# Polarization and Toxicity in Political Discourse Online

by

Martin Saveski

B.Sc., Staffordshire University (2010)
M.Sc., University Pierre and Marie Curie (2013)
M.Sc., Polytechnic University of Catalonia (2013)

Submitted to the Program in Media Arts and Sciences, in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Author ______________________________

Program in Media Arts and Sciences
August 17, 2020

Certified by ______________________________

Deb K. Roy
Professor
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by ______________________________

Tod Machover
Academic Head
Program in Media Arts and Sciences

# Polarization and Toxicity in Political Discourse Online

by Martin Saveski

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on August 17, 2020, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

## Abstract

The web and social media promised to fundamentally change the public sphere by democratizing access to information and lowering barriers for participation in public discourse. While some of these expectations have been met, we have also seen the negative effects of the web and social media, amplifying people's tendency to self-sort and polarize, and providing a platform for uncivil public discourse. In this thesis, we focus on two phenomena, toxicity and polarization in political discourse online.

In the first part of this thesis, we study media outlets' role in political polarization online, mainly, how the language they use to promote their content influences the political diversity of their audience. We track the engagement with tweets posted by media outlets over three years (556k tweets, 104M retweets) and model the relationship between the tweet text and the political diversity of the audience. We build a tool that integrates our model and helps journalists craft tweets engaging to a politically diverse audience, guided by the model predictions. To test the real-world impact of the tool, we partner with the PBS documentary series Frontline and run a series of advertising experiments on Twitter. We find that in five out of the seven experiments, the tweets selected by our model were indeed engaging to a more politically diverse audience, illustrating the effectiveness of our tool.

In the second part of this thesis, we study the relationship between the structure and the toxicity in political conversations on Twitter. We collect data on conversations prompted by tweets posted by news outlets and politicians running in the 2018 US midterm elections (1.18M conversations, 58.5M tweets). To investigate the link between structure and toxicity, we analyze the conversations at the individual, dyad, and group levels. We also consider two prediction tasks: (*i*) whether the conversation as a whole will become more or less toxic, and (*ii*) whether the next reply, posted by a specific user, will be toxic. We demonstrate that the structural characteristics of a conversation can be used to detect early signs of toxicity, both at the individual and the group level.

Thesis Supervisor: Deb K. Roy
Title: Professor, Program in Media Arts and Sciences

# Polarization and Toxicity in Political Discourse Online

by

Martin Saveski

This doctoral thesis has been reviewed and approved by the following committee members:

Deb K. Roy ________________________________________

Thesis Committee Chair<br>
Professor<br>
MIT Media Lab

Lada Adamic ________________________________________

Thesis Reader<br>
Director, Computational Social Science<br>
Facebook

Dean Eckles ________________________________________

Thesis Reader<br>
Associate Professor of Marketing<br>
MIT Sloan and MIT Institute for Data, Systems & Society

*Dedicated to Dedo Dule*

# ACKNOWLEDGMENTS

This thesis would not have been possible without the help and support of many people.

First, I would like to thank my advisor, Deb Roy. Deb has provided opportunities, guidance, and support throughout my Ph.D. Most of all, I would like to thank Deb for trusting me and giving me the freedom to work on topics that I found exciting while providing support and creating an environment where I had the best chance to succeed.

I want to thank my committee members, Lada Adamic and Dean Eckles, for their guidance throughout this process. I feel very fortunate to have had a chance to interact and collaborate with you during my Ph.D.

I want to thank my collaborators at Frontline: Katherine Griwert, Ben Abrams, Pam Johnston, and Raney Aronson-Rath. Thank you for inspiring us to think about the link between language and audience diversity, testing our tools and helping us improve them, and composing all the tweets that we ended up testing. Seeing your excitement about the project enabled me to keep going even when I felt discouraged and unmotivated.

The work presented in Chapter 2 was a collaboration with Doug Beeferman and David McClure. Doug's ability to break down a problem and ask the right questions is extraordinary. David is one of the most talented software engineers I have met, and peer-programming together was a real pleasure. It was a privilege to work with both of you. The work presented in Chapter 3 was a collaboration with Brandon Roy. Working with Brandon on that project was one of the most enjoyable experiences of my Ph.D. I will miss our endless conversations at the office and the time spent scribbling ideas on the board. I also want to thank Bridgit Mendler, who

first started exploring her own Twitter conversations and inspired Brandon and I to study Twitter conversations systematically.

I want to thank all past and present members of the Laboratory for Social Machines: Nabeel Gillani, Eric Chu, Prashanth Vijayaraghavan, Ivan Sysoev, Soroush Vosoughi, Anneli Hershman, Julina Nazare, Sneha Priscilla, Ann Yuan, Lauren Fratamico, Shayne O'Brien, Marc Exposito, William Brannon, Belen Saldias, Maggie Hughes, Bridgit Mendler, Nazmus Saquib, Alex Siegenfeld, Mina Soltangheis, Perng-hwa Kung, Sophie Chou, Neo Mohsenvand, Russell Stevens, William Powers, Andrew Heyward, Heather Pierce, Keyla Gomez. Special shout out to Eric Chu, for being the best officemate and hotel roommate; Nabeel Gillani, for his kindness and support; Soroush Vosoughi, for being the big brother and advising me in navigating the graduate program; Heather Pierce and Keyla Gomez, for brightening my days.

I would also like to thank my peers at the Media Lab who significantly contributed to my experience at the lab: Alejandro Noriega Campero, Abdullah Almaatouq, Xavi Benevides, Bianca Datta, Cristian Jara Figueroa, Natasha Jaques, Juliana Cherston, David Ramsay, Michiel Bakker, Yan Leng, and many others.

I did some of the most exciting research during my internships outside of the lab. I want to thank Ya Xu and Weitao Duan, who hosted me at LinkedIn, and Farshad Kooti, Carlos Diuk, and Lada Adamic, who hosted me at Facebook.

Throughout my Ph.D., I benefited from the financial and data support by Twitter. Most of the work in this thesis would not have been possible without the extended access to Twitter data.

I was very fortunate to meet a wonderful group of friends when I first arrived at MIT, many of whom have stayed my friends ever since. Special thanks to my WHPs: Lukas Murmann, Alba Luengo, David Alvarez Melis, Judith Amores, Tal Wagner, Sirma Orguc, Yunus Terzioğlu, Lea Verou, Chris Lilley, Thrasyvoulos Karydis, Anastasia Grigoropoulou, Andrew Kirby, Tugce Yazıcıgil Kirby, Viirj Kan, Alexandros Charidis, and Valerio Varricchio. Thank you for all the great memories. It would not have been the same without you.

I want to thank Edmond Awad and Peter Beshai for adding so much joy to my life and always being there for me. Thank you, Habibis!

I would like to thank my girlfriend, Lisa. I am immensely grateful for her support and patience over the last two and a half years.

Lastly, I want to thank my parents, Mirjana and Vlado Saveski. My gratitude for your love and support is simply immeasurable. Knowing that I can always count on you has given me the courage and confidence to take on any challenge in life. Fala vi!

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1 INTRODUCTION

In the imagination of many, the web and social media had the potential to realize Habermas's idealized vision of the public sphere, a place where people come together to discuss the news of the day, form public opinion, and hold the state accountable. Unlike English coffee houses, the web is accessible from anywhere and to anyone, not just to the privileged few. While the web and social media have made it easier to access information and engage with one another, we have also started to see their negative effects. Two phenomena, in particular, have attracted much attention, especially in the wake of the 2016 presidential election in the US: political polarization and antisocial behavior online.

Studies of political polarization have traditionally focused on policy preferences. There is overwhelming evidence that political elites in the US are getting more polarized over the last four decades [58, 65], but lack of consensus on whether the general public is more ideologically polarized [29, 42]. More recently, scholars have started studying polarization in terms of affect—feeling positive sentiment for one's own group and negative sentiment toward those identifying with opposing groups—instead of ideology [45, 46]. While partisans over the last 30 years consistently give enthusiastic ratings to their own party, both Democrats and Republicans report that they like the members of the other party less and less [45]. Many attribute this increase in animosity, at least in part, to the web and social media, blaming them for two drivers of polarization: (1) *echo chambers*, making it easier for individuals to be exposed only to information from like-minded individuals [89], and (2) *filter bubbles*, algorithmic content curation based on users' past behavior giving more visibility to content that confirms the users' worldviews [75]. In this thesis, we focus on one aspect of political polarization on social media: the role that news outlets play in the process. In particular, we are interested in how the language they use to promote their content affects the political diversity of their audience. As more than two-thirds of Americans get at least some of their news on social media [63], it is increasingly important to understand how the outlets' framing and presentation of the news on social media influences who engages with them.

Antisocial behavior is another phenomenon that hinders the potential of social media to support rich and vibrant public discourse. Antisocial behavior is an umbrella term that includes trolling, bullying, and harassment. Surveys suggest that these behaviors are very prevalent: 66% of Americans report that they have witnessed harassment online, and 41% say they have personally experienced it [27]. These behaviors are often exacerbated by the fact that people tend to be less inhibited in their online interactions [87]. Early studies argued that people engaging in antisocial behaviors online have unique personality traits [12, 79] and motivations [4, 40, 84]. However, more recent work shows that situational factors, such as the individual's mood or the surrounding context of a discussion, can trigger antisocial behaviors [17]. This suggests that even ordinary people can exhibit these behaviors under the right circumstances. In this thesis, we focus on toxicity in political conversations, rude and disrespectful comments that may make users leave the discussion. More specifically, we are interested in the social conditions that are more likely to lead to toxic behaviors. We posit that the social structure in which the conversation participants are embedded affects their behavior. Toxicity can impede the healthy discussion that is at the core of the democratic process, and understanding which factors lead to toxic behaviors is essential.

Next, we describe our approach to studying polarization and toxicity in political discourse online.

## LANGUAGE AND POLITICAL POLARIZATION

In the first part of this thesis, we study media outlets' role in political polarization online. In particular, how the language they use to promote their content online influences the political diversity of their audience. Beyond analyzing the relationship between language and audience diversity, we partner with journalists from the documentary series Frontline and apply our models to help them reach a more politically diverse audience on Twitter. Frontline is a world-renowned investigative journalism program that produces in-depth documentaries on various domestic and international issues. The series has been on the air since 1983 and has won every major journalism and broadcasting award, including 93 Emmy Awards and 24 Peabody Awards. As a PBS (Public Broadcasting Service) documentary series, their goal is not just to maximize engagement with their content, but also to reach as wide an audience as possible, across the political spectrum.

To study the relationship between language and audience diversity, we tracked all tweets posted by Frontline and five major news outlets that span the full political spectrum (New York Times, CNN, Wall Street Journal, Fox News, and Breitbart) over three years, collecting over 566K tweets and 104M retweets. To measure the political diversity of the audience of each tweet, we consider the users who retweeted the tweet, and then calculate their political alignments in terms of how often they share content from left- and right-leaning websites.

We use this data to model the relationship between the tweet text and the political diversity of the audience. Recent advances in deep learning have revolutionized the field of natural language processing, obtaining extraordinary results on a wide range of tasks, from question answering to general language understanding. We apply these state-of-the-art techniques to train machine learning models that, given an input text, accurately predict the expected audience diversity.

We integrate the prediction models into a web application that allows users to input tweet drafts and get instant predictions of the expected audience diversity. The application's goal is to supplement the journalists' writing process by allowing them to iterate on the tweet text based on the model predictions and help them craft tweets that are more likely to reach a diverse audience. In addition to the model predictions, we also highlight relevant words in the tweet drafts and surface semantically similar historical tweets that were engaging to a diverse audience.

Finally, together with Frontline, we run a series of advertising experiments on Twitter to test whether our models can be effectively used to select tweets that are engaging to a more politically diverse audience. In each experiment, we select a pair of tweets—one predicted to be engaging to a politically diverse audience, and another predicted to be engaging to a more homogeneous group of users—and measure the engagement of left- and right-leaning users with each tweet.

## THE STRUCTURE OF TOXIC CONVERSATIONS

In the second part of this thesis, we investigate the relationship between structure and toxicity of political conversations on Twitter. We were motivated by the simple idea that communication is a social act and that the relationships between conversation participants will influence their behaviors. The goal of this study is twofold: (*i*) to understand the relationship between the conversational structure and toxicity after the conversation has unfolded, and (*ii*) to evaluate the predictive value of the structural view of the conversations in forecasting future toxicity as the conversations unfold.

We tracked the conversations prompted by tweets posted by five major news outlets (New York Times, CNN, Wall Street Journal, Fox News, and Breitbart) over one year, and 1,430 politicians who ran for office in the 2018 US midterm elections over four months. We collected more than 1.18M conversations containing >58.5M tweets posted by >4.4M users. Using a machine learning model, we annotated the conversation tweets as toxic or non-toxic: we considered a tweet to be toxic if it is a rude, disrespectful, or unreasonable comment that may make users leave a discussion.

To capture the social and conversational structure of the conversations, we describe each conversation using three different representations: (*i*) *reply tree*, which encodes the relationships between individual replies, where two tweets are connected to each other if one was posted in reply to the other, (*ii*) *reply graph*, which encodes the interactions between users, where one user is connected to another if they replied to one of their tweets, (*iii*) *follow graph*, which encodes the social relationships among the users, where one user is connected to another if they follow them on Twitter.

To study the relationship between the conversational structure and toxicity after the conversation is over, we analyze the conversations at three levels: individual, dyadic, and group level. At the individual level, we analyze the users' behavior across many conversations; at the dyad level, we investigate how the probability of a toxic reply varies depending on the relationship between the two conversation participants; and at the group level, we study how the social and conversational structure among the conversation participants influences the overall toxicity of the conversation.

To test the utility of the structural representation of the conversations in forecasting toxicity, we consider two prediction tasks. In the first task, we aim to predict whether the conversation will become more or less toxic than expected, given the initial stages of the conversation. In the second task, we attempt to predict whether the next reply posted by a specific user will be toxic, given the conversation so far and the user's relationship with the current conversation participants.

Through this study, we make both significant theoretical contributions, advancing our understanding of the social factors that contribute to toxicity online; and practical contributions, demonstrating that models based on the structural characteristics of a conversation can be used to detect early signs of conversation derailment and, potentially, steer conversations in a less toxic direction.

## THESIS OUTLINE

The remainder of this thesis is organized as follows:

**CHAPTER 2:** We study the relationship between language and political polarization.

**CHAPTER 3:** We investigate the relationship between the structure and toxicity in political conversations.

**CHAPTER 4:** We examine how the two phenomena that are the main subject of this thesis—polarization and toxicity—relate to each other.

**CHAPTER 5:** In the final chapter, we summarize our contributions and propose future directions for research.

# 2 LANGUAGE AND POLITICAL POLARIZATION

## 2.1 INTRODUCTION

News outlets have the potential to bring people with different political views together and either create a shared reality or reinforce the existing political divides. With more than two-thirds of Americans using social media as their primary news source, how news outlets present their news online has an enormous impact on who engages with them. In this chapter, we investigate the role that news outlets play in political polarization online and, in particular, how the language they use to promote their content online influences the political diversity of their audience.

We track the tweets posted by five news outlets over three years and measure the political diversity of the users who engage with them. Based on this data, we model the relationship between the tweet text and the audiences' political diversity. To test our models in the real world, we partner with the documentary series Frontline. Like other programs, Frontline uses social media to promote their films. However, as a PBS program, their goal is not just to maximize engagement, but also to reach a politically diverse audience. We build a web application that integrates our models and allows Frontline's journalists to craft more bridging tweets, guided by the model predictions.

To test whether the model predictions can be effectively used to compose more bridging tweets, we run a series of advertising experiments promoting Frontline's tweets. In each experiment, we select a pair of tweets—one predicted to be more bridging, and another predicted to be less bridging—and measure the engagement of left- and right-leaning users with each tweet.

The rest of this chapter is organized as follows. In Section 2.2, we review previous work relevant to the current study. We describe the data collection process in Section 2.3 and evaluate the quality of our estimates of the users' political alignment in Section 2.4. We define the prediction task and evaluate different models in Section 2.5, and describe the web application that allows journalists to interact with the models in Section 2.6. In Section 2.7, we explain the setup and discuss the results of the advertising experiments we ran in partnership with Frontline. Finally, we conclude and summarize our findings in Section 2.8.

## 2.2 RELATED WORK

We start by reviewing papers that are most closely related to different aspects of this part of the thesis, namely: (*i*) studies that examine the relationship between language and virality, (*ii*) the differences in language used by different groups when discussing the same topics/issues, (*iii*) the challenges of running randomized experiments using advertising campaigns.

**PREDICTING VIRALITY.** The problem of predicting the political diversity of the audience based on the language used in a post is most related to past work on predicting the popularity/virality of a post. Since the early days of social media, there have been many studies that aim at predicting the popularity of a post before it is posted, using features related to the post language, poster characteristics, and posting time [2, 38, 44, 48, 78].

As recent work [62] has highlighted, it is challenging to systematically compare these studies since many differ in the experimental setup and outcome measures. While most studies report that it is possible, at least to some extent, to predict popularity *a priori*, recent evidence suggests that there is a limit to the prediction accuracy, even if infinite amounts of data are available [62]. Here, we highlight several papers that have focused on the content characteristics and have studied this question most rigorously.

Berger and Milkman [9] study how emotion in online content influences virality. They collect 7,000 New York Times articles and compare the emotions evoked by those that made it to the most-emailed list vs. those that did not. They find that more positive and more negative content is more viral than content that does not evoke any emotion and that positive content is more viral than negative content. Their results hold even after controlling for how prominently the article was displayed on the web and printed version of the newspaper, when it was published, and who it was written by. Looking at specific emotions, they find that virality is driven by more than just valence: content that evokes high-arousal emotions (awe, anger, anxiety) is more popular regardless of the valence. To confirm that high-arousal content indeed causes virality, they run a series of randomized lab experiments. They show subjects (*a*) high vs. low amusement, (*b*) high vs. low anger, (*c*) high vs. low sadness version of a story, and ask them how likely they are to share the story with others. The results confirm their observational analysis: content that evoked high-arousal (high amusement, high anger) is more likely to be shared, while content that evoked more of a deactivating emotion (high sadness) is less likely to be shared.

Reis et al. [26] provide some more evidence for the findings by Berger and Milkman. They analyze 70k headlines produced by four global media outlets and study how sentiment relates to popularity. They find that the majority of the news, across the four outlets, have negative sentiment and that the strength of the sentiment correlates with popularity—extremely positive or extremely negative headlines attract more attention than more neutral ones.

Tan et al. [90] study the effect of wording tweets on their virality, i.e., how often they are retweeted. To control for author and topic effects, they collect pairs of tweets posted by the same user and containing the same URL, but with different wording. They analyze the linguistic differences between the successful (i.e., more retweeted) and unsuccessful tweets within each pair and find that tweets that are retweeted more tend to be longer (where length is a proxy for informativeness), use language similar to the user's previous tweets, conform to the poster's followers' expectations (measured by the user's and user's followers' language models) and read like news headlines (using New York Times headlines as a reference). More retweeted tweets also tend to be more general (use indefinite articles more often), more readable, and contain more emotional (both positive and negative) words (corroborating the findings in [9]). They also test how good human raters are at predicting which of the two tweets is more retweeted and find that they achieve an average accuracy of 61.3%, better than random guessing, but far from perfect. Using a logistic regression model with the custom linguistic features (most described above), they were able to outperform the human raters, achieving 63% accuracy. Finally, combining the custom features with the bag-of-words representation of the text unigrams and bigrams, they were able to achieve even higher performance of 66% accuracy.

Gligorić et al. [32] study how the brevity of a message affects its popularity. They take a sample of 60 long tweets (250 characters long) and ask crowd workers to shorten them to 9 different lengths while preserving the meaning of the original message. Then, they show another group of crowd workers pairs of tweets, the original tweet and a shortened one, and ask them which one they think will get more retweets. They find that there are significant benefits of brevity: the concise versions of the tweets were, on average, more successful than the original tweets and the optimal reduction is between 10%-20%. Comparing the language of the original tweets and the more concise versions, they find that verbs and negations (part of speech that carries essential information) are disproportionally more likely to be preserved. They also find that words that describe affect are much more likely to be preserved and that the effect is much stronger for negative (anger, sadness, anxiety) than positive emotions.

**subgroup receptivity.** Demszky et al. [24] analyze the relationship between language and political polarization in tweets related to 21 mass shootings in the US. They study four aspects: topic choice, framing, affect, and illocutionary force. They classify each user who tweeted about one of these events as a Democrat or a Republican (based on whom they follow) and study the difference in the language used by each group. They cluster tweets by topic and find that the most polarizing topics, across events, are related to the shooter's identity & ideology (more discussed among Republicans) and laws & policy (more discussed among Democrats). Looking at the use of specific terms, they find a strong relationship between polarization and the race of the shooter, e.g., the term "terrorist" is more likely to be used by Democrats when the shooter is white and more used by Republicans when the shooter is a person of color. Analyzing the affect of the language, they find that positive sentiment, sadness, and trust are more likely to be expressed by Democrats, while fear and disgust are more likely to be expressed by Republicans. They also find that in the aftermath of a tragic event, Democrats are much more likely to use phrases associated with illocutionary force (should, must, have to, and need to), reflecting on what should have happened or what should happen to prevent such events.

Lakkaraju et al. [57] study how the interaction between the language used to present the content and the target audience affects the popularity of the content. They analyze photos posted multiple times on Reddit, on different communities (subreddits), with different titles. This allows them to disentangle the photo's intrinsic quality from other factors, such as the title and the community. They propose a predictive model of popularity that has two components: (*i*) a community component capturing when (time of day), how many times, and to which communities the photo was previously submitted, and (*ii*) a language component that captures the quality of the title. They find that the community model alone has higher predictive power than the language model alone ($R^2 = 0.56$ vs. $R^2 = 0.14$) and that combining the two leads to the best performance

($R^2 = 0.64$). Looking at the language model, they find that popular titles use language that is familiar to the community (relative to other posts in the same subreddit) but is different from the titles accompanying the photo in previous posts. They also find that nouns and adjectives impact the success of a title more than verbs and adverbs.

**ADS AS EXPERIMENTS.** Eckles et al. [28] highlight the challenges of running field experiments using social media advertising platforms. They argue that using the standard features of an online advertising platform to compare how different versions of an advertisement perform does not create a randomized experiment. Their key argument is that users are not randomly assigned to different versions of the advertisement. In fact, advertising platforms optimize the campaign performance by showing advertisements to users who are more likely to fulfill the campaign objective (e.g., clicks, app installs, etc.) and the groups that end up being exposed to each version of the advertisement are not comparable. Moreover, users may be shown both the treatment and control versions of an advertisement, and users in one condition (campaign) may be shown the advertisement more times than users in another condition. To illustrate their point, they analyze the data from studies by Matz et al. [64], which uses Facebook advertisements to test the effectiveness of psychologically targeted messages. Their analysis shows a severe imbalance of the user characteristics among the users who were exposed to different advertisements, thus calling into question the internal validity of the experiments.

