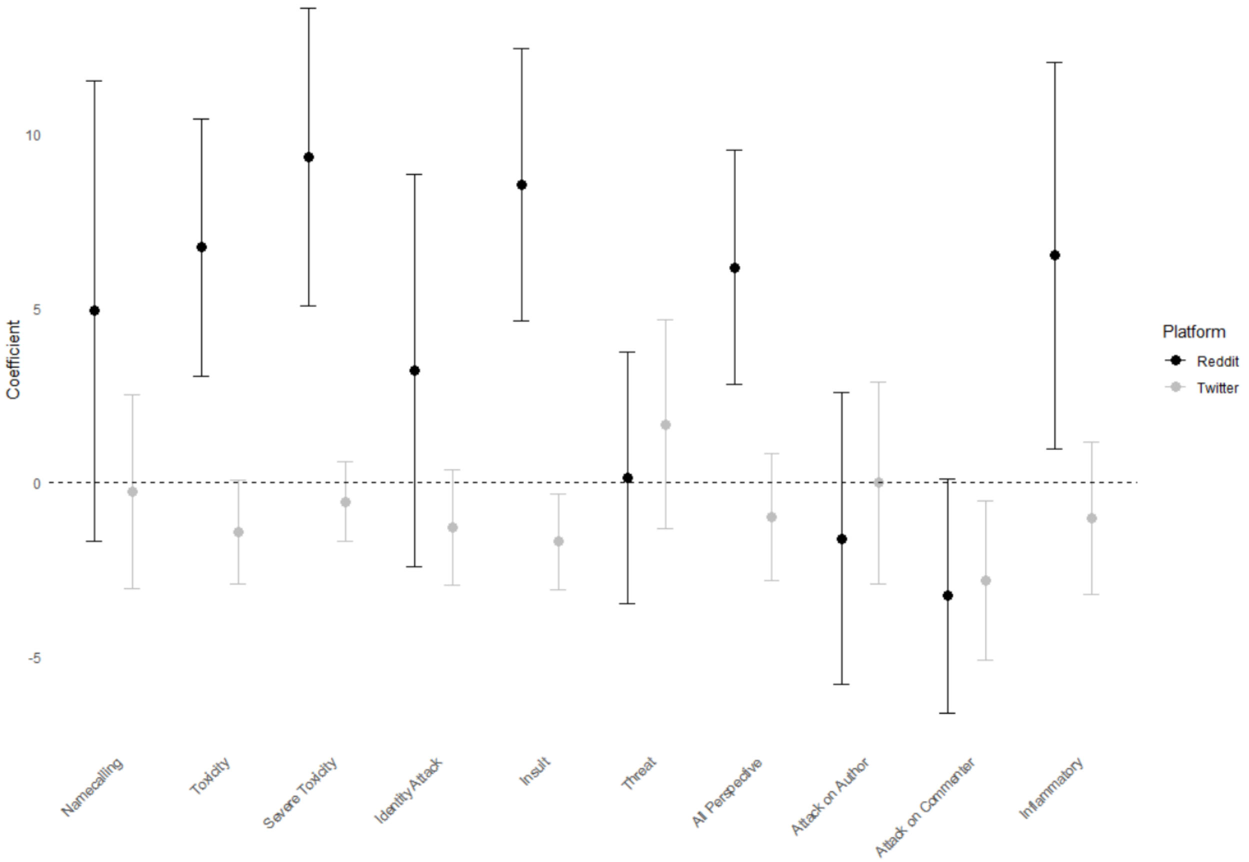
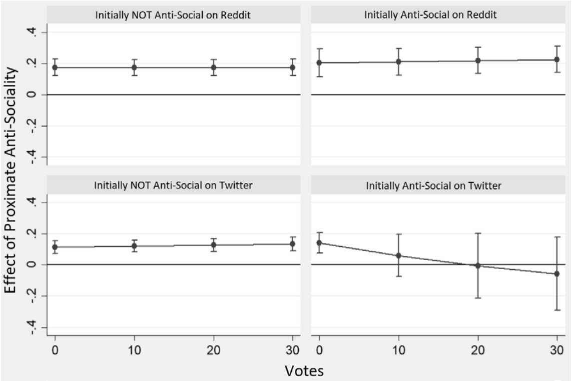