An alternative to running experiments using an ad platform is running a survey on a crowdsourcing platform and asking workers which messages they are more likely to engage with or which messages they believe are more likely to be retweeted (similar to [32]). Recent empirical evidence suggests that self-reported willingness to share information indeed correlates with actual sharing behavior on Twitter [69]. It is worth noting that this kind of experiment is more likely to have lower external validity.

## 2.3 DATA

### 2.3.1 Data Collection

To systematically study the relationship between language and audience diversity, we collect a large number of tweets posted by news outlets and data related to the users who engaged with those tweets. We use the news outlets' tweets to characterize the language and the user data to characterize the audience and their political alignment in particular.

We tracked the tweets of five major news outlets and Frontline over the course of three years and three months, from Jan 2017 to March 2020. We selected the New York Times, CNN, Wall Street Journal, Fox News, and Breitbart as they have large followings on Twitter, their tweets consistently receive a lot of engagement, and together they cover the full political spectrum [6, 13]. We collected all tweets posted by the outlets and all of their retweets. We used the Twitter PowerTrack (also known as the Firehose) to capture the data in real-time between May 2018 and March 2020, and we ran batch jobs using the Twitter Historical PowerTrack to collect past tweets published between Jan 2017 and May 2018. Due to the limit of the number of tweets that we could ingest per month, we were unable to consider a larger set of outlets.

Figure 2.1 shows the daily volume of tweets per outlet over the data collection period. In total, we collected 566k tweets and the corresponding 104M retweets. We note the drop in the volume of tweets posted by Fox News after Nov 8, 2019, when they stopped tweeting in protest against Twitter after a group of demonstrators posted the home address of Tucker Carlson, one of the network's show hosts. To avoid any bias due to data censoring, we excluded from the analysis tweets posted over the last week of the data collection period, but counted the new retweets of tweets posted prior to that. Figure 2.2 shows the total number of tweets per outlet.

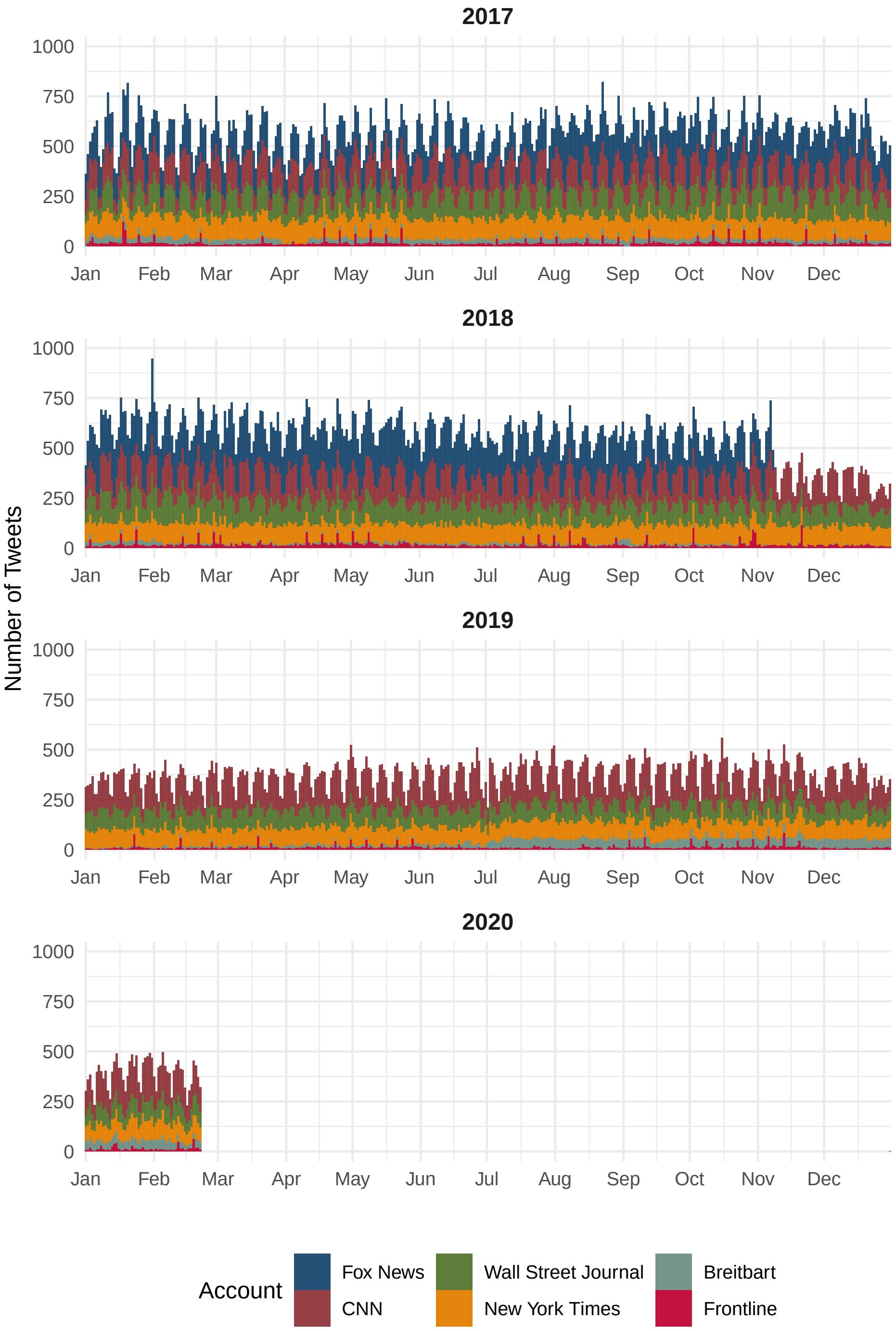


**Figure 2.1:** Daily volume of tweets posted by each account over the full data collection period.

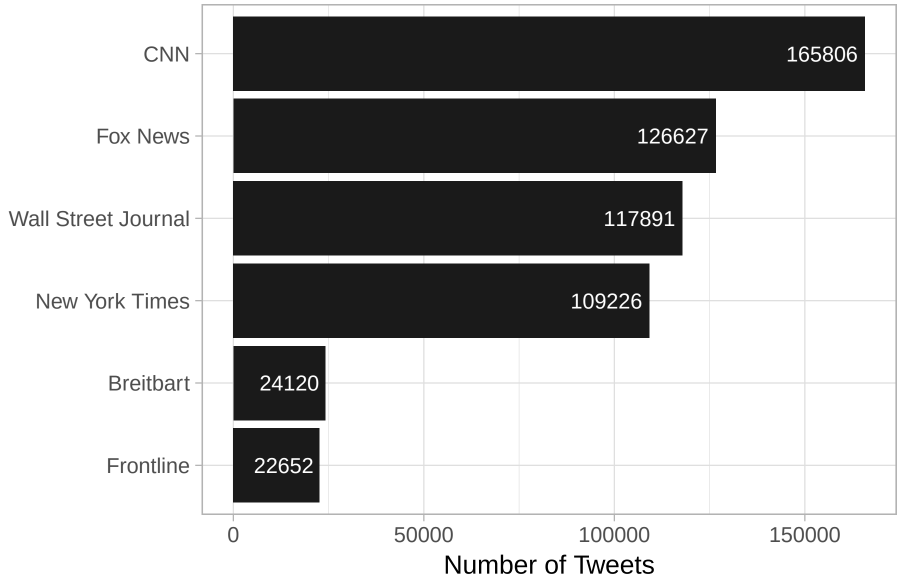


**Figure 2.2:** Number of tweets per account.

To characterize the users who engaged with each tweet, we compute summary statistics of the political alignments of the users who retweeted the tweet. To measure the political diversity of the audience, we compute the entropy of the retweeters' (discretized) political alignments, and to measure the overall alignment of the audience, we compute the mean of their (numerical) political alignments. We consider only retweets as they are a clear sign of agreement and endorsement of the content. We decided to exclude quote tweets, i.e., retweets with a commentary, which can be used to express disagreement with the original tweet. To ensure that we have a good estimate of the tweets' audience characteristics, we filter out tweets with less than three retweeters whose political alignment we could estimate. Figure 2.3 shows the distribution of average alignment and entropy per tweet for each outlet.

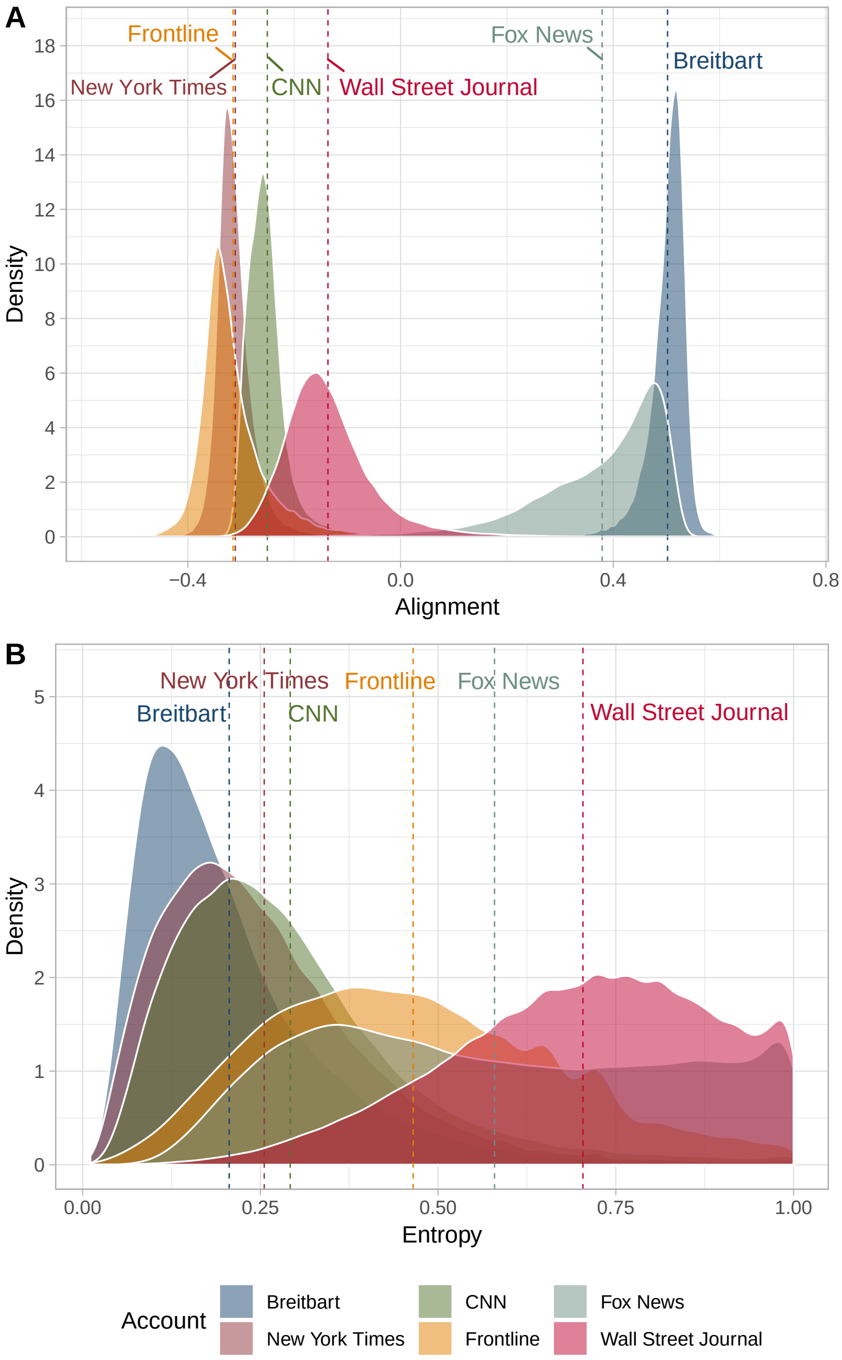


**Figure 2.3:** Distribution of political alignment (A) and entropy (B) per tweet for each account. The vertical dashed lines represent the average alignment/entropy per outlet. The tweet alignment scores are computed as the average of the retweeters alignment scores, and the entropy is computed on the distribution of left- vs. right-leaning retweeters. The user alignments scores and classifications were estimated based on the retweets' media-sharing patterns.

### 2.3.2 Measuring User Alignment

To measure the users' political alignments, we analyze the links that they share in their tweets and retweets. We build on previous work by Bakshy et al. [5, 6] which demonstrates that left- and right-leaning users share significantly different content. Based on their analysis—grounded in self-reported political leaning of the users—they released the political alignment of the 500 most shared domains with scores ranging from -1 (left-leaning) to +1 (right-leaning).

To assign a political alignment score to each user, we take the URLs of the content that they tweeted, look up the political alignment of each URL domain, and take the average. To obtain a binary classification for each user, we threshold the average alignment score at zero, classifying users with negative average alignment score as left-leaning and users with positive average alignment scores as right-leaning.

To find tweets posted by the users, we use a 3-year snapshot of the Twitter Decahose, which includes a 10% sample of all public tweets posted between January 2017 and December 2019. We use the Decahose snapshot instead of retrieving the user's tweets through the Twitter REST API for two reasons. First, the Rest API limits the number of user tweets we can retrieve to the 3,200 most recent ones. This would prohibit us from getting a longitudinal view of the users' tweeting behavior, especially for active users. A second more practical reason is that the REST API returns only the shortened URLs of the links included in the users' tweets, which we would have to expand in order to match them with the domain alignments by Bakshy et al. The Decahose, on the other hand, provides the expanded URLs that we can readily use.

Beyond using the Decahose to calculate user alignment scores, we also used it to expand the set of news outlets we consider in our analysis. However, as we detail in Section 2.5.6, including this additional data did not lead to more accurate models.

## 2.4 POLITICAL ALIGNMENT SCORE EVALUATION

Since we heavily rely on the estimated political alignment scores in rest of the analysis, in this section, we evaluate the quality of the alignments in three different ways: (*i*) we compare the breakdown of left- vs. right-leaning users for each US state against the proportion of Republican votes in 2016 and 2018 elections, (*ii*) we compare the alignments obtained using the users' media-sharing patterns against the estimates of another method that relies on the users' follow relationships, and (*iii*) we run a survey on Amazon Mechanical Turk where we ask a group of left- and right-leaning users whether they would share sample tweets, and compare their responses against the estimated alignment scores of the Twitter users who actually retweeted the sample tweets.

### 2.4.1 Comparison with the Share of Two-Party Vote per State

To validate our user ideology classifications, we compare the proportion of users classified as right-leaning for each US state against the Republican share of the two-party vote in the 2016 presidential and 2018 midterm elections. We use the same methodology as several previous studies [7, 8, 24].

To infer the users' states, we used the location field in their profile information. Since the field allows any text input, we used a few simple heuristics to extract the user's state: (*i*) we check whether the location field ends with a state abbreviation (e.g., "Cambridge, MA"), (*ii*) we search for a full state name anywhere in the field, (*iii*) if we find matches in both cases, rule *i* takes precedence, which avoids errors in cases such as: "Washington, DC" and "Kansas City, MO". We designed these rules to be high precision and not necessarily high recall.

We used all users in the Decahose whose location we could infer and for whom we had enough data to estimate their political alignment. This sample consisted of 2.3M users. To calculate the proportions of Republican

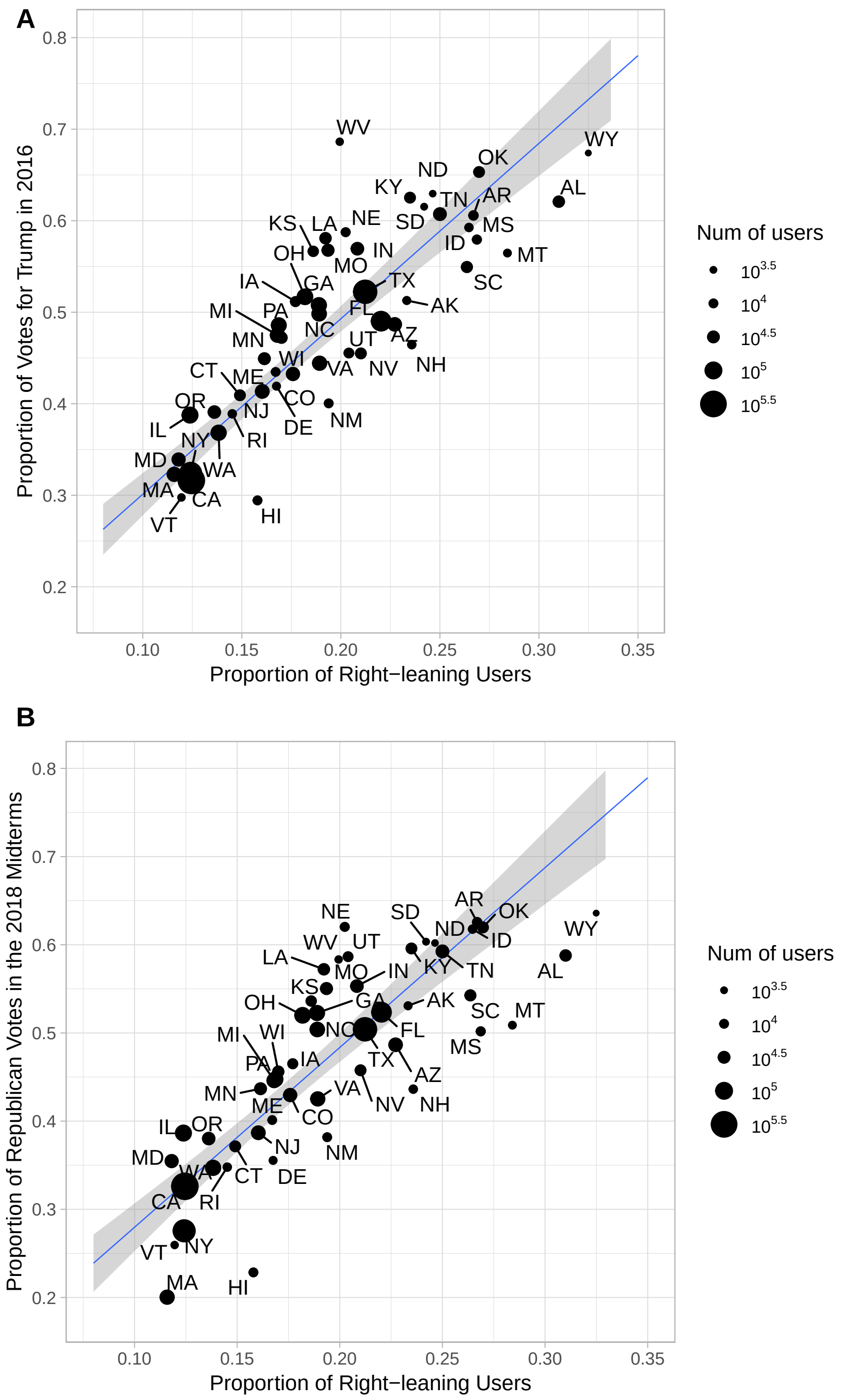


**Figure 2.4:** Comparison of the proportion of right-leaning users in our dataset for each US state against the proportion of right-leaning votes in the 2016 presidential elections (A) and the 2018 midterm elections (B). The lines are fitted using linear regression weighted by the number of users per state (A: $R^2 = 0.81$, B: $R^2 = 0.78$).

votes, we relied on the election results data curated by the MIT Elections lab [22, 23]. In the case of the 2018 midterm elections, we consider only the results of the elections for the House of Representatives, as only a third of the Senate seats were on the ballot.

Figures 2.4 show the results. We observe a strong correlation between the fraction of users classified as right-leaning in our dataset and both the proportion of votes for Donald Trump in 2016 ($R^2 = 0.81$ in a weighted linear regression adjusting for the number of users per state, Figure 2.4A) and the proportion of votes for Republican candidates for the House of representatives in 2018 ($R^2 = 0.78$ in a weighted linear regression, Figure 2.4B). This analysis suggests that, despite the fact that the Twitter users in our dataset are a highly self-selected sample of the population, the distribution of right-leaning users in our dataset is very similar to that of right-leaning voters across states.

### 2.4.2 Comparison with Network-Based User Alignments

Next, we compare the alignments we calculated based on the users' sharing patterns with other approaches of calculating political alignment.

Barberá et al. [8] propose a method for inferring the users' alignment by considering who they follow. Their key idea is that users are more likely to follow accounts that align with their political views, and in particular, to follow accounts that have an unambiguous ideological leaning such as presidential candidates, legislators, and media outlets. Their method works as follows: (a) they construct an adjacency matrix which indicates whether a user $i$ (rows) follows a political account $j$ (columns), (b) they apply Correspondence Analysis which uses SVD to factorize the matrix by its most important dimensions, (c) they consider the first component to be an estimate of the political alignment of the user, and (d) they standardize the estimates to have a mean of zero and standard deviation of one to ease interpretation. They demonstrate that since ideology is the most salient

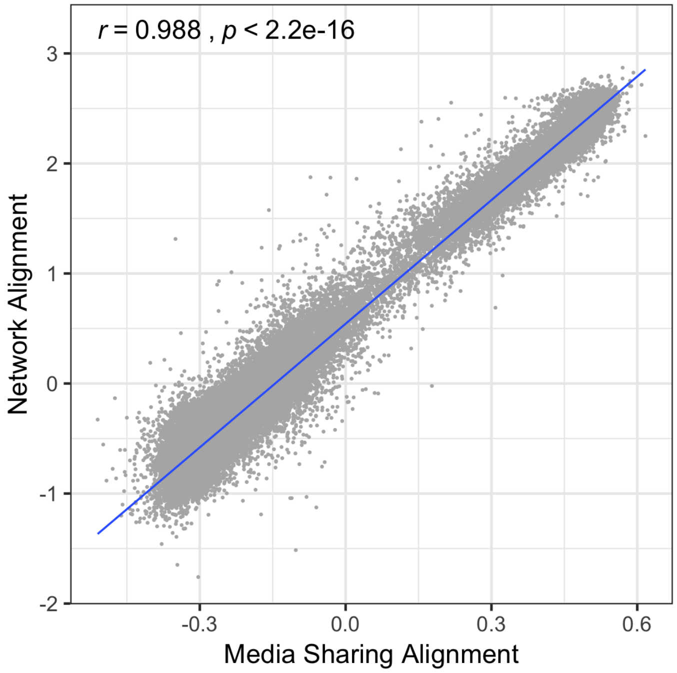


**Figure 2.5:** Comparison of the tweet alignment scores calculated based on the users' media-sharing patterns against those calculated based on their follow relationships.

dimension discriminating among the political accounts in the columns, the first component indeed aligns well with ideology.

We compare the tweet alignment scores calculated based on the users' media-sharing patterns against those calculated based on their follow relationships. The follow graph based alignments were computed based on the state of the follow graph in July, 2018[1]. We use only tweets posted by the outlets for which we had enough information to compute both types of alignments, i.e., tweets with at least three retweeters whose alignments we could estimate. We find a very strong correlation between the alignments computed using the two methods: Pearson $r = 0.99$ and Spearman $\rho = 0.80$ (Figure 2.5).