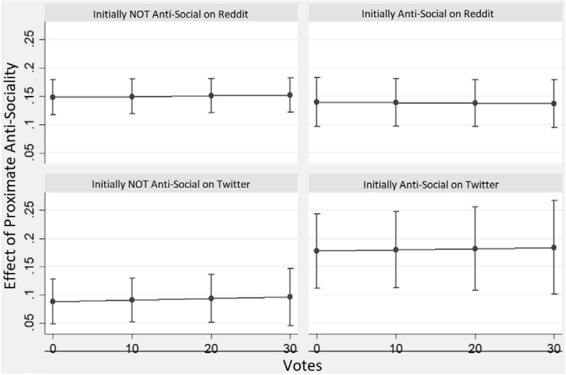
Figure 10

Testing TNSB on Reddit and Twitter During the January 6th Capitol Riots

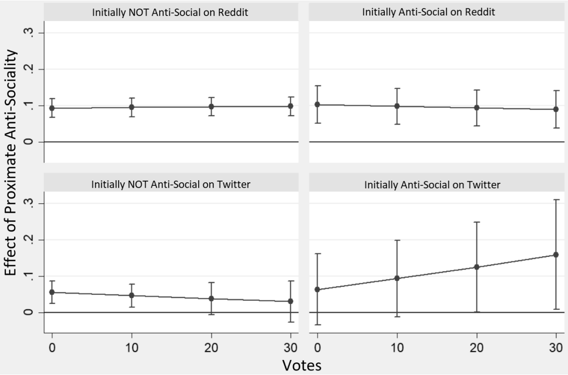
Automated Namecalling



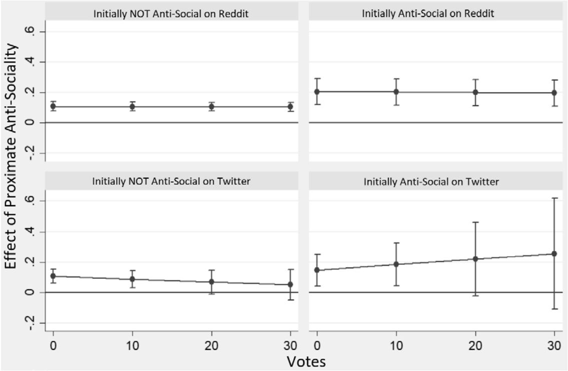
Perspective (Toxicity)



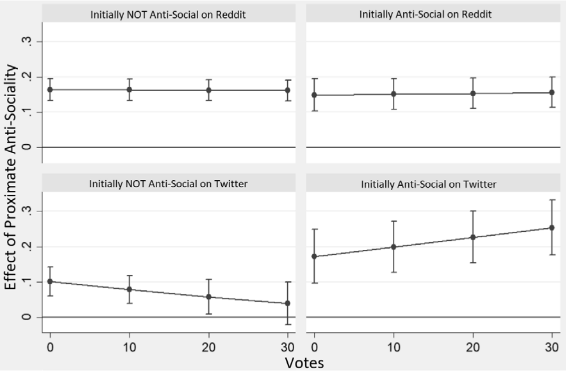
Perspective (Severe Toxicity)



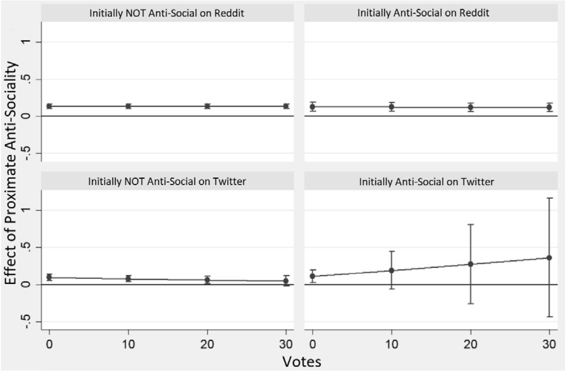
Perspective (Identity Attack)



Perspective (Insult)

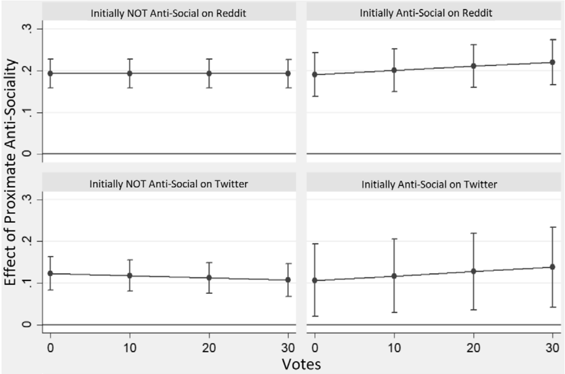


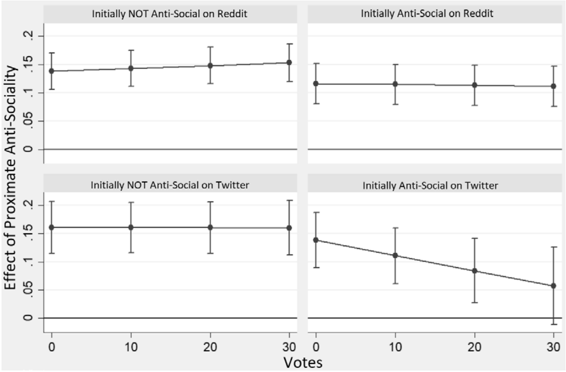
Perspective (Profanity)



Perspective (Threat)

New York Times (Inflammatory)





End Notes

[1] https://huggingface.co/civility-lab. Accessed November 22, 2023.

[2] More information on the attributes from the Perspective API, including those trained on the *New York Times* comments, can be found here: https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages. Accessed November 22, 2023.

[3] That the rate of anti-sociality in comments starting a triplet does not differ substantially from that of comments in general should not be viewed as evidence for a lack of social influence. This is because early comments will not necessarily start a triplet, as the triplets we constructed to have the particular structure that allows for a test of our framework (that is, they originate and end with the same commenter with a different commenter sandwiched between them). For example, if the identities of the contributors of a string of comments are 1234546, then there will only be one applicable triplet (454) and it will not be found at the beginning of the set.

[4] Upon further investigation, we discovered that the default input length limitations can be circumvented with additional code. However, a complete reanalysis of the data was not possible at the time of this writing.

[5] More information about OpenAI's API can be found here: https://platform.openai.com/overview. Accessed November 22, 2023.