While the alignment estimates are highly correlated, we observe that in 93% of the tweets, the alignments based on the media sharing patterns give us more information to estimate the alignments of the tweets in our dataset.

[1] We thank Pablo Barberá for kindly providing the user alignment scores to us.

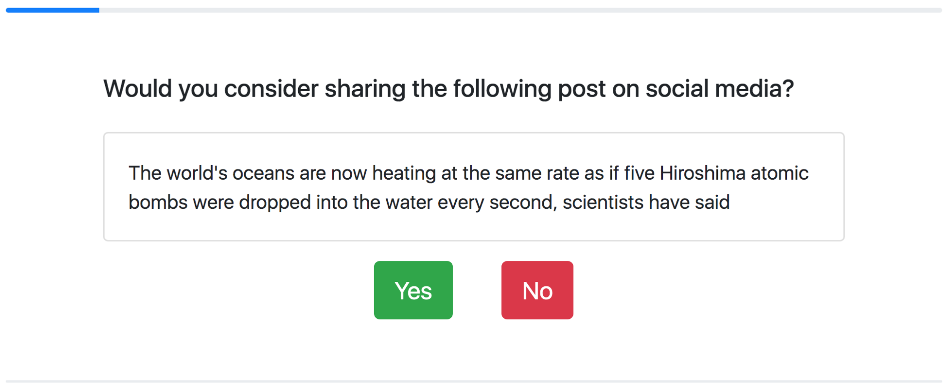
We find that we can infer the user alignments of more retweeters per tweet and that we have enough information to compute the alignments of more tweets (i.e., more tweets have at least three retweeters whose alignment we could infer). This is perhaps because users who have retweeted one of the outlets' tweets are more likely to have also shared links to other outlets that are in the list of domains we use to infer the alignments.

### 2.4.3 Mechanical Turk Survey

To further validate our tweet alignment scores, we run a Mechanical Turk survey asking left- and right-leaning participants whether they would consider sharing sample tweets and compare their responses against the inferred alignment scores. We build on recent work which has shown that self-reported willingness to share political news in online surveys conducted on Mechanical Turk correlates with actual sharing on Twitter [69].

We sampled ten tweets for each outlet and Frontline, excluding Fox News whose account was inactive at the time due to their boycott of Twitter. To ensure that the tweets included in the survey were about topics that were relevant at the time, we considered only tweets posted by the outlets during the six weeks prior to the survey (i.e., February, 2020). To make sure that the sample tweets capture the variation of tweet alignments, we took a stratified sample for each outlet: we first computed the deciles of the outlet's tweet alignment distribution and then sampled one tweet from each decile. We also ensured that the tweets are self-contained in that they (*i*) do not require any additional context to understand, and (*ii*) do not reveal the organization that posted them.

The survey included 25 questions. We first explained to the participants that they will be shown a series of social media posts posted by major news outlets and that they will be asked whether they will consider sharing them. Then, we showed them one tweet at a time and asked them "Would you consider sharing the following post on social media?" (Figure 2.6). We worded the question in a similar way as Mosleh et al. [69]. We did not



Would you consider sharing the following post on social media?

The world's oceans are now heating at the same rate as if five Hiroshima atomic bombs were dropped into the water every second, scientists have said

**Figure 2.6:** Sample question of the Mechanical Turk survey used to validate the tweet alignment scores.

show the images or the headlines associated with the tweets. To avoid any bias due to ordering effects, we randomized the order of the questions and the order of the response buttons. We also included three attention checks, randomly placed among the other questions, that simply asked the participants to press the "yes" or "no" button.

To ensure high-quality responses, we invited only participants from the US that have completed at least 100 tasks and have a high approval rate (>98%). One of the main advantages of Mechanical Turk is that it allows us to recruit participants that have self-identified as "Liberal" or "Conservative". We administered the survey such that, for each sample tweet, we obtained 50 responses by "Liberal" and 50 responses by "Conservative" participants. We compensated them 60 cents for completing the survey or roughly 9$ per hour. Each participant could take the survey only once. The protocol was approved by the MIT Institutional Review Board.

For each tweet, we compute the fraction of right-leaning survey participants out of all survey participants who said that they would consider sharing the tweet. We compare these values against the average political alignment scores and the fraction of users classified as right-leaning that actually retweeted the tweet. As we described above, we used the users' media sharing patterns to infer their alignment and to classify them into left- vs. right-leaning.

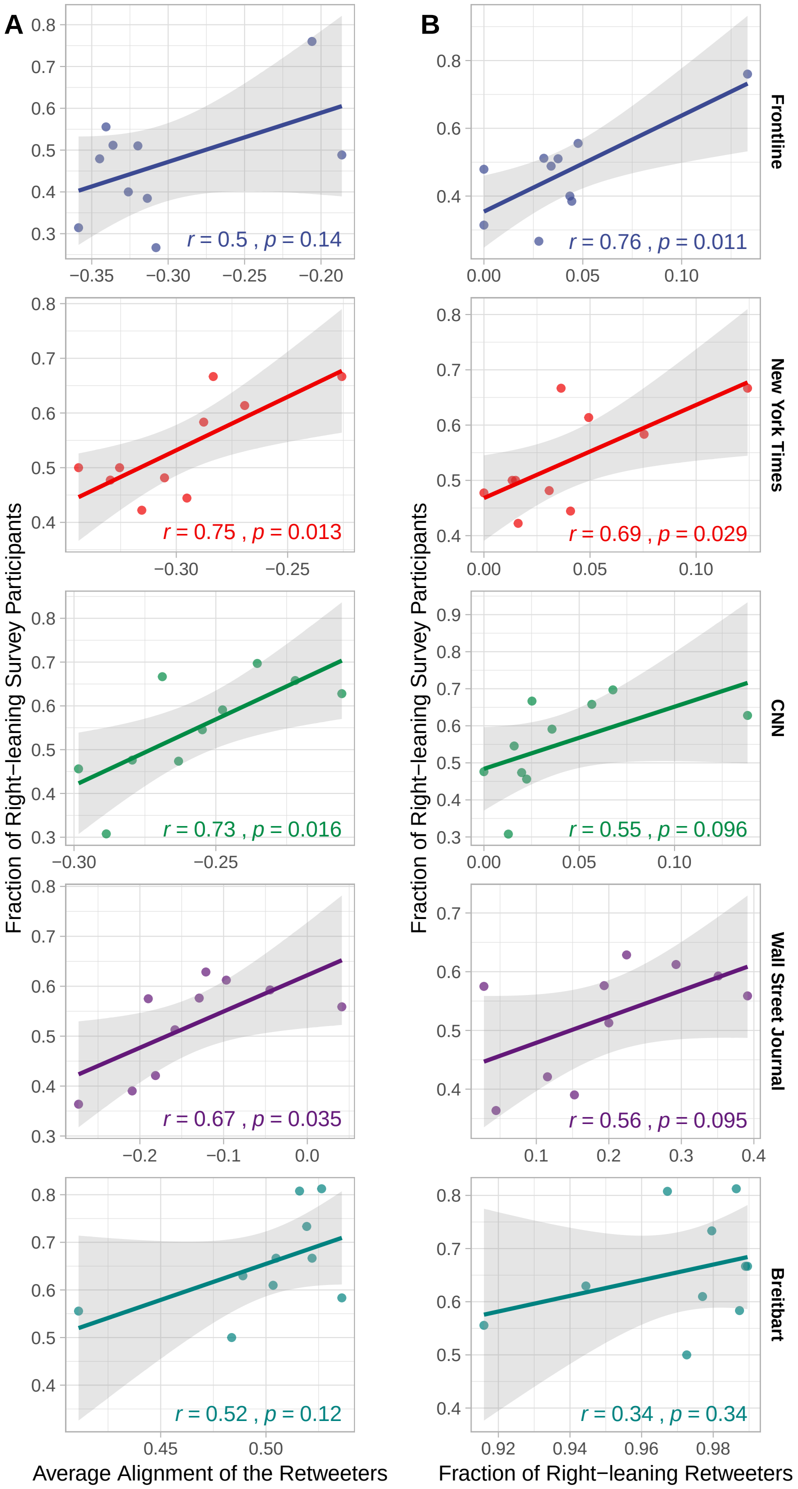


**Figure 2.7:** Comparison of the fraction of right-leaning survey respondents who said that they would consider sharing a sample tweet against the average alignment of the users who actually retweet the tweet (A) and the fraction of right-leaning users who retweeted the tweet (B).

We find a positive correlation between the fraction of right-leaning survey participants who said that they would share the tweets and the average alignment of the retweets, ranging between $r = 0.5$ and $r = 0.75$ when we consider the tweets of each outlet individually, and $r = 0.53$ when we consider all tweets together (Figure 2.7A). We also find a positive correlation between the survey responses and the fraction of retweeters that were classified as right-leaning, ranging between $r = 0.34$ and $r = 0.76$ by outlet and $r = 0.50$ in overall (Figure 2.7).

In summary, we find a strong alignment between the tweet metrics based on the retweeters' inferred political leaning and the survey responses, despite the fact that the survey respondents were shown only the tweet text.

## 2.5 PREDICTIVE MODELING

### 2.5.1 Measuring Audience Diversity

**DESIDERATA.** Before we discuss the pros and cons of any specific choices, it is worth outlining our goals and constraints in measuring audience diversity. First, the main goal is to adopt a measure that will allow us to quantify the extent to which both left- and right-leaning users engaged with a tweet. Second, the measure needs to be intuitive and easy to explain to non-experts, such as the journalists who will use the predictive models to compose new tweets. Third, we need a measure that we can use both in our predictive modeling and when running advertising experiments that test whether choosing tweets with predictive models actually leads to higher audience diversity.

**CLASS DEFINITIONS.** As we will discuss in more detail later, we are much more constrained in what we can measure during the advertising experiments. For instance, we can only specify which users the advertisements could be shown to and measure the overall engagement of the group. Although we can have more granular measurements of the users' alignments, we will not be able to know the identities of the individual users who engaged with the content. This inherently constrains us to a definition of audience diversity based on a categorical definition of the users' alignment. As such, we can run separate advertising campaigns for each category of users, measure their engagement, and calculate the diversity.

The most natural categorization of the users is to classify them as left- and right-leaning. This classification also leads to an intuitive way of measuring the diversity of a group of users: the group is most diverse if there is an equal number of left- and right-leaning users and least diverse if the group consists of only left- or only right-leaning users.

Alternatively, we can classify the users into more granular classes, e.g., far left, left, center, right, far right leaning. The benefit of this classification is that although we still group users into discrete categories, we preserve more information from the continuous alignment score[2]. However, such classification complicates the definition and interpretation of a diversity measure. First, it is unclear how to define diversity under this classification. Should a group of only centrist users have the same diversity as a group with an equal number of left- and right-leaning users? Should we weigh the far-left and far-right leaning users more than the left- and right-leaning users and by how much? Is a group of an equal number of far-left, left, and centrist users more diverse than a group of an equal number of left, centrist, and right-leaning users? Second, even if we answer these questions, explaining and interpreting such a definition will be much harder and require explaining the subjective choices in defining the measure.

With these trade-offs in mind, we opt for the binary classification of the users into left- and right-leaning categories.

**DIVERSITY MEASURE.** Given the binary classification of the users to left- and right-leaning, one way to measure the diversity of a group of users is to define a discrete random variable $X$ that takes two values, $left$

[2] By the same token, this increases the measurement error: users are now even more likely to be assigned to an incorrect class due to an error in our user alignment measurements.

and *right*, with respective probabilities $p_{left}$ and $p_{right}$, and compute its entropy. The entropy is commonly used to measure diversity and is often referred to as the Diversity Index or Shannon Index. It is defined as:

$$H(X) = -p_{left} \log_2(p_{left}) - p_{right} \log_2(p_{right}).$$

It is maximized when ($H = 1$) when $p_{left} = p_{right}$ and minimized ($H = 0$) when $p_{left} = 0$ or $p_{right} = 0$ (Figure 2.8a).

We also considered using cross-entropy and Kullback–Leibler divergence. The benefit of these measures is that they would allow us to account for the distribution of left- and right-leaning users per outlet. We can define another random variable $Y$ that captures the overall audience of an outlet and takes two values, $left$ and $right$, with respective probabilities $q_{left}$ and $q_{right}$. We can estimate $q_{left}$ and $q_{right}$ by calculating the number of left- and right-leaning followers or the number of left- and right-leaning retweeters across many tweets. The cross-entropy is defined as:

$$H(X,Y) = -q_{left} \log_2(p_{left}) - q_{right} \log_2(p_{right}),$$

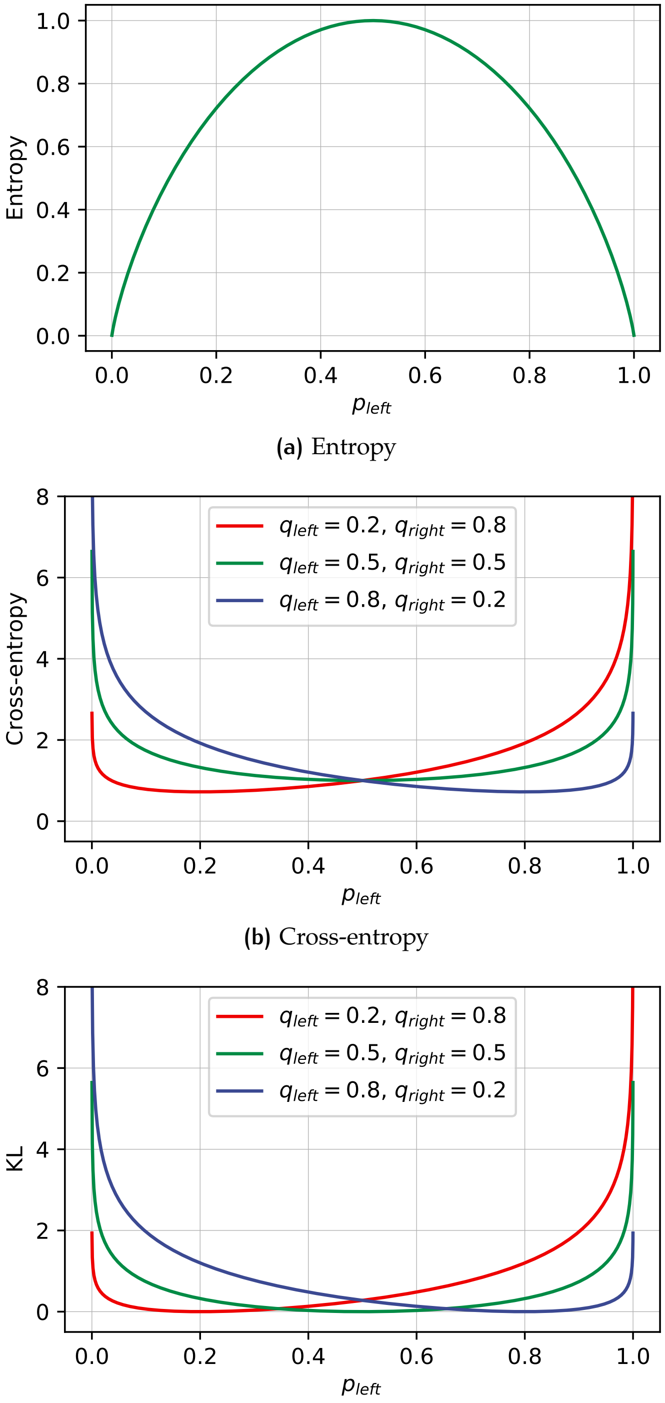
and the KL divergence is defined as:

$$D_{KL}(X||Y) = -q_{left} \log_2 \left(\frac{q_{left}}{p_{left}}\right) - q_{right} \log_2 \left(\frac{q_{right}}{p_{right}}\right).$$

The entropy, cross-entropy, and the KL divergence are related to each other:

$$H(X,Y) = H(X) + D_{KL}(X||Y).$$

As Figures 2.8b and 2.8c show, the cross-entropy and the KL divergence measure how much the tweet distribution ($p$) departs from the outlet distribution ($q$). Their values increase when the tweet alignment distribution ($p_{left}$ vs. $p_{right}$) is both more and less evenly split than $q$ ($q_{left}$ vs. $q_{right}$) and do not capture whether the tweet is more or less bridging than we would expect given the outlet distribution. This makes them an unsuitable choice for a diversity measure in this case.



**(a)** Entropy

**(b)** Cross-entropy

**(c)** KL Divergence

**Figure 2.8:** Entropy, cross-entropy, and KL divergence as a function of $p_{left}$, i.e., the breakdown of left- vs. right-leaning retweeters. Note that $p_{right} = 1 - p_{left}$. In the case of the cross-entropy and the KL divergence, $q$ ($q_{left}$ and $q_{right}$) represents the overall audience distribution of the outlet.

**ESTIMATION.** To estimate $p_{left}$ and $p_{right}$ for each tweet, we use Maximum Likelihood Estimation with Laplace Smoothing [61], i.e., we add a pseudocount of one to the number of observed retweets by left- and right-leaning users:

$$p_{left} = \frac{n_{left} + 1}{n_{left} + n_{right} + 2}, \qquad p_{right} = \frac{n_{right} + 1}{n_{left} + n_{right} + 2},$$

where $n_{left}$ and $n_{right}$ is the number of observed retweets by left- and right-leaning users, respectively.

This estimation approach also has a Bayesian interpretation: it is equivalent to using a Beta distribution—Beta(1,1), in particular—as the conjugate prior for the parameters of a Binomial distribution [31]. The smoothing has the most significant effect on the estimates of tweets with a small number of retweets, and its effect diminishes as the number of observed retweets increases. Note that we still consider only tweets with at least three retweets. From a practical perspective, the smoothing allows us to distinguish between polarizing tweets with a few retweets (e.g., $n_{left} = 3$, $n_{right} = 0$) and polarizing tweets which have a lot of retweets (e.g., $n_{left} = 100$, $n_{right} = 0$); in the latter case, we have much more information that suggests that the tweet is indeed very polarizing.

### 2.5.2 Defining the Prediction Target

Next, we investigate different ways of defining the target variable of our prediction problem. We consider three possibilities: (*i*) regression task where we predict the entropy of the retweeters' political alignment distribution, (*ii*) classification task where we predict whether the tweet's entropy is below or above the median entropy of all tweets in the dataset, and (*iii*) classification task but we predict whether the tweet's entropy is below or above the median entropy of all tweets posted by the outlet.

The first approach is most intuitive; since our diversity measure is continuous, it is natural to think of the prediction problem as a regression. The benefit of the second approach is that by casting the problem to classification, we train the model to distinguish between bridging and non-bridging language, rather than to quantify exactly how bridging the content is, which is a much harder learning problem. The classification setup is also less sensitive to measurement error, i.e., the measurement error needs to be large for a data point to be placed in the wrong class. Finally, the third approach allows us to take into account the fact that different outlets have different compositions of Twitter followers, which can affect the diversity of the users who engage with their tweets. By defining the target with respect to the outlet's distribution of tweet audience diversity, we train the models to predict whether the tweet will have a more or less diverse audience than expected for the outlet.

Since the three approaches lead to different prediction tasks, we cannot simply use the typical model evaluation metrics to compare the resulting models. For instance, it is unclear how to compare the mean squared error of the regression models against the classification accuracy of the classification models. To resolve this issue, we consider how we will eventually use the models to make decisions. Mainly, we are interested in using the model to choose between a pair of tweets or rank a set of candidate tweets by their expected audience diversity[3]. Intuitively, given a pair of two random tweets, we would like the model to be accurate at predicting which one of the two tweets will be more bridging. It turns out that this is exactly what the Kendall's $\tau$ rank correlation coefficient measures [51, 52]. Given $n$ paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ the Kendall $\tau$ coefficient is defined as:

$$\tau = \frac{C - D}{\sqrt{(C + D + X_0)(C + D + Y_0)}},$$

where $C$ is the number of *concordant* pairs (i.e., $x_i > y_i \wedge x_j > y_j$, or $x_i < y_i \wedge x_j < y_j$), $D$ is the number of *discordant* pairs (i.e., $x_i > y_i \wedge x_j < y_j$, or $x_i < y_i \wedge x_j > y_j$), and $X_0$ and $Y_0$ are the number of pairs tied only

[3] These two use-cases are very similar as any ranking problem can be solved by making many pairwise choices.

Table 2.1: Comparison between the three approaches of defining the prediction target in terms of the agreement between the rankings based on the model predictions and the ground-truth measured using Kendall's $\tau$ rank correlation coefficient. We measure the agreement of the predictions per outlet and overall, i.e., considering the predictions for all outlets together.

| | (*i*)<br>Regression | (*ii*)<br>Classification<br>$>$ global median? | (*iii*)<br>Classification<br>$>$ outlet median? |
|---|---|---|---|
| Frontline | **0.22** | 0.18 | 0.21 |
| New York Times | 0.22 | 0.20 | **0.24** |
| CNN | 0.29 | 0.28 | **0.30** |
| Wall Street Journal | 0.16 | 0.11 | **0.21** |
| Fox News | 0.45 | 0.31 | **0.46** |
| Breitbart | 0.07 | 0.06 | **0.10** |
| Overall | **0.48** | 0.44 | 0.22 |

on the $X$ and $Y$ variables, respectively. The coefficient ranges between $+1$ (perfect agreement) and $-1$ (perfect disagreement).

To compare the three approaches, we split the dataset into an 80% training set and a 10% validation set using stratified random sampling, preserving the distribution of the number of tweets per outlet. We use the training set to fit three models, one for each approach, then we use the fitted models to score the validation set, and finally, we compute the ranking correlation between the model predictions and the observed (i.e., ground-truth) audience diversity. We train BERT models [25] but specify different loss functions and target variable definitions: we use root mean squared error loss for the regression (approach (*i*)) and binary cross-entropy loss for the classification tasks (approaches (*ii*) and (*iii*)). We will provide more details about the model in the next section.

Table 2.1 shows the ranking correlation between the model predictions and the ground-truth per outlet and overall (i.e., considering the predictions for all outlets together). We find that the regression model (*i*) performs best when we consider the data from all outlets together. Perhaps not surprisingly, when we consider each outlet individually, the classification model with “above outlet median” as a target (*iii*) performs best for all outlets, except for Frontline. For most outlets, the regression model performs only slightly worse, and always better than the classification model with “above global median” as a target (*ii*). Overall, the regression model strikes a good balance between achieving good performance for all outlets together and for each outlet individually. Moreover, it performs best for Frontline, the account for which we will use to model to guide our decisions in the advertising experiments.

Based on this analysis, in the rest of this chapter, we formulate the prediction problem as a regression task.

### 2.5.3 Learning Methods

Now that we have decided how to measure audience diversity and formulate the prediction problem, we describe the different learning methods we use for prediction. We consider a wide variety of models from simple linear models on TF-IDF representations of the input text to the state-of-the-art neural network approaches for natural language processing.

Before we apply the models, we preprocess the input text (i.e., the outlets’ tweets) by converting them to lower-case, removing punctuation, and replacing numbers with “#”. We also remove any URLs and Twitter @mentions; however, we keep hashtags as they may carry important semantic information.

**TF-IDF + LINEAR MODELS.** We start with simple linear models. We consider several ways of representing the text in vector space. (*i*) We tokenize the (preprocessed) text and build a vocabulary of all uni-grams or all uni-grams and bi-grams. (*ii*) We build a vocabulary of all character n-grams of size three to five, either by including white-spaces or respecting the token boundaries. (*iii*) We also test a more sophisticated tokenization technique called SentencePiece [56], which breaks up tokens into subtoken units, selected by analyzing the full corpus. By representing more complex tokens using simpler sub-tokens, this approach allows us to handle any input, even tokens not seen in the training set, and to control the vocabulary size. Regardless of how we build the vocabulary—unigrams/bi-grams, character n-grams, or sentence pieces—we always limit the vocabulary size to 32,000.

Next, given the vocabulary, we encode the tweets using TF-IDF feature representations. We also test whether standardizing the features helps. We scale the features to unit variance but do not center them in order to avoid breaking the sparsity structure of the data.

We test five model types: (*i*) Linear Regression, (*ii*) Ridge Regression, (*iii*) Lasso, (*iv*) Elastic Net, (*v*) Support Vector Regression. For each model, except Linear Regression, we tune the strength of the L1 / L2 regularization parameter, $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$, and when training Elastic Nets we give equal weight to the L1 and the L2 regularization.

**WORD EMBEDDINGS.** We also consider models based on pre-trained word embeddings. We use the word2vec embeddings, trained on 6 billion tokens of Google News articles using language modeling as a training objective [67]. We chose word2vec instead of other pre-trained embeddings (e.g., Glove [77] or FastText [11]) as they were trained on a very similar domain.

To obtain tweet embeddings, we consider three different ways of aggregating the word embeddings. (*i*) We compute the average embedding of all the vectors associated with the words in the tweets. (*ii*) We use *smooth inverse frequency* weighting, a theoretically motivated approach to aggregating word embedding inspired by the traditional TF-IDF weighting [1]. It works by computing the average embeddings and then removing the projections of the average embeddings on their first singular vector. It is often considered as a "tough to beat" baseline for sentence representations. (*iii*) We use a self-attention mechanism: we compute an average of the word embeddings weighted by the words' attention scores [60]. To learn the attention scores, we use a two-layer neural network followed by a softmax: the network takes the individual word embeddings as inputs and outputs a score for each embedding. The weights of the attention network are free parameters that we learn as part of the training. Self-attention has been successfully used in a variety of NLP tasks including reading comprehension [15], textual entailment [74], and abstractive summarization [76].

After we aggregate the word embeddings, we feed the tweet representation into a series of fully-connected layers with ReLU activations [70], followed by a prediction layer. We tune several aspects of the learning procedure: whether we freeze or fine-tune the word2vec embeddings, the size of the layers in the attention network (64, 128, 256), and the number (0, 1, 2) and the size (128, 256, 512) of the fully-connected layers.

**RECURRENT NEURAL NETWORKS.** Recurrent neural networks (RNNs) are particularly suitable for natural language processing as they can be used to encode sequences of arbitrary length and to capture dependencies between the tokens in a sequence. They are designed to process sequences one token at a time, taking into account the contextual information encoded in the preceding tokens. We consider two types of RNN architectures: Long Short-Term Memory units (LSTMs) and Gated Recurrent Units (GRUs). LSTMs [43] were designed to solve the problem of vanishing gradients which made effectively training simple RNN architectures on longer sequences practically impossible. GRUs [18] are a simplified version of the LSTMs that is much less computationally expensive and still achieves comparable performance [19]. In both cases, we train bi-directional RNNs (i.e., we process the sequence left-to-right and right-to-left) and use the pre-trained word2vec token embeddings as input.

RNNs output a representation/embedding of each token as they process the sequence. We consider three ways of aggregating the embeddings to obtain a tweet embedding. (*i*) We concatenate the last outputs of the RNN in both directions, left-to-right and right-to-left. Since the RNNs capture contextual information as they process each token, we expect the final outputs to capture longer dependencies. (*ii*) We compute the mean embedding of the RNN outputs for each token. (*iii*) We use self-attention [60] and compute the mean embeddings of the RNN outputs weighted by the learned attention scores. Once we aggregate the token embeddings, we feed the tweet embedding into a series of fully-connected layers with ReLU activations [70].

We tune the network architecture by testing how the performance changes as we vary the size (128, 256, 512) and the number of RNN layers (1, 2, 3, 4), the pooling mechanism (last embedding, mean embedding, attention with different parameters of the attention network) and the number (0, 1, 2), size (128, 256, 512) and dropout rate (0.0, 0.3) of the fully-connected layers.

**BERT.** Bidirectional Encoder Representations from Transformers or BERT [25] is a language representation model that processes tokens in relation to all other tokens in the sentence, unlike RNN-based models that process tokens in order, one token at a time. At its core, BERT consists of a series of Transformer encoder layers [93] which consist of multiple "heads", fully-connected layers with self-attention. For every input token in a sequence, each head computes key, value, and query vectors, which are used to create a weighted representation. The outputs of all heads in the same layer are combined and feed into a fully-connected layer [80]. The key advantages of Transformers are that they are better at modeling long-range dependencies and that they can be trained in parallel, making it feasible to train very large models with hundreds of millions of parameters. BERT has been used to achieve state-of-the-art results in numerous NLP benchmarks and has been integrated into Google Search, leading to significant improvements in understanding and ranking search queries[4].

To adopt BERT for our task we average the token embeddings of the last Transformer layer and add a fully-connected layer with a dropout of 0.1 and ReLU activations, followed by a prediction layer. We initialize

[4] https://blog.google/products/search/search-language-understanding-bert/

the network with the pre-trained BERT model and fine-tune it using our dataset.

### 2.5.4 Experimental Protocol

To train the models, we split the dataset into 80% training, 10% validation, and 10% test sets using stratified random sampling to preserve the same distribution of tweets per outlet. Due to the long training times of some models, we were unable to run cross-validation. To tune the model architectures and hyper-parameters, we train the models on the training set, evaluate the model variations of the validation set to choose the best model, and measure its performance on the test set. Due to the large number of hyper-parameter combinations and long training times, it was too computationally expensive to perform a grid search. Instead, we first optimized one parameter at a time to find the most promising values and then tested the combinations of those values.

To train the neural models we evaluate the performance on the validation set after every epoch and stop training and select the last best model if the performance has not improved in the last 5 epochs for BERT and 10 epochs for all other models or if we have reached the maximum number of epochs, 15 for BERT and 100 for all other models. We train the networks using mini-batches of size 32 for BERT and size 64 for all other models. To prevent the gradients from exploding, we clip them to a unit L2 norm after every mini-batch [36]. We use the Adam optimizer [54] setting $\beta_1 = 0.9$, $\beta_2 = 0.999$, and L2 weight decay of 0.01. We consider the following learning rates for BERT, $lr \in \{2 \times 10^{-5}, 3 \times 10^{-5}, 4 \times 10^{-5}, 5 \times 10^{-5}\}$ as recommended in [25], and a larger set of values for all other models, $lr \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$. We use Mean Squared Error as a loss function for all neural models. In addition to the Mean Squared Error, we also report the Mean Absolute Error of the model predictions to ease the interpretation of the results.

### 2.5.5 Results

In Figure 2.9, we show the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) of the models on the test set using the hyperparameters that performed best on the validation set. To put the results in perspective, we use two constant predictors as baselines: the mean and the median of the tweets' audience diversity in the training set. The mean minimizes the MSE, and the median minimizes the MAE of the target variable in the training set.

We find that all linear models trained on the TF-IDF representations of the tweet text perform significantly better than the baselines. All models have a similar performance with MAE ranging between 0.18 and 0.184. We observe that tokenizing the text using SentencePiece tokenization and including all uni-grams and bi-grams in the vocabulary leads to the best performance for all linear models except for SVR, which works slightly better with regular tokenization.

We find that the mean word embedding of the tweet tokens is not a good predictor of the tweet audience diversity. In fact, the models based on averaging the word2vec embeddings perform worse than the linear models trained on TF-IDF representations. Moreover, using the *smooth inverse frequency* weighting does not improve the performance. However, using self-attention, i.e., learning different weights for each word embedding, leads to significantly better results, improving over the linear models trained on TF-IDF representations.

We observe that the Recurrent Neural Network models (RNNs) work slightly better than the word2vec embeddings with self-attention. The two variants, GRUs and LSTMs, achieve very similar results. Both models perform best when aggregating the RNNs outputs using self-attention and using only one RNNs layer instead of stacking multiple layers.

Finally, we find that the fine-tuned BERT model performs best, significantly outperforming the RNN models. It achieves a MAE of 0.14 and MSE of 0.036. We observed that the model performs well with different

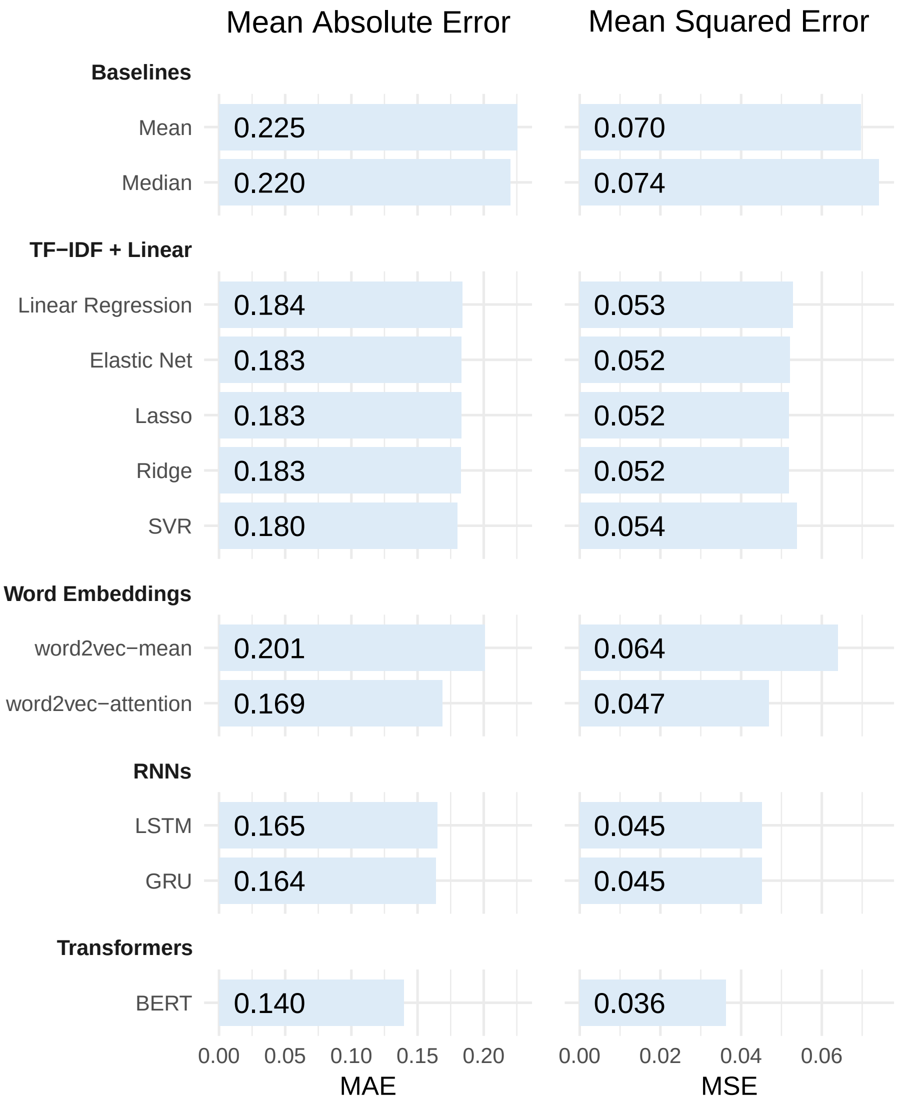


**Figure 2.9:** Performance of the different regression models predicting the tweets' audience diversity given the tweet text.

learning rates and different sizes of the final fully-connected layer that we added to the network architecture. We use this model for all analyses presented in the rest of this chapter.

### 2.5.6 BERT Model Variations

Next, we describe three variants of the BERT model that we experimented with but, unfortunately, did not lead to significant improvements in performance.

**$p_{left}$ AND $p_{right}$ AS TARGETS.** Instead of using the audience diversity (i.e., the entropy of the political leaning of the users who engage with the tweet) as a target variable, we used the distribution of left- and right-leaning users (i.e., $p_{left}$ and $p_{right}$) as target variables. We used softmax to transform the two model outputs to probabilities ($p_{\hat{left}}$ and $p_{\hat{left}}$) and the KL-divergence as a loss function.

In the original formulation, due to the nature of the entropy function, the model has no way of distinguishing whether the lack of diversity is because the tweet is less interesting to left- or right-leaning users. Thus, the intuition behind this formulation is that having $p_{left}$ and $p_{right}$ as outputs will provide more specific supervision to the model and improve its performance.

While we trained the model using $p_{left}$ and $p_{right}$ as outputs, we evaluated it by computing the entropy of $p_{\hat{left}}$ and $p_{\hat{left}}$ and measuring how close it is to the observed entropy. This allowed us to compare the new formulation with the original formulation. To our surprise, the model performed slightly worse, achieving a MAE of 0.153 and MSE of 0.043.

**PRE-TRAINING WITH RETWEETERS' OUTLET-SHARING DISTRIBUTIONS.** Next, we tried even a more extreme version of the formulation above. Instead of using $p_{left}$ and $p_{right}$, we used the mean distribution of outlets that the users share as a target. For each user, we compute the distribution of how often they share each of the 500 outlets, which we used to compute their alignment, resulting in a 500-dimensional vector per user. Then, we take the outlet-sharing distributions of all the users who shared the tweet and compute the mean vector for each tweet.

We train the model in two phases. In the first phase, we pre-train the model using the users' mean outlet-sharing distributions as a target. As before, we use softmax in the last layer to obtain probabilities as outputs and the KL-divergence as a loss function. In the second phase, we take the pre-trained model, replace the last layer with a new fully-connected layer, and fine-tune the model using the audience diversity as a target (single output, with MSE as a loss function). The intuition was that pre-training the model with more granular supervision will lead the model to more promising regions of the model space. The performance of this model was only slightly better, MAE: 0.137, MSE: 0.037.

**EXPANDING THE TRAINING SET.** Lastly, we tried expanding the training set by including tweets from the 3-year snapshot of the Decahose (Section 2.3.2). We included only tweets by the Twitter accounts of the 500 outlets we used to compute the user alignments. We added 2.4M new tweets to the training set, increasing its size by more than six times. Note, however, that since we have only 10% of the tweets' retweets, we have noisier estimates of the audience diversity of each tweet. Moreover, tweets are more topically diverse than those in the original training set. To make a fair comparison, we evaluated the model on the same train/validation/test splits as the original model but added the new data to the training set. It achieves a MAE of 0.16 and MSE of 0.044. One reason for the lower performance might be a slight difference in the distribution of audience diversity in the two datasets.

Since none of these variations led to significant improvements, we use the original model in the rest of our analyses.

## 2.6 WEB APPLICATION

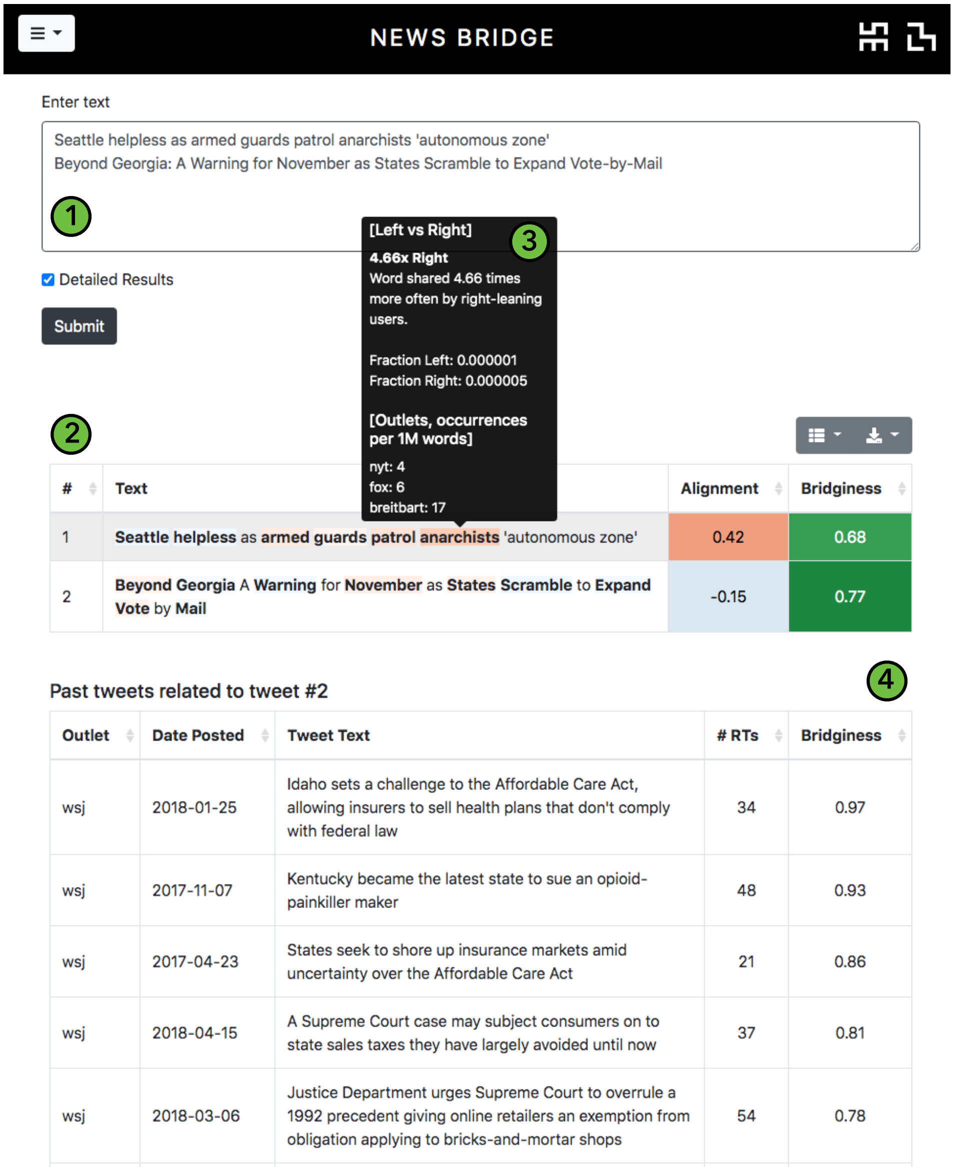
To make the models more easily accessible to Frontline's journalists, we build a web application that surfaces the model predictions. The goal of the application is to allow the journalists to quickly iterate on tweet drafts based on the model predictions and to help them select candidate tweets from the film transcripts.

### 2.6.1 Main Interface

Figure 2.10 shows a screenshot of the main page of the application. The users can enter draft tweets in the input box (Figure 2.10#1), press submit, and get the model predictions in the results table. The input can consist of multiple tweets, separated by a new line, and each tweet can consist of multiple sentences that will be scored together.

The results table (Figure 2.10#2) shows the tweet text and the predicted "bridginess" score, which is a user-friendly name for the audience diversity measure using the entropy of the distribution of left- and right-leaning users, as we detail in Section 2.5. The color of the table cells containing the scores varies from light-green for non-bridging tweets to dark-green for bridging tweets.

Based on feedback from the journalists, we also added tweet alignment score predictions. The goal of these scores is to supplement the bridginess scores. To make the alignment predictions, we trained a BERT model equivalent to the one we used to make the bridginess predictions, but instead of the entropy we used the retweeters' average political alignment as a target (we provide more details about the model in Section 2.5.3). To include the alignment predictions in the results table, the user needs to check the "Detailed Results" checkbox before submitting the text. Similar to the bridginess scores, we vary the color of the table cells that contain the alignment scores on a gradient from blue (if the score is negative, i.e.,



NEWS BRIDGE

Enter text

Seattle helpless as armed guards patrol anarchists 'autonomous zone'
Beyond Georgia: A Warning for November as States Scramble to Expand Vote-by-Mail

☑ Detailed Results

Submit

[Left vs Right]

**4.66x Right**
Word shared 4.66 times more often by right-leaning users.

Fraction Left: 0.000001
Fraction Right: 0.000005

**[Outlets, occurrences per 1M words]**

nyt: 4
fox: 6
breitbart: 17

| # | Text | Alignment | Bridginess |
|---|---|---|---|
| 1 | **Seattle helpless** as **armed guards patrol anarchists** 'autonomous zone' | 0.42 | 0.68 |
| 2 | **Beyond Georgia** A **Warning** for **November** as **States Scramble** to **Expand Vote** by **Mail** | -0.15 | 0.77 |

Past tweets related to tweet #2

| Outlet | Date Posted | Tweet Text | # RTs | Bridginess |
|---|---|---|---|---|
| wsj | 2018-01-25 | Idaho sets a challenge to the Affordable Care Act, allowing insurers to sell health plans that don't comply with federal law | 34 | 0.97 |
| wsj | 2017-11-07 | Kentucky became the latest state to sue an opioid-painkiller maker | 48 | 0.93 |
| wsj | 2017-04-23 | States seek to shore up insurance markets amid uncertainty over the Affordable Care Act | 21 | 0.86 |
| wsj | 2018-04-15 | A Supreme Court case may subject consumers on to state sales taxes they have largely avoided until now | 37 | 0.81 |
| wsj | 2018-03-06 | Justice Department urges Supreme Court to overrule a 1992 precedent giving online retailers an exemption from obligation applying to bricks-and-mortar shops | 54 | 0.78 |

Figure 2.10: News Bridge web application, a screenshot of the main interface.

the tweet is predicted to be more engaging to left-leaning users) to red (if the score is positive).

### 2.6.2 Explanations

After the initial deployment of the tool, the main feedback we received was that while the scores are informative, it is often unclear why the model made the predictions it did. To address this, we show two kinds of explanations to supplement the predictions: (*i*) we highlight certain words in the tweets and display relevant corpus statistics, and (*ii*) we show historical tweets that are semantically similar to the input tweet and have a high bridginess score.

**WORD HIGHLIGHTING.** We compute how often each word is retweeted by left- and right-leaning users[5]. More specifically, we compute the probability $p(word|left\text{-}leaning)$ (and $p(word|right\text{-}leaning)$) as the ratio between the number of times the word appears in retweets by left- (right-) leaning users and the total number of words in all retweets by left- (right-) leaning users. We highlight the words of the input text in blue or red depending on whether they are more likely to be retweeted by left- or right-leaning users, and we set the brightness of the color in proportion to the ratio between the two quantities ($p(word|left\text{-}leaning)$ and $p(word|right\text{-}leaning)$). We also compute how often each word appears in tweets posted by each of the five news outlets and Frontline. When the user hovers over the word, a pop-up shows the different word statistics (Figure 2.10#3).

Beyond highlighting the words based on simple word statistics, we also considered two other, more sophisticated approaches. (*i*) A common way of visualizing which words in the text were most important in the model prediction is to use the self-attention weights. However, recent studies have

[5] We consider only retweets of tweets in our dataset, i.e., posted by the five news outlets and Frontline.

shown that attention weights do not provide meaningful explanations for the model predictions [47]. It is worth noting that this is still a topic of active debate [94]. (*ii*) We also considered using Integrated Gradients [88], a technique that aims to explain the relationship between a model prediction and the input features. However, we found that it takes about 10 to 20 seconds to compute the integrated gradients on a single prediction of our model. This is likely due to the fact that the BERT model has a very large number of parameters. Since the goal of this tool is to allow journalists to quickly iterate on the tweet text based on the model predictions, we decided that increasing the latency of the predictions would significantly degrade the user experience.

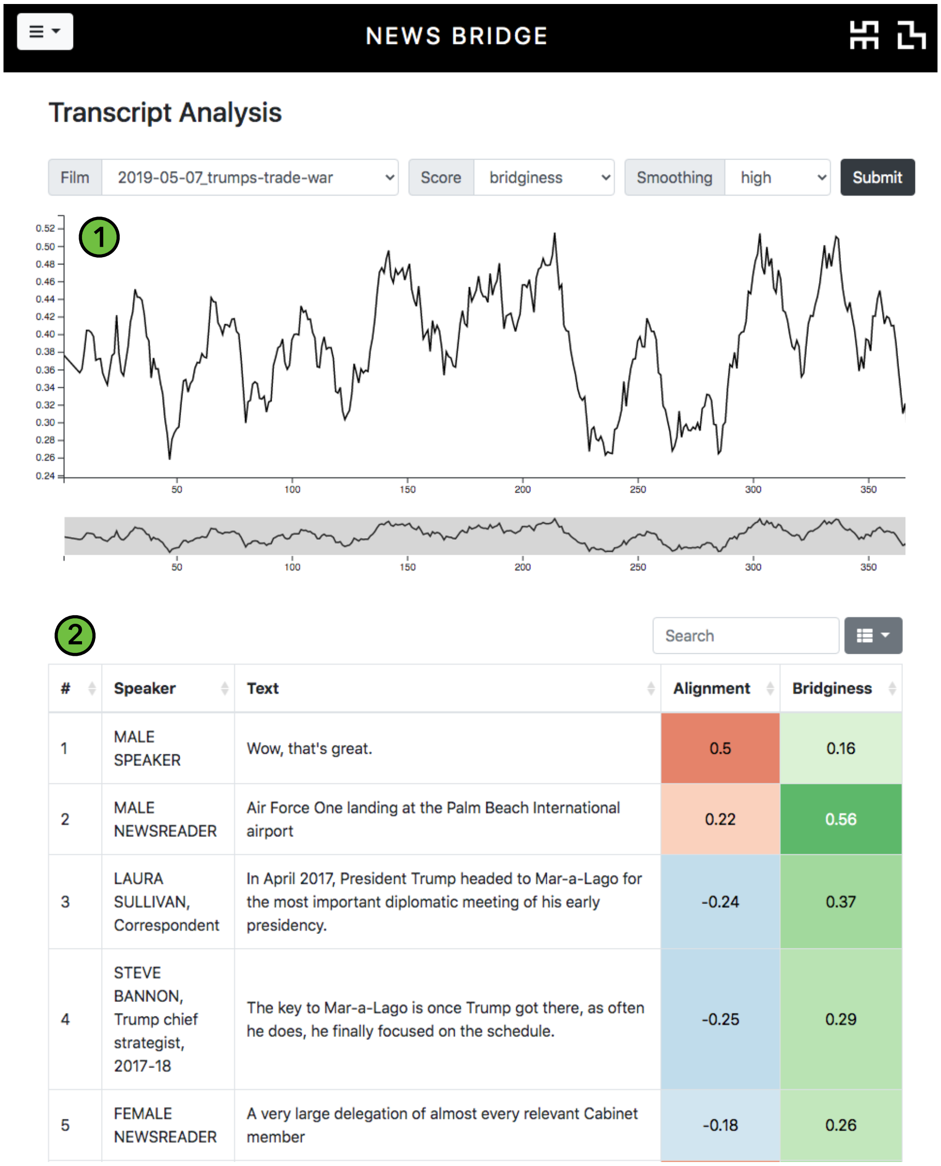
**SIMILAR HISTORICAL TWEETS.** One of BERT's main advantages is that it models the relationships between all the words in the sentence together. As a result, highlighting individual words is unlikely to fully explain its predictions. Therefore, in addition to providing word statistics, we also show similar historical tweets that were bridging. The goal is to show the user sample tweets that look similar to the model but have a higher bridginess score. To represent the tweets, we use the embeddings generated in the last layer of the BERT model. We save the embeddings of all tweets in the dataset and given the embedding of the input tweet, we find the nearest neighbors in the embedding space. To index and search the tweet embeddings efficiently, we use Faiss [49], a library for similarity search of dense vectors developed by Facebook Research. When the user clicks one of the rows in the results table, we show the ten most similar tweets to the input tweet, including when they were posted, by which outlet, how many retweets they received, and their bridginess score (Figure 2.10#4).

### 2.6.3 Transcript Analysis

To streamline the selection of bridging tweets, we also analyze the transcripts of the Frontline documentaries and show the results through an interactive interface. The tweets posted by the Frontline's Twitter account often include quotes from the documentaries and our models can be used to guide the selection of quotes or film segments that might be engaging to a politically diverse audience.

To analyze each transcript, we parse the transcript segments and use the BERT models to predict the expected bridginess and alignment of each segment. To provide an overview of the predictions, we plot the scores (bridginess/alignment) against the segment number (Figure 2.11#1). To make it easier to identify relevant regions of the transcript, we apply the Savitzky-Golay filter [81, 82] which smooths the curve by using local least-squares polynomial approximation. The user can select the level of smoothing (none, low, medium, or high smoothing), which corresponds to a different size of the window used to fit the local polynomial approximation. Below the plot, we show a table of the analysis results (Figure 2.11#2) that includes the segment number, the speaker (as specified in the transcript), the segment text, and the predicted bridginess and alignment. The user can sort the table based on the predicted scores and search for keywords in the text. Moreover, the plot at the top is interactive, and the user can zoom and focus on specific regions of the transcript. The table below automatically updates to show only the segments in the selected region.

The application includes transcript analyses of 125 films, from most recent ones to films going back to 2014. We received very positive feedback from the journalists about this feature of the web app.



Transcript Analysis

Film: 2019-05-07_trumps-trade-war
Score: bridginess
Smoothing: high

| # | Speaker | Text | Alignment | Bridginess |
|---|---|---|---|---|
| 1 | MALE SPEAKER | Wow, that's great. | 0.5 | 0.16 |
| 2 | MALE NEWSREADER | Air Force One landing at the Palm Beach International airport | 0.22 | 0.56 |
| 3 | LAURA SULLIVAN, Correspondent | In April 2017, President Trump headed to Mar-a-Lago for the most important diplomatic meeting of his early presidency. | -0.24 | 0.37 |
| 4 | STEVE BANNON, Trump chief strategist, 2017-18 | The key to Mar-a-Lago is once Trump got there, as often he does, he finally focused on the schedule. | -0.25 | 0.29 |
| 5 | FEMALE NEWSREADER | A very large delegation of almost every relevant Cabinet member | -0.18 | 0.26 |

**Figure 2.11:** News Bridge web application, a screenshot of the transcript analysis interface.

## 2.7 ADVERTISING EXPERIMENTS

Next, we test whether the predictive models we developed can be effectively used to compose more bridging tweets. In partnership with Frontline, we ran seven advertising experiments on Twitter between May and June of 2020. In each experiment, we selected a pair of tweets—one that was predicted to be more bridging (treatment) and one that was predicted to be less bridging (control)—and measured the engagement of left- and right-leaning users with each tweet. While the advertising experiments presented in this section are not randomized experiments (as we discussed in Section 2.2), they are the only way to run experiments on the platform and to measure how thousands of Twitter users respond to the test tweets. In the rest of this section, we explain how the experiments were set up and discuss the results.

### 2.7.1 Campaign Setup

One of the challenges of measuring the audience diversity of the test tweets is that the Twitter advertising platform only reports aggregate metrics of engagement and does not report the engagement levels by users with different political leanings. To address this issue, we run two campaigns for each test tweet—one targeting only left-leaning users and another targeting only right-leaning users—and measure the aggregate engagement of each group. Thus, in each experiment, testing a pair of tweets, we run four campaigns in total (Figure 2.12).

We ran all experiments for five days, Wednesday to Sunday, since Frontline airs new documentaries on Tuesday evenings[6]. We used "awareness" as a campaign objective, which unlike other advertising objectives, optimizes for reach and does not explicitly optimize for engagement. The idea behind setting this objective was to minimize the interference of the

[6] Except for one experiment, which we had to relaunch two days later after we discovered a misspelling in one of the test tweets.

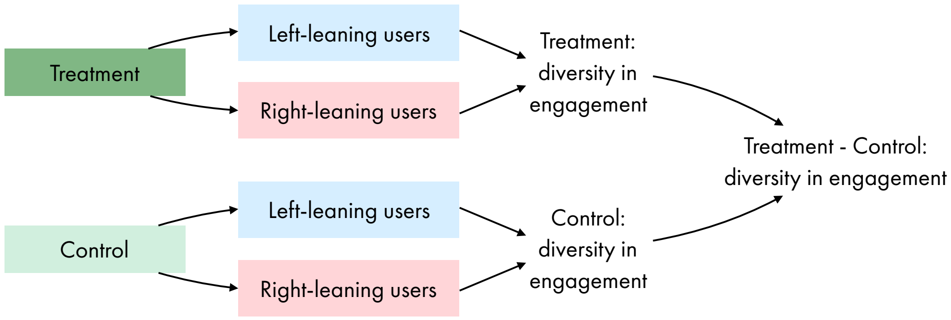

Treatment
Left-leaning users
Right-leaning users
Treatment: diversity in engagement
Control
Left-leaning users
Right-leaning users
Control: diversity in engagement
Treatment - Control: diversity in engagement

**Figure 2.12:** Advertising experiments setup.

advertising engine as much as possible such that the test tweets are not shown only to users who are likely to engage with them.

We set the budget for each campaign to $250 and capped the daily budget to $50 to spread out the experiments[7]. To make sure that the test tweets are shown to as many users as possible, we bid at $15/1000 impressions, significantly higher than the range recommended by the advertising platform, $3.5 - $6 / 1000 impressions. Also, to ensure that users in one group are not exposed to the advertisements more often, we limit the number of impressions per user to one.

We used "promoted-only tweets" which are shown only to the users selected to see the advertisements; they do not appear on the account (@frontlinepbs) page, the followers' timelines, or in the search results. This allows us to make sure that users are not exposed to both the treatment and control tweets. Promoted tweets are the same as organic tweets in every aspect except that they have a "Promoted" label at the bottom of the tweet.

To ensure that the test tweets are shown only to left-leaning or only to right-leaning users, we used "tailored audiences", i.e., we uploaded a list of user ids that the tweets could be shown to. Next, we discuss how we selected the audiences.

[7] Since we used tailored audiences (i.e., uploaded a list of user ids), we never reached the daily budget limit.

### 2.7.2 Audience Selection

To select the users who will be exposed to the test tweets, we consider only the followers of @frontlinepbs and the followers' followers (i.e., two hops away from the Frontline account[8]). Initially, we included only Frontline followers, but to make sure that enough users are eventually exposed to the tweets, we had to expand this set. Out of these users, we further restrict to users whose political alignment we can estimate based on their media-sharing patterns.

We randomly select 200k users for each experiment, 100k left- and 100k right-leaning users, and, within each group, we randomly assigned half of them to treatment and half of them to control (i.e., treatment-left: 50k, control-left: 50k, treatment-right: 50k, control-right: 50k). We make sure that by design, an equal proportion of @frontlinepbs followers vs. followers of followers is assigned to each treatment arm.

Once we make the assignments, we perform a number of balance checks to ensure that there are no systematic differences between users in different groups. We test for balance between three pairs of groups: (*i*) left-leaning users in treatment vs. control, (*ii*) right-leaning users in treatment vs. control, (*ii*) all users in treatment vs. all users in control. We consider the following user characteristics: number of posts, likes, followers, friends, tenure on Twitter, and the numerical estimate of their political alignment[9]. We run two types of covariate balance analysis. (*i*) We regress the user characteristics on the treatment assignment using logistic regression and ensure that none of the coefficients are statistically significant. (*ii*) We use a permutation test, i.e., we compare the log-likelihood of the logistic model regressing the user characteristics on the treatment assignment with its empirical distribution under random reassignments of treatment that fol-

[8] Due to the limits of the Twitter API, we did not include the followers of @frontlinepbs followers with more than 5,000 followers. They constituted only 3.85% of all @frontlinepbs followers.

[9] We logged the number of posts, likes, friends, and followers, since their distributions were highly skewed.

low the same randomization scheme. To obtain the empirical distribution, we measure the log-likelihood of 10k reassignments[10].

Once we have selected the audiences, we upload them to the Twitter advertising platform. While we upload a list of 50k user ids, only 15k-18k (i.e., 32%-37%) of them can be targeted. According to the documentation, inactive users are excluded from the tailored audience, but it is unclear what criteria are used to determine whether a user is active. Furthermore, when we use the tailored audiences to run a campaign, only 8.5k-9k (i.e., 17%-18%) of the users are shown the test tweets. There might be several factors that lead to this: (*i*) the users may simply be less active and have not logged on Twitter when we ran the campaign, (*ii*) the users might be very desirable and many other campaigns may have bid for them, (*iii*) Twitter's algorithm has predicted that they are less likely to engage with the tweet and has given less priority to our campaign. There could be other factors we are not aware of. Since we have no control over who will be exposed to the test tweets, this is where our experiments depart from A/B tests.

We note that left-leaning users are both more likely to be part of the tailored audience and more likely to be shown the test tweets. Among other reasons, this might be because they are more active or because the advertising engine predicts that they have a higher affinity to engage with Frontline content. While there are differences between the left- and right-leaning subgroups, we observe that when we consider the treatment and the control groups as a whole (i.e., combining the left- and right-leaning users of each group) both the size of the audience and the number of impressions are very similar.

### 2.7.3 Content Selection

All test tweets were composed by the journalists at Frontline. We were not involved in the selection process and only provided guidance on how

[10] We followed the procedure outline here: [59].

Table 2.2: Tweets used in the advertising experiments. Two tweets per experiment: Treatment (T) and Control (C) tweet.

| # | Tweet text |
|---|---|
| 1 / T | What does the response to COVID-19 look like in Washington DC, versus in Washington State — where the first U.S. case was confirmed? |
| 1 / C | As the coronavirus outbreak continues to spread, FRONTLINE traces the different responses by the federal govt. in DC and the govt. in Washington State, where the first known US case of the virus was identified in January. |
| 2 / T | Plastic has wide-ranging applications, from packaging to clothing to home furnishings. But when we're done with it, how much can actually be recycled? Journalist Laura Sullivan investigates. |
| 2 / C | It's estimated that no more than 10% of plastic has ever been recycled. FRONTLINE and NPR investigate why — and examine how plastic companies continue to promote recycling as environmentally friendly. |
| 3 / T | Get a rare and intimate window into health workers' battles against the coronavirus, from inside a hard-hit hospital in Italy in "Inside Italy's COVID War." |
| 3 / C | "I know my colleagues, they're struggling... They're like walking ghosts, because they don't want to have to make the decision whether to intubate someone or not." Go inside a hospital in Italy at the height of the coronavirus crisis. |
| 4 / T | Doctor Francesca Mangiatordi works inside an Italian hospital hard hit by the coronavirus. Her son says, she's "like Captain America." Follow her story in "Inside Italy's COVID War." |
| 4 / C | "There's a fear of becoming infected and in turn infecting the ones who are closest to me." - Dr. Francesca Mangiatordi, who works in a hospital in a region that's at the epicenter of Italy's coronavirus outbreak. |
| 5 / T | FRONTLINE PBS goes inside the U.S. response to the coronavirus in our documentary "Coronavirus Pandemic" |
| 5 / C | "I've covered science stories for nearly 30 years, but this felt more like science fiction." Veteran science reporter Miles O'Brien goes inside how the coronavirus outbreak unfolded in the United States. |
| 6 / T | "Once you start connecting the dots, you see that Amazon is building all of the invisible infrastructure for our futures," says quantitative futurist and author Amy Webb. |
| 6 / C | Jeff Bezos is the richest person on the planet. "Amazon Empire" explores his rise — and the rise of Amazon. |
| 7 / T | How is artificial intelligence disrupting life as know it — for good and for ill? FRONTLINE PBS investigates in "In the Age of AI." |
| 7 / C | China has set its sights on leading the world in artificial intelligence by 2030. Here's a look at the promise and perils of AI. |

to use our tools. The tweets posted by Frontline go through the same editorial scrutiny as other content published by Frontline and need to be approved before publication. None of the tweets used in the experiments were previously published, i.e., posted on Twitter.

All test tweets included a link to the relevant documentary and a promotional image. To avoid confounding effects, we used the same promotional image in both the treatment and control tweets. The timing of the experiments was also determined by Frontline's schedule. Since we ran the experiments in the midst of the coronavirus pandemic, four out of the seven experiments include tweets about documentaries related to the pandemic. We list all test tweets, without the promotional images, in Table 2.2.

### 2.7.4 Results

We measure the diversity of the audience engagement of the treatment and control tweets using the definition explained in Section 2.5.1. We focus on overall engagement and do not analyze the number of likes and retweets individually since there are too few such interactions to make meaningful comparisons. Figure 2.13 shows the results.

We find that in five out of the seven experiments, the treatment tweets achieved higher audience diversity than the control tweets, matching our model's predictions. The average difference in audience diversity between the treatment and control tweets is 0.005. The difference is not statistically significant ($p = 0.28$), which is not surprising given the small sample size ($n = 7$).

We observe that across the seven experiments, both the treatment and the control tweets have a high audience diversity, i.e., entropy values close to one (Figure 2.13A). This is partly due to the shape of the entropy function: around 0.5, small changes in the balance between left- and right-leaning users ($p_{left}$ and $p_{right}$) lead to even smaller increases in entropy. In mathematical terms, the slope of the first derivative of the entropy

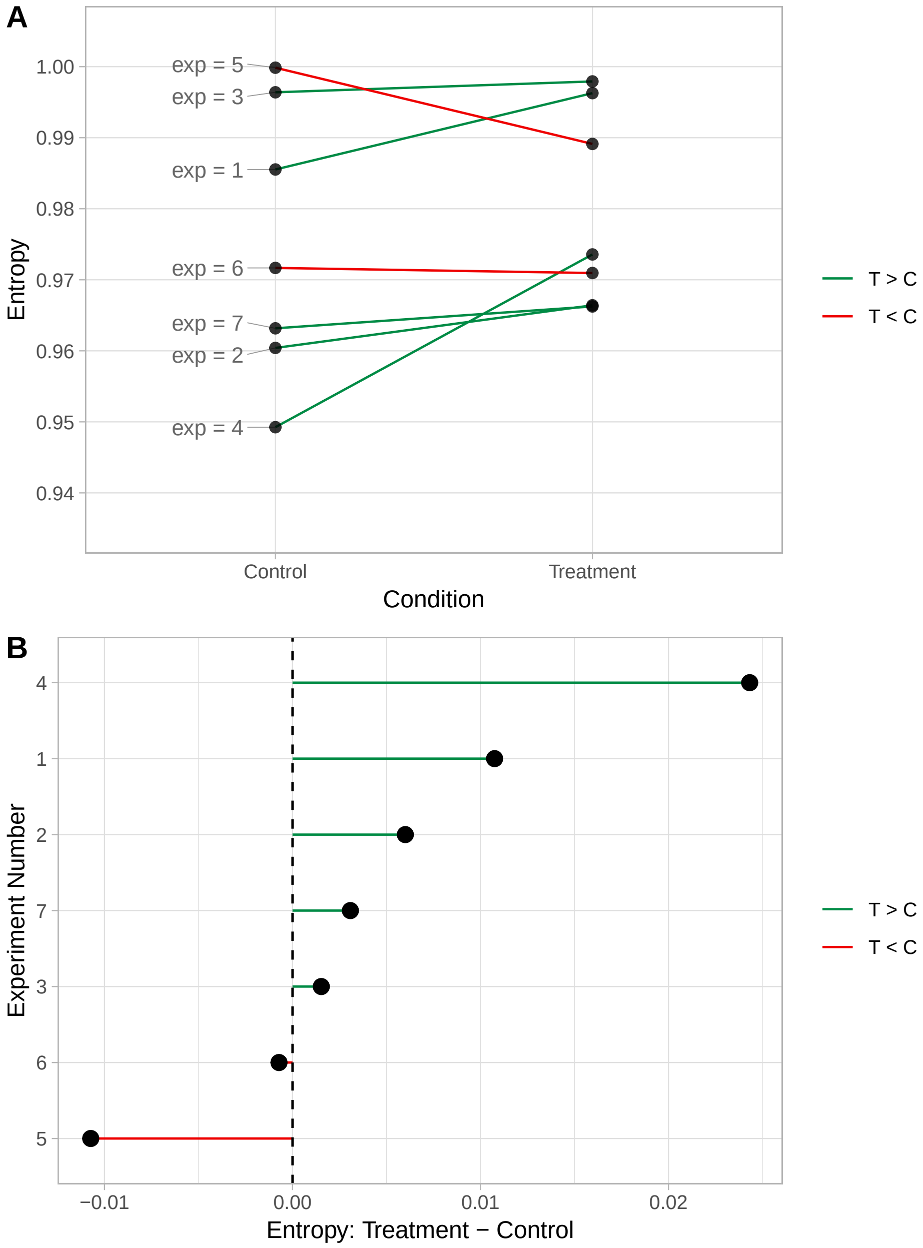


**Figure 2.13:** Comparison between the audience diversity of the treatment and control tweets in the advertising experiments.

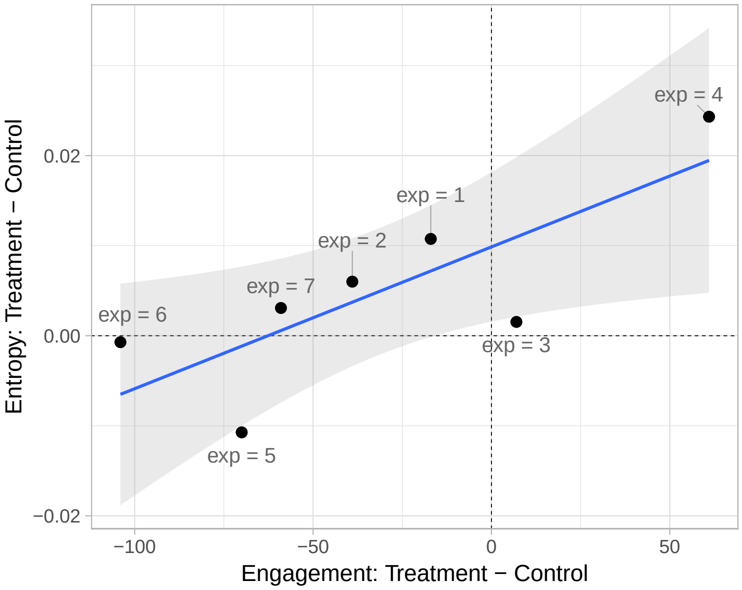


**Figure 2.14:** Comparison between the difference in engagement and audience diversity of the treatment and control tweets in the advertising experiments.

around 0.5 is much smaller than at the extremes (0 or 1). For instance, if the breakdown of engagement with the control tweet is $p_{left} = 0.45$ and $p_{right} = 0.55$ (entropy = 0.9927) and the treatment tweet achieves perfect balance in audience engagement, i.e., $p_{left} = 0.5$, $p_{right} = 0,5$ (entropy=1), then that would be an increase in entropy of only 0.007.

We also analyze the trade-off between engagement and audience diversity. While we have too small a sample size to make definitive conclusions, we find a positive correlation (Pearson $\rho = 0.79$, $p = 0.034$) between the difference in audience diversity and the difference in engagement between the treatment and control tweets (Figure 2.14). However, we note that while in five out of seven experiments, the treatment tweets achieved higher audience diversity, in only two out of seven experiments, they achieved higher engagement.

### 2.7.5 Limitations

While we are optimistic about the results of the advertising experiments, we would like to point out some of their limitations. The main caveat of our experiments is that the users who were exposed to the test tweets were not randomly selected.

We used different features of Twitter's advertising platform so that our experiments are as close as possible to randomized experiments: (*i*) we used tailored-audiences and promoted-only tweets to make sure that users are not exposed to both the treatment and control tweets; (*ii*) we limited the number of exposures per user to one to ensure that users in one group are not exposed to the test tweets more often; (*iii*) we placed high bids to increase our chances reaching as many users from our tailored audiences as possible. However, with the features currently available on the Twitter advertising platform, we were unable to remove the influence of the advertising engine, which optimizes the overall value of the platform. More specifically, due to algorithmic predictions or market forces, the advertising engine may show the test tweets to users who are more likely to engage with them, instead of a random subset of users. Thus, we cannot eliminate the possibility that the higher audience diversity of the treatment tweets is not due to differences in the tweet content, but due to differences in the delivery of the advertisements.

Rerunning the experiments using randomized assignment administered by the advertising engine—when such a feature is available on Twitter—is our key priority for future work.

## 2.8 CONCLUSION

In this chapter, we studied the relationship between the language used in tweets posted by news outlets and the political diversity of the users who engaged with them. We collected 566k tweets by five news outlets and

Frontline over more than three years. To measure each tweet's audience diversity, we compared the breakdown of left- vs. right-leaning users who shared the tweet. Using this data, we trained models that, given the tweet text, predict the audience diversity. We considered various ways of defining the prediction task and evaluated different prediction models. We then integrated the best model into a web application, which allowed Frontline journalists to craft more bridging tweets, guided by the model's predictions. Finally, in partnership with Frontline, we ran seven advertising experiments to test whether the model predictions can be effectively used to compose more bridging tweets. We found that in five out of the seven experiments, the tweets selected by our model were indeed engaging to a more politically diverse audience. While we are optimistic about the results of the experiments, we caution that the advertising experiments are not A/B tests and that we cannot rule out the possibility that the results are influenced by Twitter's advertising engine.

# 3 THE STRUCTURE OF TOXIC CONVERSATIONS

## 3.1 INTRODUCTION

With millions of people taking to social media to participate in public discussions, platforms like Twitter, Facebook, and Reddit have become the virtual public squares. They allow users to share their views and prompt conversations about issues that they care about. In the case of Twitter, a user can post a tweet and any user who sees the tweet can reply, thereby bringing the content into their own network of followers, sharing their point of view and reactions to the original tweet, and broadening the conversation. This chain reaction of replies propagates in complex ways through the Twitter network and can lead to conversational exchanges between people who may be tightly connected to one another online and in the real world, or equally to people who have never met and have little to no connection on Twitter. This potential for large-scale conversations across diverse sets of people holds promise for supporting a rich and vibrant public discourse, but also permits degradation of civility between people. A 2017 study by Pew Research study finds that 41% of Americans have been personally subjected to harassment online, and an even larger share (66%) has witnessed these behaviors directed at others [27].

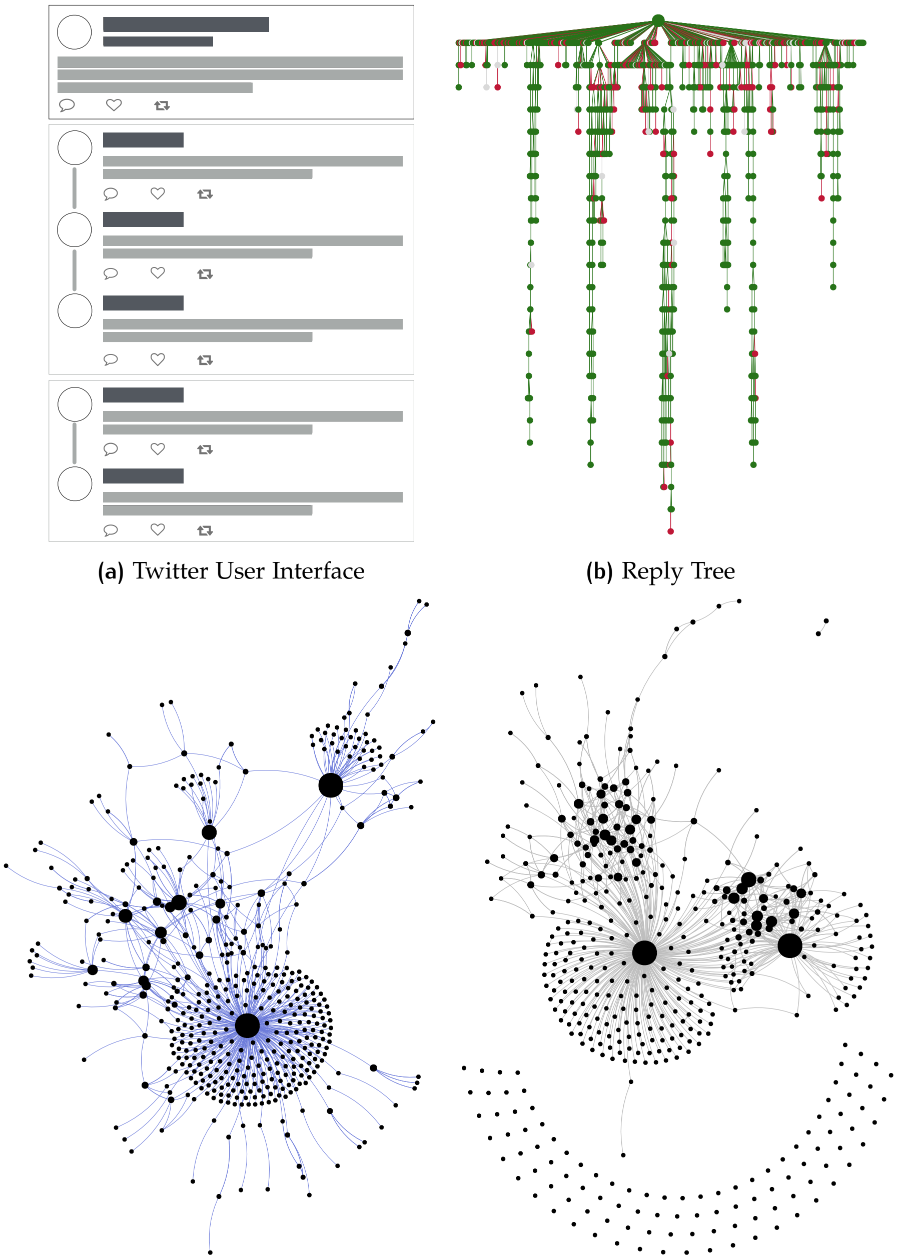
In this chapter, we investigate the relationship between toxicity and the conversational structure on Twitter. We focus on political conversations prompted by tweets that are posted by or mention five major news outlets (CNN, New York Times, Wall Street Journal, Fox News, and Breitbart) and 1,430 candidates running for office in the 2018 midterm elections in the

US. We collect a comprehensive sample of 1.18M conversations containing >58.5M tweets posted by >4.4M users.

We represent the structure of the conversation in three different ways: (*i*) using the reply tree (Figure 3.1b) which encodes the relationships between posts, where two posts are connected if one is a reply to the other; (*ii*) the reply graph (Figure 3.1c), a directed graph which captures the conversational interactions between users, where two users are connected if one replied to the other; and (*iii*) the follow graph (Figure 3.1d), which captures the social connections among the conversation participants, where one user is connected to another if they follow them.

The goal of this study is twofold: (a) to understand the relationship between the conversational structure and toxicity after the conversation has unfolded, and (b) to test the value of the structural view of the conversations in forecasting future toxicity, as the conversation unfolds. A greater understanding of the relationship between structure and toxicity of conversations can guide changes to the platform design to improve the health of the conversations, and the predictions of future toxicity can be used to moderate conversations automatically and prevent further toxicity. To study the link between structure and toxicity, we analyze the conversations at three levels: individual, dyad, and group level. To measure the predictive power of the structural characteristics of the conversations, we consider two prediction tasks. In the first task, we predict whether the conversation will become more or less toxic, based on the initial stages of the conversation. In the second task, we aim to predict whether the next reply posted by a specific user will be toxic, given the conversation so far and the user's relationship with the current conversation participants.

The rest of this chapter is organized as follows. In Section 3.2, we review previous work. In Section 3.3, we describe the data collection and in Section 3.4, we evaluate the tweet toxicity annotations. In Section 3.5, we analyze the structure of toxic conversations at the individual, dyad, and group level. We consider the task of predicting future toxicity in Section 3.6 and the task of predicting the toxicity of the next reply in Section 3.7.



**(a)** Twitter User Interface

**(b)** Reply Tree

**(c)** Reply Graph

**(d)** Follow Graph

**Figure 3.1:** Four views of a Twitter conversation started by a @foxnews tweet. **(a)** A sketch of the conversation as experienced by the conversation participants through the Twitter UI. **(b)** Reply tree, the root node is the tweet that prompted the conversation and the remaining nodes are replies. The red nodes represent tweets classified as toxic. **(c)** Reply graph, a user-centric view of the reply tree in which two users are connected if one replied to the other, and **(d)** the graph of the follow relationships between the conversation participants. The size of the nodes in the **c** and **d** is proportional to their PageRank.

## 3.2 RELATED WORK

We start by reviewing previous studies that have analyzed the structural and linguistic aspects of online conversations. Some analyze the general characteristics of the conversations while others focus on studying the characteristics that are most related to specific conversation outcomes, such as conversation growth or derailment.

**STRUCTURAL ANALYSIS.** The initial work on online discussions focused on discussion forums where users can comment on a post by adding a reply at the end of the discussion thread, or replying to one of the existing replies, creating a nested structure that can be represented as a tree. Gomez et al. [34] study Slashdot, a technology news forum, and report the distribution of different conversation tree characteristics, including number of posts/posters, width, and depth. Interestingly, they find that most conversations start with a wide first layer of comments, followed by an even wider second layer, and then exponentially smaller layers. Gonzalez-Bailon et al. [35] also studied Slashdot but focused on political discussions. They find that, compared to other discussion categories, political conversations tend to engage a larger number of participants and tend to have wider (i.e., have a large number of comments at any depth of the tree) and deeper (i.e., have more nested comments) conversation trees.

Backstrom et al. [3] develop models for algorithmic curation of online discussions on Facebook and Wikipedia. They focus on two prediction problems: (*i*) length prediction, i.e., predicting the total number of comments a discussion will reach, and (*ii*) re-entry prediction, i.e., predicting whether a user who already participated in the discussion will contribute again. Note that at the time of the study, Facebook comments had a linear structure, i.e., users could not start subthreads. The experiments were set up such that all prediction features are derived from a prefix of the conversation (e.g., the first five comments) and are used to predict how the conversation will unfold (e.g., whether it will grow larger than the median conversation size). They find that temporal features (e.g., the time it took for the first five comments to arrive) and arrival pattern features (e.g., the number of unique users with the first five comments) are most predictive of whether the conversation will grow significantly. They also find that discussions tend to be longer and that the users are more likely to return to the discussion if the first few users participating in the discussion are connected to each other in the social graph; however, they report that social features are high precision, but low recall.

In a more recent study, Zhang et al. [98] analyze online discussions on public Facebook pages. At the time of this study, Facebook had switched to a user interface that allows one-level threading, i.e., comments and replies (threads). They focus on predicting antisocial behaviors, in particular, whether the initiator of a conversation thread will be blocked or will block another conversation participant (i.e., prevent any further interactions with them) later in the thread, given the first ten replies in the thread. For each conversation thread, they create a graph where the nodes are users and the edges encode either reply or reaction (e.g., like) interactions and compute statistics of the degree, edge type, and subgraph distributions, which they use as input features in a predictive model. They find that the proportion of participants that post a reply, as opposed to a reaction, is positively correlated with blocking. In contrast, a higher propensity to react is negatively correlated with blocking.

Coletto et al. [21] build models that distinguish between controversial and non-controversial Twitter conversations, after the conversations have unfolded. They collect data on conversations started by the posts of 18 highly popular Twitter accounts and consider all conversations prompted by political/news accounts as controversial and sports/entertainment accounts as non-controversial. Using news vs. entertainment as a proxy for controversial and non-controversial is perhaps the main limitation of the study. To represent each conversation, they consider both the reply and the social graph among the conversation participants and compute the frequency of dyadic and triadic subgraphs, as well as temporal features.

They find that the temporal features (e.g., the average time between replies) and the frequency of dyadic subgraphs are most predictive and observe that the frequency of triadic subgraphs only marginally improves the predictive accuracy.

**LINGUISTIC ANALYSIS.** Zhang et al. [99] analyze Wikipedia's talk page discussion looking for linguistic cues that are predictive of whether a conversation will derail. Given the first non-toxic exchange (comment and a reply), they aim to predict whether the rest of the conversation will become toxic. (It is worth noting that the average length of the conversations in their dataset is 4.6 comments, thus one exchange, i.e., two comments on average constitute a large portion of a conversation.) To represent the first exchange, they use hand-crafted features that capture markers of politeness (e.g., positive: saying thanks, or negative: asking direct questions or starting a sentence with you/your/yours) and conversation prompt types (e.g., coordination, moderation, opinion) inferred by clustering conversations. They find that conversations in which the first comment poses a direct question or starts with a second person pronoun are more likely to derail. In contrast, conversations in which the initial exchange contains greetings or gratitude are less likely to become toxic. They build a model that, given a pair of first exchanges, predicts which conversation will derail. Using the politeness and prompt type features, they achieve an accuracy of 61.6%; while better than a bag-of-words baseline (56.7%), the relatively low performance demonstrates the difficulty of the task.

Chang and Danescu-Niculescu-Mizil [50] consider the same task of forecasting conversation derailment. They analyze two datasets: an extended version of Wikipedia's talk pages dataset [99] and conversations from the subreddit ChangeMyView using the moderators' interventions as indicators of conversation derailment. They draw on ideas from neural network approaches for training conversational agents [83, 85]. They model conversations using a hierarchical neural network model that first uses a Recurrent Neural Network (RNN) to embed each comment, modeling the sequence of comment tokens, and then models the context between individual comments with a second RNN that takes as input the sequence of comment embeddings. Since these approaches require a lot of training data, they first train the model on an unsupervised dialog generation task and then fine-tune the model on the supervised task, i.e., predicting conversation derailment. Unlike previous work, they do not use a fixed prefix (e.g., the first exchange) but make a prediction after every comment. They consider a prediction correct only when the model predicts a derailment before it happens. Their experimental results show that the neural network approach achieves an accuracy of 66.5% on Wikipedia and 63.4% on Reddit, significantly better than random guessing (50%). They also outperform the model based on hand-crafted linguistic features (Wikipedia: 58.9%, Reddit: 54.4% accuracy) described above [99]. Interestingly, they report that in 50% of the cases where the model made the correct prediction, it would have made the right call at least three hours before the conversation derailed, suggesting that this might give moderators enough time to intervene.

Hessel and Lee [41] build models that predict whether a Reddit post is controversial using data from 6 subreddits. Their approach combines both the linguistic and structural properties of the discussions. They consider a post as controversial if it receives both many upvotes and many downvotes. First, they evaluate the predictive power of using information available at the time of posting (post text, time, author) and find that combining time-related features and encoding the post text using a pre-trained neural language model (BERT [25], followed by mean-pooling and PCA) leads to accuracy between 65.3% and 69.3% across the 6 subreddits (random guess would lead to 50% accuracy). Next, they test how the text, rate, and tree structure of the comments posted in the early stage of the conversations (first 15 to 180 minutes) affect the predictive performance. They find that in 5 out of 6 subreddits, the performance significantly increases in less than three hours of observation. The improvement gains come primarily from the comments' text features, followed by the tree structure and the commenting rate features. Finally, they analyze how the models trained on one subreddit perform on another. They find that text features are most predictive in-domain but do not generalize well, while the structural and rate features have lower predictive power in-domain but transfer better across subreddits.

**THE PRESENT WORK.** In this work, we focus on studying how the structure of a conversation is related to toxic behavior, as opposed to predicting whether a conversation is controversial [21, 41] or whether the conversation initiator will block or will be blocked by another user [98]. In contrast to previous work, we analyze and model toxic behavior both at the individual and at the group (i.e., conversation) level. Moreover, unlike studies of discussion forums such as Slashdot [34, 35] or Reddit [41, 50], studying conversations on Twitter allows us to observe and analyze the social relationships among the conversation participants, in addition to the conversational structure (i.e., the reply tree). Finally, while we do not propose new methods for modeling the linguistic characteristics of the conversation to predict toxicity, we compare the predictive power of the structural features with features related to the content of the tweets, extracted using existing models.

## 3.3 DATA

**ACCOUNT SELECTION.** To capture a wide variety of political conversations, we collected conversations prompted by major news outlets and candidates who ran for office during the 2018 midterm elections in the US. We selected five news outlets that span the full political spectrum—New York Times and CNN on the left, Wall Street Journal in the middle, and Fox and Breitbart on the right [6, 13]—and have Twitter accounts with a large number of followers. We collected both the conversations started by tweets posted by these accounts and by tweets posted by others that @mention these accounts.

We tracked the news accounts for one year, from May 2018 to May 2019, capturing 510k conversations (32.6M tweets, 2.4M users), and the accounts of the midterm candidates for five months, one month leading up to the election and four months after, capturing 676.8k conversations (25.8M tweets, 2M users). Figure 3.2 shows the daily volume of conversations and Table 3.1 shows the summary statistics of the two datasets.

We followed both the personal accounts that the candidates used during their campaigns and their official accounts created after their inauguration[1]. We obtained the personal Twitter accounts of the candidates from Ballotpedia[2], and the official accounts from the `congress-legislators`[3] Github repository maintained by journalists from GovTrack, ProPublica, MapLight, FiveThirtyEight, and others. 1,430 of 3,339 candidates had a Twitter account.

Taken together, the two datasets include a large number of conversations over a long period of time. Moreover, the collected conversations vary in several important ways. They capture discussions prompted by a politically diverse set of accounts, including both left- and right-leaning news outlets and midterm candidates. Some conversations are started by highly influential accounts such as the news outlets and the candidates with a

[1] This explains the larger volume of conversations after January, 2019.
[2] `https://ballotpedia.org/`
[3] `https://github.com/unitedstates/congress-legislators`

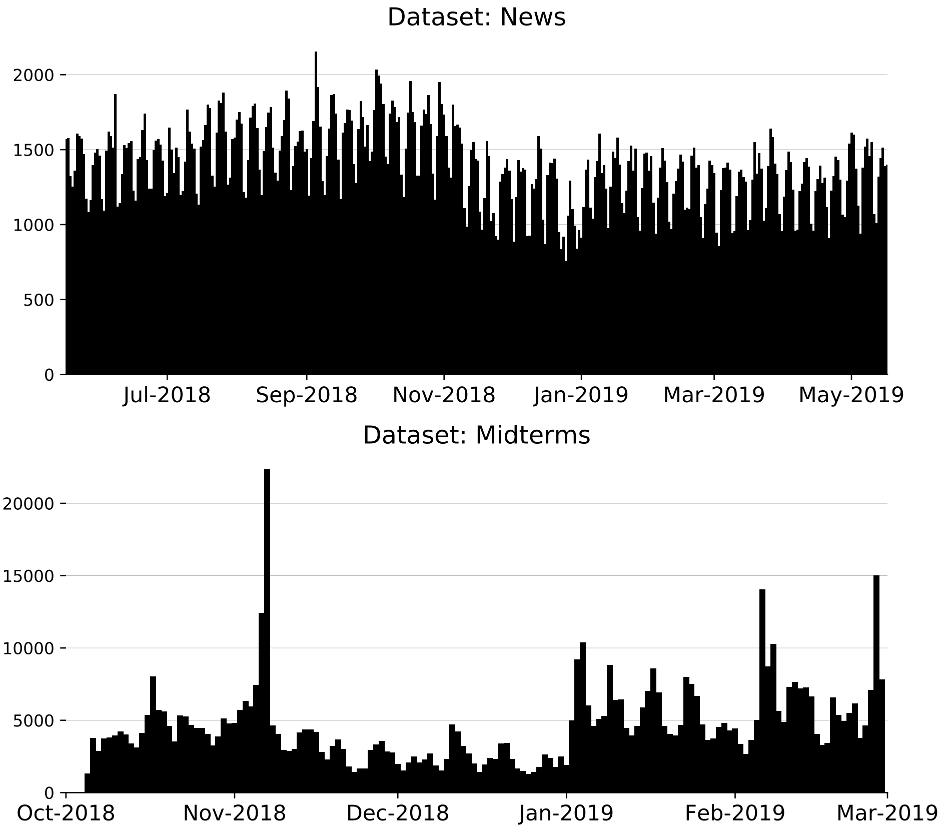


Figure 3.2: Number of conversations per day in the news and the midterms dataset.

large number of followers, others by ordinary users who @mentioned the news outlets or the candidates in their tweets. The data also reflects conversations on issues relevant at the local level through the candidates running for a seat in the House, at the regional level through the candidates for a seat in the Senate, and at the national level through the news outlets.

**DATA COLLECTION PIPELINE.** The key technical challenge in collecting tweets related to the same conversation is that the Twitter APIs only provide a link from the reply to the original tweet, but not vice versa. Thus, given a root tweet, one cannot simply query for all subsequent replies. To overcome this issue, we rely on the fact that every time a user replies to a tweet, they implicitly at-mention all users that posted or were mentioned earlier in the reply chain. (These mentions are considered a tweet prefix,

**Table 3.1:** Summary statistics of the news and the midterms dataset.

| **Dataset** | **\|Conversations\|** | **\|Users\|** | **\|Tweets\|** | **% Toxic Tweets** |
|---|---|---|---|---|
| News | 510,001 | 2,394,190 | 32,600,609 | 21.09% |
| Midterms | 676,839 | 2,013,918 | 25,874,622 | 20.22% |
| Total | 1,186,840 | 4,408,108 | 58,475,231 | 20.70% |

they are not part of the tweet body and do not count towards the character limit.)

To string together the reply and build the complete reply trees (Figure 3.1b), we scan the full dataset and use the reply-to field to recursively link posts to replies. We retain only reply trees rooted in tweets that are either posted by or @mention the selected accounts. As we are interested in studying conversations, we exclude tweets with no replies and strings of replies by only one user.

To collect the social graphs of the users who participated in these conversations, we set up a daily job that scans all tweets collected in the last 24 hours, compiles a list of all users that posted at least one tweet, and using the Twitter REST API downloads each user's list of friends and followers[4]. We do not collect data on accounts that are protected. Note that if the same user participates in multiple conversations over multiple days, we will have multiple snapshots of their friends and followers. This allows us to have an accurate and comprehensive view of the social connections at the time the users participated in the conversations. It is worth noting that limits on social graph endpoints of the Twitter API are the main reason that we limited the number of tweets/conversations we collected per day.

[4] The user's friends are outgoing edges, and the user's followers are incoming edges in the Twitter follow graph.

## 3.4 TOXICITY ANNOTATIONS

**PERSPECTIVE API.** To label tweets for toxicity, we used Google's Perspective API. We chose this API because its models are trained on Wikipedia comments, which like tweets, are short and informal. The initial Perspective API model was trained on 100K comments each annotated 10 times and was reported to be as accurate as the aggregate performance of three annotators [97]. Since then, the model has been retrained[5] on a larger dataset and modified to address some of its weaknesses reported by other researchers (e.g., [37]).

**EVALUATING THE TOXICITY PREDICTIONS.** Since the rest of our analysis relies on the Perspective API's toxicity annotations, we thoroughly assess their quality. To do so, we deployed an Amazon Mechanical Turk annotation task to obtain human toxicity labels on randomly selected tweets. Beyond assessing the quality of the annotations, we also relied on the human annotations to tune the Perspective API score threshold that we used for classifying a tweet as toxic or nontoxic. (The API returns a toxicity score, rather than a binary toxicity label.)

The Mechanical Turk annotation task consisted of five randomly selected tweets. We showed an input label next to each tweet for the annotators to select between "toxic" and "nontoxic." To avoid any annotation bias due to ordering effects, we randomized the order of the labels between tasks (i.e., batches of five tweets), but kept the order consistent within a task. To help clarify what constituted a toxic tweet, we provided the annotators with simple instructions. We used the same definition of toxicity as the Perspective API: "a rude, disrespectful, or unreasonable comment that may make you leave a discussio" [97]. To ensure the quality of the labels, we recruited only annotators from the US with high performance on previous Mechanical Turk tasks. We compensated them 20 cents per task (i.e., labeling five tweets). Before the annotators started the task, we warned

[5] We use the most recent version of the Toxicity model (TOXICITY@6), released in Sep, 2018.

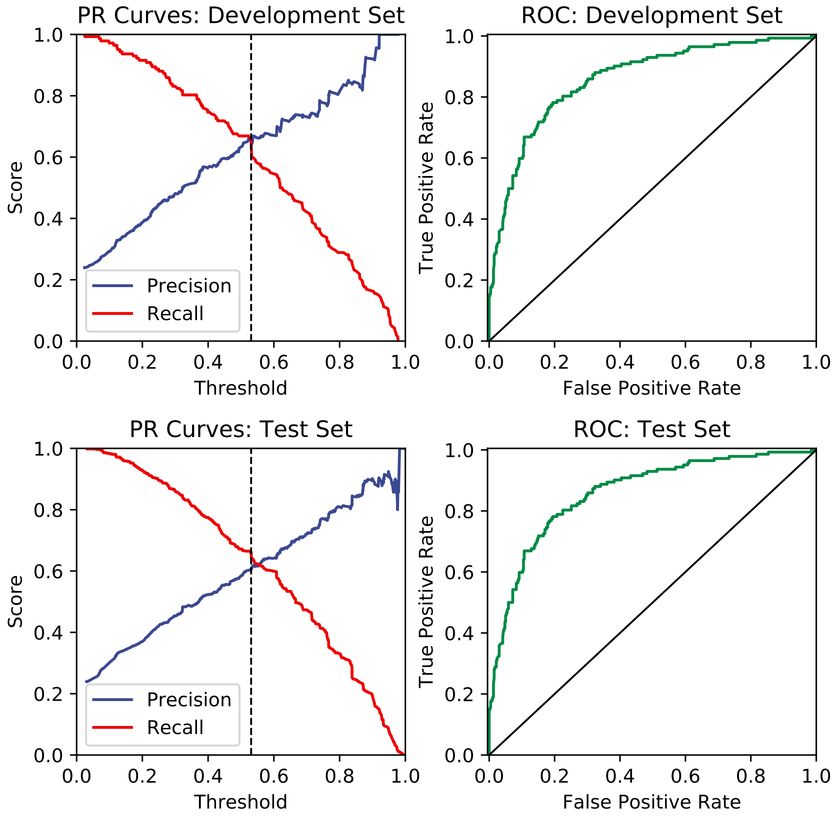


**Figure 3.3:** Precision-recall and ROC curves of Perspective API tweet toxicity classifier with respect to the Mechanical Turk tweet toxicity labeling obtained by a majority vote, for the development set (top) and test set (bottom). The vertical dashed line represents the chosen threshold, $T = 0.531$.

them that they might see offensive content. The protocol was approved by the MIT Institutional Review Board.

We randomly sampled 3,000 tweets for annotation from the first five months of the news dataset. We ensured that the sample is representative of the overall distribution of toxicity scores, as predicted by the Perspective API (K-S test, $D = 0.01$, $p = 0.89$). Each tweet was independently labeled by three different workers so that we can measure the inter-annotator agreement and use a voting scheme to obtain a single "ground-truth" label. To assess the inter-annotator agreement, we used Krippendorff's $\alpha$ [55] and found a fair agreement between the annotations, $\alpha = 0.32$. To obtain a single label for each tweet, we used a majority vote.

Table 3.2: Performance of the Perspective API toxicity classifier evaluated against the majority vote of three human labels.

| | Development | Test | Test (consensus) |
|---|---|---|---|
| Accuracy | 0.835 | 0.819 | 0.914 |
| AUC | 0.860 | 0.861 | 0.947 |
| Precision | 0.652 | 0.605 | 0.646 |
| Recall | 0.648 | 0.661 | 0.840 |
| F1 | 0.650 | 0.632 | 0.730 |
| N | 600 | 2,400 | 1,435 (59.8%) |

Next, we tuned the Perspective API toxicity score threshold above which we consider a tweet to be toxic, and measured the quality of the predictions. We used a random sample of 600 annotated tweets (20%) as a development set on which we chose the threshold, and the remaining tweets as a test set. We picked a threshold ($T = 0.531$) that strikes a balance between precision and recall on the development set. Figure 3.3 shows the Precision-Recall and ROC curves for both the development and the test set. On the test set, this threshold yields a classification accuracy of 0.82, AUC of 0.86, and an F1 score of 0.63. When we consider only the subset of the test set in which annotators reached a consensus, all measures of the prediction performance increase significantly, accuracy: 0.91, AUC: 0.95 , F1: 0.73. Table 3.2 summarizes all performance metrics.

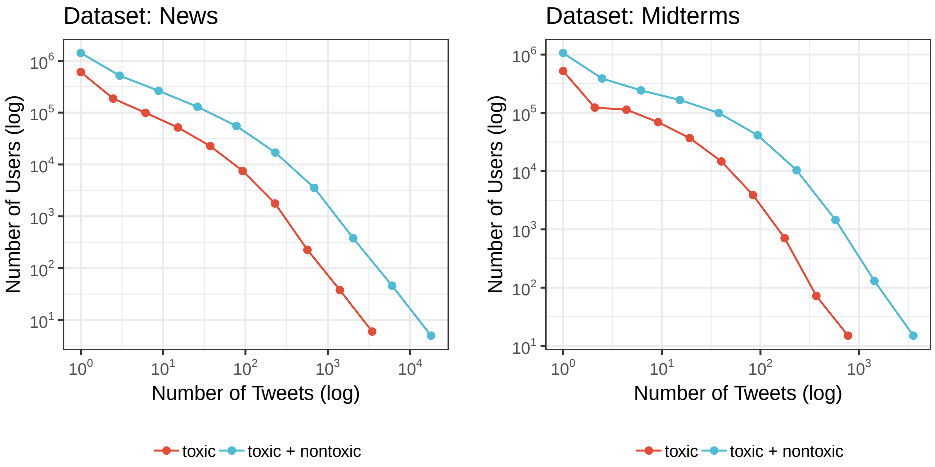


**Figure 3.4:** Distribution of the number of tweets and the number of toxic tweets per user.

## 3.5 ANALYSES

Next, we investigate conversations at multiple scales. First, we explore user characteristics and their propensity for toxic behavior. Second, we investigate the dyadic relationship between users by considering a tweet and the reply. Finally, we look at the overall conversational structure, including the tree of replies, the network of user replies, and their follow relationships.

### 3.5.1 Individual Level

We start by analyzing the distribution of tweets and toxic tweets per user in the two datasets. There are 32.6M tweets by 2.4M users in the news and 25.9M tweets by 2M users in the midterms dataset.

In Figure 3.4, we bucket users in logarithmically-sized buckets by the number of tweets and toxic tweets they posted ($x$ axis) and show the number of users that fall into each bucket ($y$ axis). As one may expect, we find that both distributions are long-tailed, i.e., there are many users who posted a few tweets and a few users who posted many tweets. Out of all

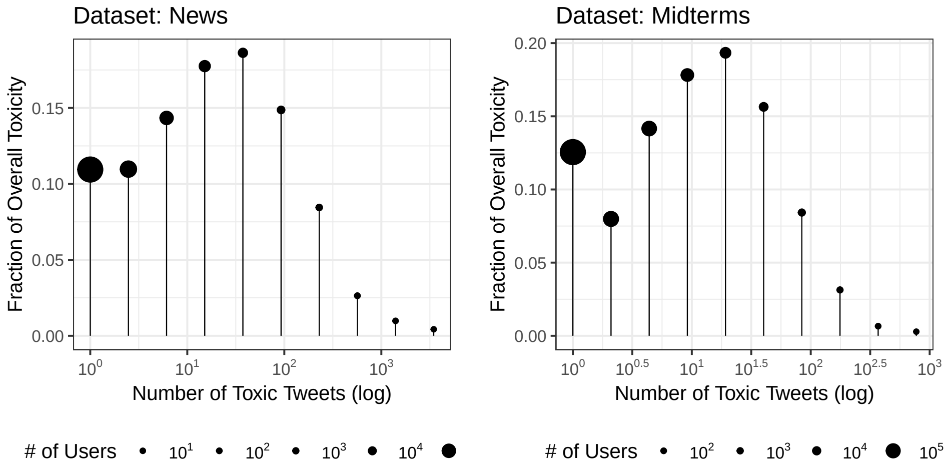


**Figure 3.5:** Fraction of toxic tweets contributed by the users in each toxicity bucket.

users, 44.71% in the news and 38.85% in the midterms dataset posted only one tweet. Most users—59.26% in the news and 56.15% in the midterms dataset—did not post any toxic tweets.

**DISTRIBUTION OF TOXICITY.** Next, we look at how the overall toxicity is spread across the users that posted at least one toxic tweet. In particular, we are interested in whether the toxicity is concentrated among a small number of users or dispersed across the population. This has important implications on how the platform might approach reducing toxic behavior. For instance, if only a small fraction of users are toxic, one may hope that changing their behavior or altogether removing them from the platform may disproportionately reduce the overall toxicity and significantly improve the experience for the rest of the users on the platform.

In Figure 3.5, we bucket users in logarithmically-sized buckets by the number of toxic tweets they posted (same as in Figure 3.4) and compute what fraction of toxic tweets (out of all toxic tweets in the dataset) was posted by users in each bucket. We scale the size of each point by the number of users that fall within each bucket to provide a visual reminder that the buckets contain a different number of users.

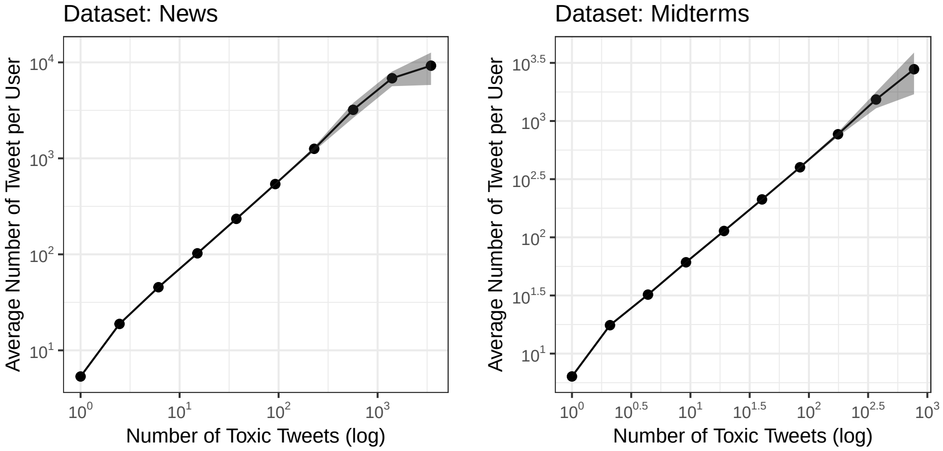


**Figure 3.6:** Average number of tweets (toxic and nontoxic) posted by users in each toxicity bucket.

We find a very similar pattern in the two datasets: buckets containing moderately toxic users account for the largest fraction of toxic tweets, ranging from 15% to 18% per bucket. While there are more users in the lower toxicity buckets, the higher number of toxic tweets per user in the medium toxicity buckets leads to a larger number of toxic tweets. This suggests that the toxicity is not concentrated among a few highly-toxic users, but it is rather dispersed across many low to moderately toxic users.

**ACTIVITY LEVELS OF TOXIC USERS.** We analyze how often users in each of the toxicity buckets participate in the conversations by posting reply tweets. Figure 3.6 shows the average number of tweets (toxic + nontoxic) posted by users in each toxicity bucket. In both the news and the midterms datasets, we find that users who post more toxic tweets also tend to post more tweets in general.

**RATE OF TOXICITY.** There are two main ways that the toxicity of a user can be characterized: (*i*) by looking at the absolute number of toxic tweets posted by the user, or (*ii*) by the fraction of toxic out of all tweets posted by the user. Each approach captures a different aspect of the user's behavior.

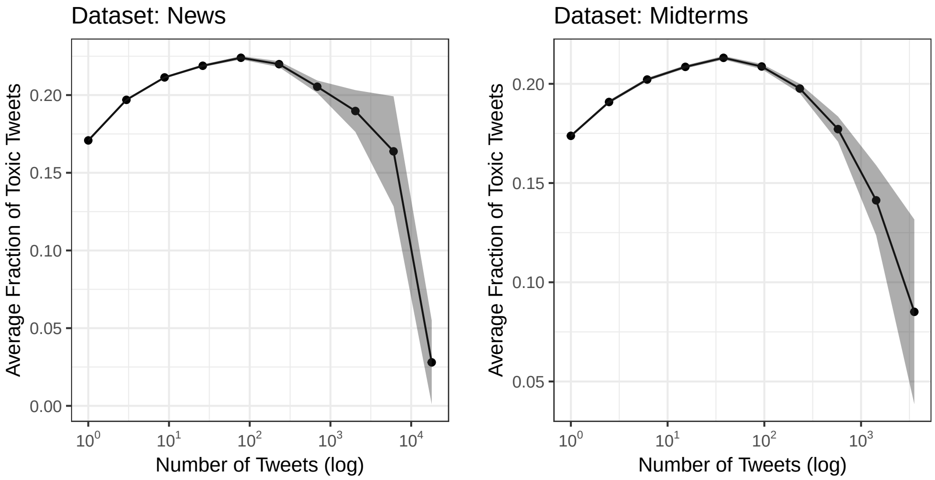


**Figure 3.7:** Average fraction of toxic tweets out of all tweets posted by users in each activity bucket.

So far, we have taken the first approach, bucketing users by how many toxic tweets they posted.

Here, we explore the second approach. In Figure 3.7, we bucket users in logarithmically-sized buckets by the number of tweets they posted and analyze how often their tweets were toxic. We find a similar pattern in both datasets, where moderately active users have a higher fraction of toxic tweets than both low- and high-activity users. We also find that highly active users have, on average, a smaller fraction of toxic tweets than lowly active users. Although, it is worth noting that the estimates for the buckets of highly active users have wider confidence intervals as fewer users belong to these buckets.

Similar to the analysis in Figure 3.5, where we computed contribution to the overall toxicity by users from different toxicity buckets, in Figure 3.8 we compute the contribution to the overall toxicity by users with different levels of activity, i.e., different number of tweets. In both datasets, we find that the toxic tweets posted by moderately active users account for the largest fraction of overall toxicity.

**HOMOPHILY.** We test whether there is homophily [66] among users within the Twitter follow graph, i.e., whether toxic users are more likely

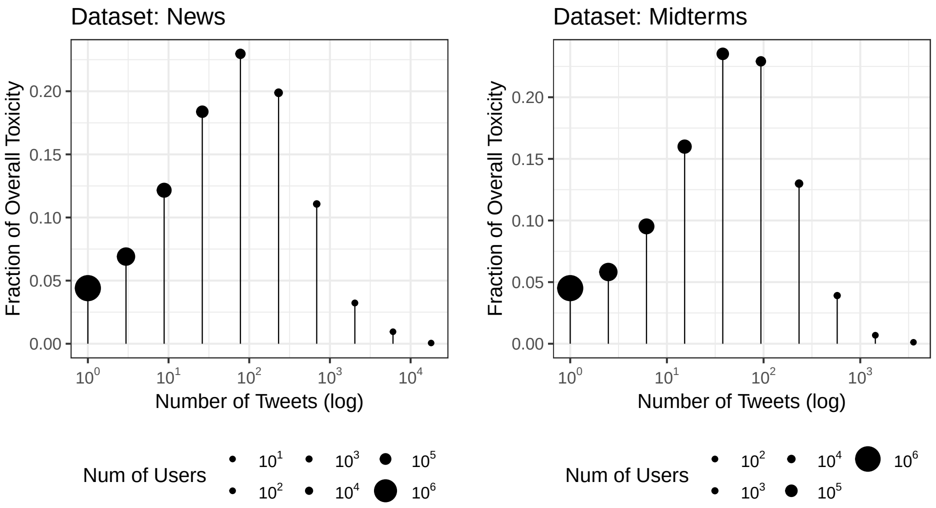


**Figure 3.8:** Fraction of toxic tweets contributed by the users in each activity bucket.

to follow other toxic users and whether nontoxic users are more likely to follow other nontoxic users.

To construct the complete follow graph among the users, we use the earliest snapshot of the user's friends (i.e., the Twitter users the user follows) captured right after they participated in a conversation for the first time. By doing so, we ensure that all of the users' follow connections were established before the users exhibited any toxic behavior captured in our dataset.

To measure the levels of homophily, we use the assortativity coefficient defined in [71] which quantifies whether nodes with the same attributes connect more or less often than we would expect by chance, i.e., in a random network. One attractive property of this coefficient is that it is defined both for categorical and numerical node characteristics. The assortativity coefficient can take values between -1: perfect disassortativity (i.e., users only connect with others that have a different characteristic from them) and 1: perfect assortativity (i.e., users connect only with others that have the same characteristics as they do).

We start by assigning users to two categories: (*i*) users that did not post any toxic tweets and (*ii*) users that posted at least one toxic tweet, and

Table 3.3: Measure of homophily (assortativity) among the users within the follow graph.

| Assortativity Coefficient Type | News | Midterms |
|---|---|---|
| **Categorical:** Users with 0 v.s. 1+ toxic tweets | 0.1496 | 0.1250 |
| **Categorical:** Users with 0 v.s. 4+ toxic tweets | 0.2280 | 0.1999 |
| **Numeric:** Num. of toxic tweets (All Users) | 0.0060 | 0.0336 |
| **Numeric:** Num. of toxic tweets (Users with 1+ toxic tweets) | 0.0047 | 0.0145 |

compute the corresponding assortativity coefficient (Table 3.3). We find that there is a low level of homophily among the users in the two datasets, 0.1496 in the news and 0.1250 in the midterms dataset. If we consider only users that did not post any toxic tweets and users that posted at least four toxic tweets, such that we exclude cases where users may be in the toxic category because a few of their tweets were misclassified, the assortativity coefficient increases slightly, but it still remains low, i.e., 0.2280 in the news and 0.1999 in the midterms dataset.

We also compute the assortativity coefficient among the users using the number of toxic tweets as an attribute (Table 3.3). This allows us to test whether users with many toxic tweets tend to follow other users with many toxic tweets. The resulting assortativity coefficients are very close to zero, 0.0060 in the news and 0.0336 in the midterms dataset. If we restrict the analysis only to users with at least one toxic tweet, the assortativity coefficients are even closer to zero. These results suggest that there is neither a positive nor negative affinity for highly toxic users to connect to other highly toxic users.

In summary, we find a low level of homophily among users with no toxic tweets and users with at least one toxic tweet. However, we find no evidence that highly toxic users (with many toxic tweets) are more likely to connect to other highly toxic users.

### 3.5.2 Dyads

Next, we focus on the relationship between toxicity and the characteristics of the reply dyads. A reply dyad $(i, j)$ consists of two conversation participants, user $i$ and user $j$, where user $j$ replied to user $i$'s tweet. We call user $i$ a parent (or a poster) and user $j$ a child (or replier), since $i$'s tweet is a parent of $j$'s tweet in the reply tree. Note that user $i$ might be a child in another dyad, e.g., $(x, i)$, or user $j$ might be a parent in a dyad $(j, y)$ (e.g., if a reply tree has a branch $(x, i, j, y)$).

We exclude reply dyads that are direct replies to the root tweet (i.e., have the news outlet as the parent user) since we are interested in understanding the relationship between the conversation participants rather than the relationship between the participants and the host news outlet. We also exclude self-replies, i.e., a chain of tweets posted by the same user, since the same user is both a dyad parent and the dyad child. This results in a total of 9.2M dyads in the news and 8M dyads in the midterms dataset.

**DYAD CHARACTERISTICS.** We define four dyad characteristics: (*i*) toxicity type, (*ii*) edge type, (*iii*) influence gap, and (*iv*) embeddedness. Each dyad can be characterized by whether the parent's post is toxic and whether the child's reply is toxic, leading to four possible dyad *toxicity types*. The dyad can also have one of four *edge types* depending on the relationship between the dyad users in the follow graph: (*i*) they may mutually follow each other (O=O), (*ii*) the child (replier) may follow the parent but not vice versa (O←O), (*ii*) the parent may follow the child (O→O), and (*iv*) they may not be connected at all (O O). (Note that, in our notation, the parent user is always on the left.) The dyad's *influence gap* is the ratio between the parent's and the child's number of followers. Finally, the *dyad embeddedness* measures the extent to which the social contexts of the dyad users overlap. We define it as the number of common friends between the dyad users, i.e., the number of users they both follow.

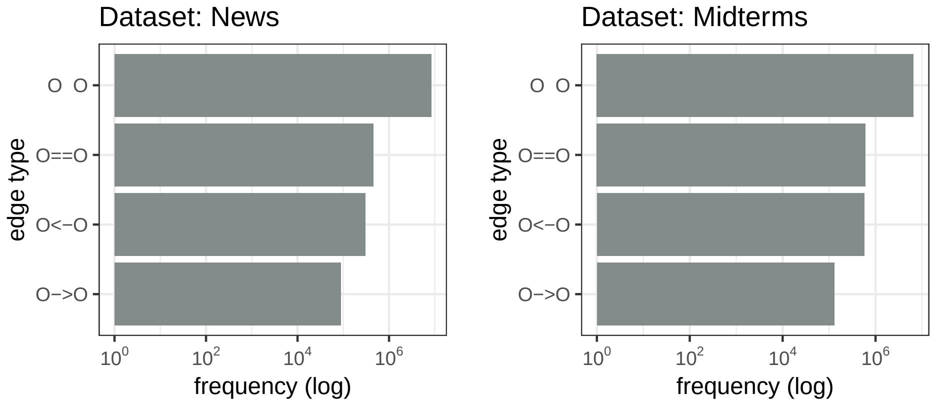


**Figure 3.9:** Frequency of edge types. Most replies occur between users who do not follow each other (O O).

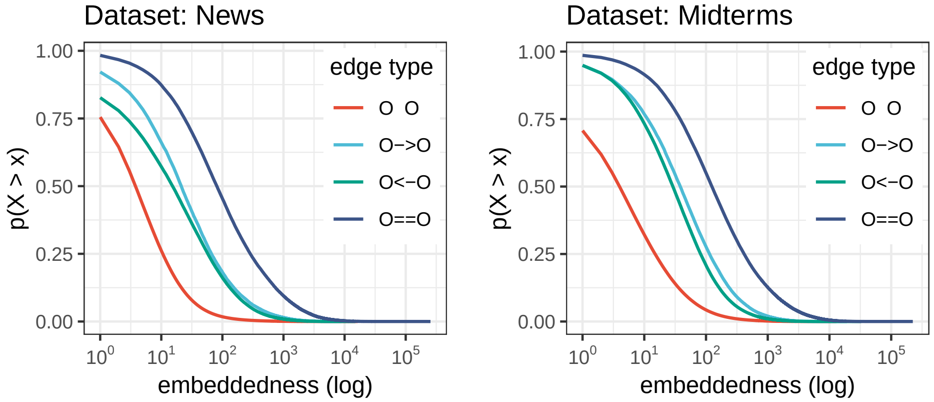


**Figure 3.10:** Complementary cumulative distribution of the number of common friends for each edge type. The $y$ axis shows the fraction of dyads that have more than $x$ common friends.

Figure 3.9 shows the frequency of dyadic interactions for each edge type. In both datasets, most dyadic interactions occur between strangers, i.e., users who do not have any follow relationship (O O, news: 90.8%, midterms: 83.4%), followed by cases where users mutually follow each other (O=O, news: 4.9%, midterms: 7.75%) or the replier follows the poster (O←O, news:3.31%, midterms: 7.24%). The most rare case is when the poster follows the replier, but not the other way around (O→O, news: 0.97%, midterms: 1.64%).

Figure 3.10 shows the complementary cumulative distribution of embeddedness, i.e., the number of common friends, for each edge type. We observe that in both datasets, dyad users who follow each other (O=O) are more likely to have more common friends. Dyad users who have a one-way follow edge (O←O or O→O) are less likely to have higher embeddedness than dyads who mutually follow each other, but are more likely to have higher embeddedness than users who do not follow each other (O O).

**TOXICITY TYPE.** We start by analyzing how the probability of a toxic reply varies depending on whether the parent post is toxic or not. We find that toxic tweets are 65% in the news and 64% in the midterms dataset more likely than nontoxic tweets to elicit toxic replies; the probability of a toxic reply given a toxic post is 0.3 in the news and 0.27 in the midterms dataset, while the probability of a toxic reply given a nontoxic post is 0.18 in the news and 0.16 in the midterms dataset. The toxicity type is the most defining characteristic of the dyad. We find that, in general, the patterns in other dyad characteristics differ significantly depending on whether the parent post is toxic or not. Therefore, in all subsequent analyses, we report how our findings differ in these two cases.

**EDGE TYPE.** Next, we look at how toxicity varies across different edge types. We find that the probability of a toxic reply varies significantly depending on the edge type (Figure 3.11). Given a toxic post, a toxic reply is more likely to come from another user who neither follows nor is followed by the parent user (news: 0.30, midterms: 0.28). The probability of a toxic reply among the other edge types (O=O, O←O, or O→O) is very similar, ranging between 0.22 and 0.24 in the news and between 0.2 and 0.24 in the midterms dataset.

Given a nontoxic post, it is more likely that a toxic reply will be posted by another user who does not have any follow relationship with the poster (news: 0.18, midterms: 0.17). However, in this case, the probability that a toxic reply comes from a user who follows the poster, but not vice versa (O←O), is higher (news: 0.158, midterms: 0.156) compared to the other two edge types (O=O, news: 0.12, midterms: 0.11; or O→O, news: 0.10,

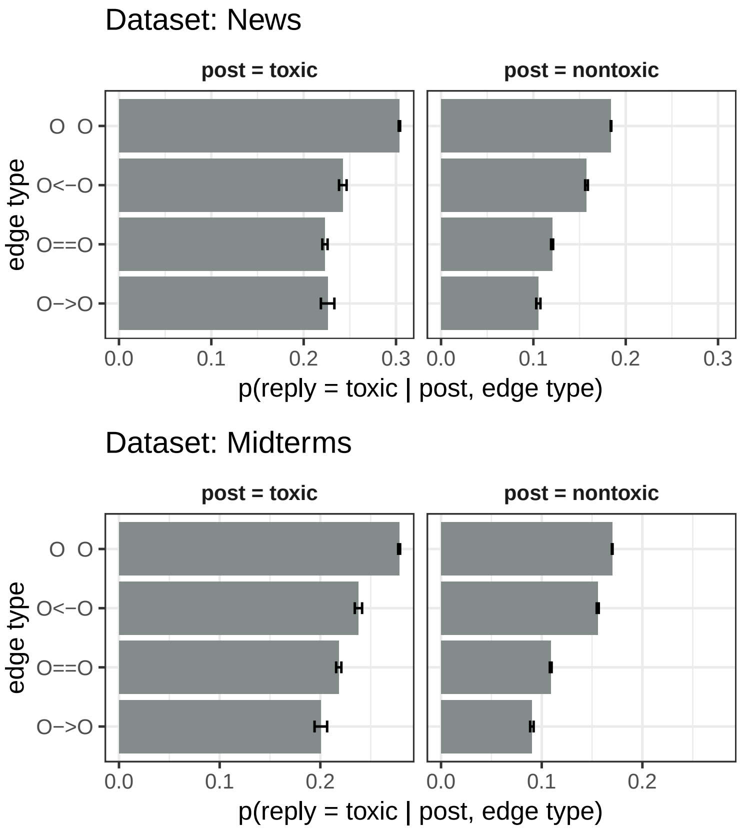


**Figure 3.11:** Probability of a toxic reply given a toxic (left) or nontoxic (right) post depending on the edge type between the users in the dyad. Toxic replies are more likely to be posted by users who do not have any follow relationship with the poster (O O), or, in the case of nontoxic posts, from users who follow the poster but are not followed back (O←O).

midterms: 0.09). This suggests that more influential users are more likely to be a target of toxic replies. We investigate this hypothesis next.

**INFLUENCE GAP.** We define the influence gap as the ratio between the parent's and the child's number of followers. Since the distribution of the number of followers is long-tailed, we compute the log of the ratio: $\log_{10}$(|parent's followers|) - $\log_{10}$(|child's followers|). Although most dyadic interactions occur among users with a similar number of followers, users are more likely to reply to tweets posted by others who have more followers than they do. In the news dataset, when the parent's post is

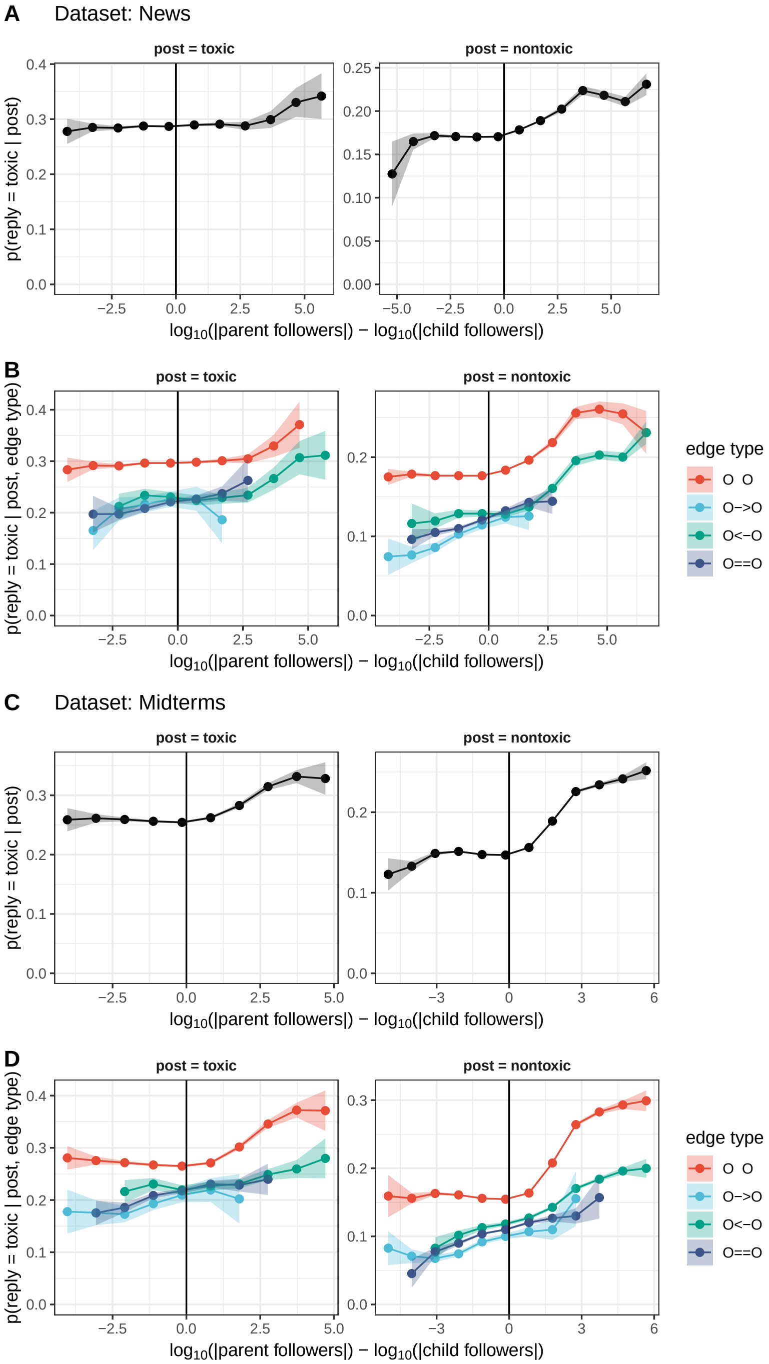


**Figure 3.12:** Probability of a toxic reply as a function of the *influence gap*, i.e., the log difference between the poster's and replier's number of followers. **(A, C)** Given a nontoxic post, a toxic reply is more likely to be posted by a user who has less followers than the poster. **(B, D)** This phenomenon is most pronounced in dyads where the users do not follow each other (O O), or the replier follows the poster but not vice versa (O←O).

toxic, the probability of a toxic reply is roughly the same, regardless of the influence gap (Figure 3.12A). In contrast, in the midterms dataset, the probability of a toxic reply increases when the parent has more followers than the child (Figure 3.12C). When the parent's post is nontoxic, then the influence gap matters even more. In both datasets, the probability that a reply will be toxic is higher when the parent has significantly more followers than the child. Interestingly, this relationship is asymmetric, i.e., the probability of a toxic reply does not decrease when the child has more followers than the parent.

We also observe a strong heterogeneity in the effect of the influence gap among the four edge types (Figure 3.12, B and D). Both when the parent is toxic and when it is nontoxic, the effect is most pronounced among dyads where the two users do not have any follow relationship (O O) and when the replier follows the poster but not vice versa (O←O).

**EMBEDDEDNESS.** We define the embeddedness of a dyad as the number of common friends between the poster and the replier. Higher embeddedness suggests that the two users have similar interests and overlapping social contexts. This, in turn, may influence the behavior of the replier: their potentially toxic behavior is more likely to be observed by others that both poster and the replier are mutually aware of and may increase the social cost of toxic behavior [20].

We find that the probability of a toxic reply significantly decreases as the embeddedness increases (Figure 3.13, A and C). This is the case regardless of whether the parent post is toxic or not. Given a toxic post, the probability of a toxic reply is 11% lower in the news (dropping from 0.315 to 0.206) and 9% lower in the midterms dataset (dropping from 0.29 to 0.2) if the poster and the replier have 100 vs. 1 common friend. Similarly, given a nontoxic post, the probability of a toxic reply goes down from 0.191 to 0.134 in the news and from 0.178 to 0.123 in the midterms dataset when the dyad users have 100 vs. 1 common friend.

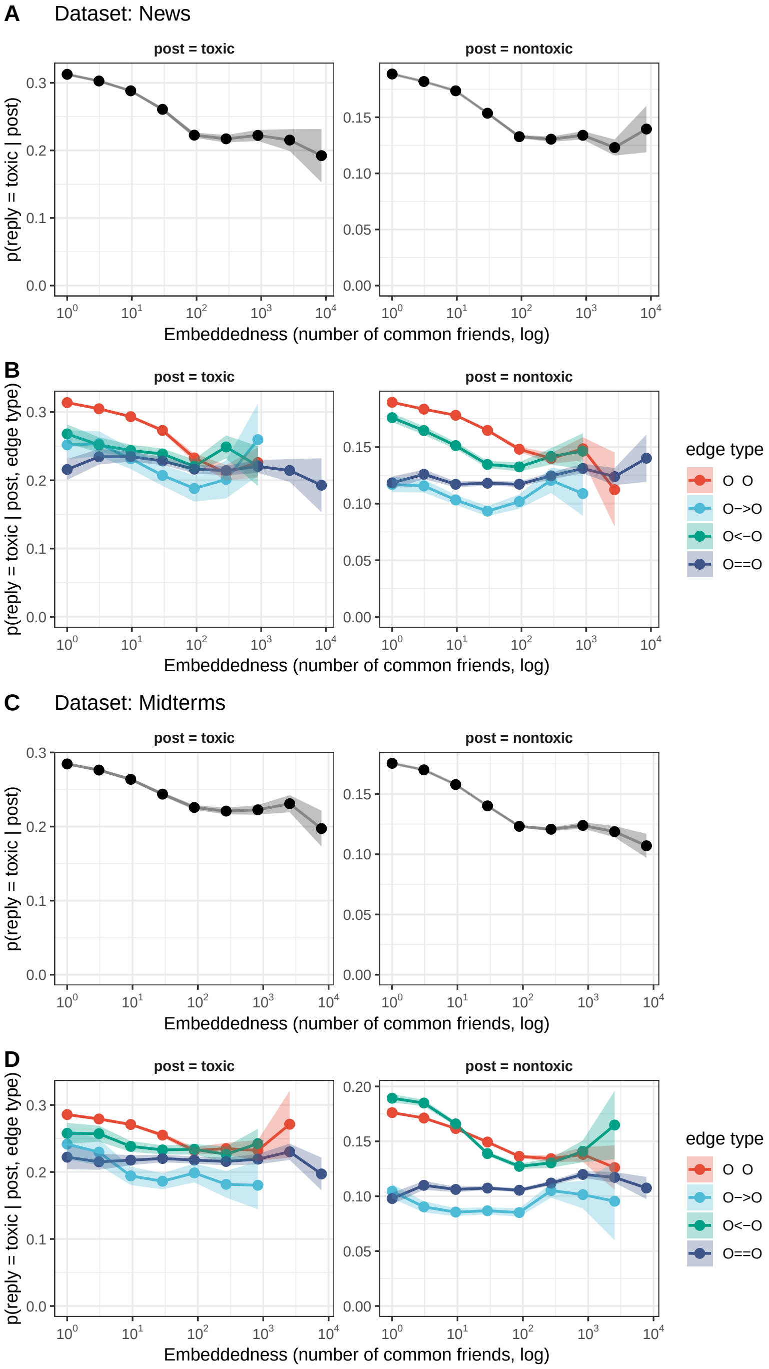


**Figure 3.13:** Relationship between *embeddedness*, i.e., the number of common friends, and the probability of a toxic reply. **(A, C)** A toxic reply is less likely to be posted by users who share more common friends with the poster, regardless of whether the post is toxic or not. **(B, D)** This effect is most pronounced among user dyads who do not follow each other (O O) and dyads where only the replier follows the poster but not vice versa (O←O).

Figure 3.13 (B and D) shows the relationship between embeddedness and the probability of a toxic reply for each edge type. In the news dataset, we find that given a toxic post, the probability of a toxic reply decreases as the embeddedness increases only for dyads where the users do not follow each other (O O). This effect is less pronounced in the midterms dataset. Given a nontoxic post, embeddedness is negatively correlated with the probability of a toxic reply, both when the dyad users do not follow each other (O O) and when only the replier follows the poster (O←O). This effect is consistent across the two datasets.

**BIFURCATIONS.** To further examine the robustness of the results presented so far, we repeated all analyses but only considered the dyads where bifurcations occurred. Bifurcations, places where the reply tree splits and the parent tweet has more than one child/reply tweet, allows us to measure the correlation between the dyad characteristics and toxicity while holding everything else constant. While the exact point estimates are slightly different and the confidence intervals larger[6], we find the same substantive results.

[6] Due to the smaller number of data points.

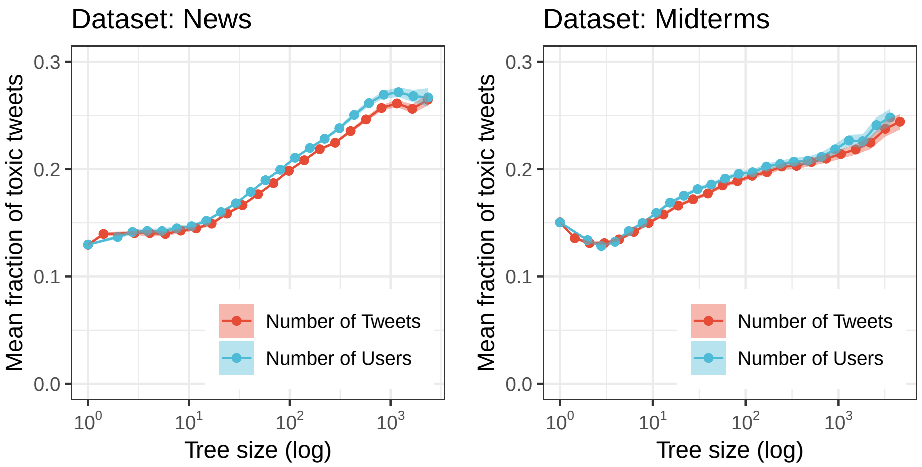


**Figure 3.14:** Relationship between reply tree size (number tweets and number of users) and the mean fraction of toxic tweets in the conversation. Larger conversations tend to be more toxic.

### 3.5.3 Reply Tree Structure

When a user posts a tweet, other users may choose to post a reply tweet, which in turn can lead to subsequent replies. The result is a reply tree of tweets, rooted in the original tweet (Figure 3.1b). Here, we investigate the relationship between the structural characteristics of the reply tree and the overall toxicity of the conversation. We define the toxicity of a reply tree as the fraction of toxic tweets. We note that the results presented are also consistent with a slightly different definition, in which we compute the mean or the median of the toxicity scores.

**TREE SIZE.** First, we consider the size of the tree in terms of the number of tweets and the number of users who posted the tweets. We find a clear, positive relationship between these two measures of tree size and its toxicity. As shown in Figure 3.14, larger trees tend to be more toxic both in the news and the midterms dataset. While the two size metrics (number of tweets and number of conversation participants) are strongly correlated (news: $\rho = 0.98$, midterms: $\rho = 0.99$), we present them both since they reflect slightly different aspects of a conversation.

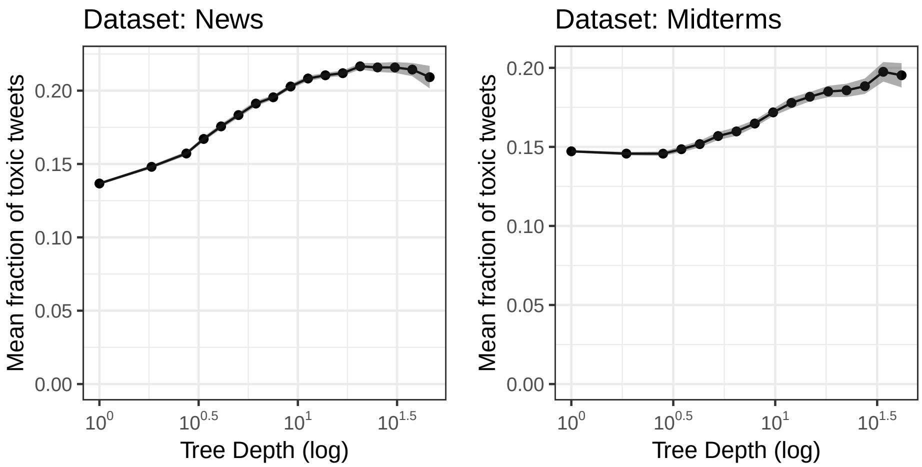


Figure 3.15: Relationship between reply tree depth and the mean fraction of toxic tweets in the conversation. Conversations with deeper reply trees tend to be more toxic.

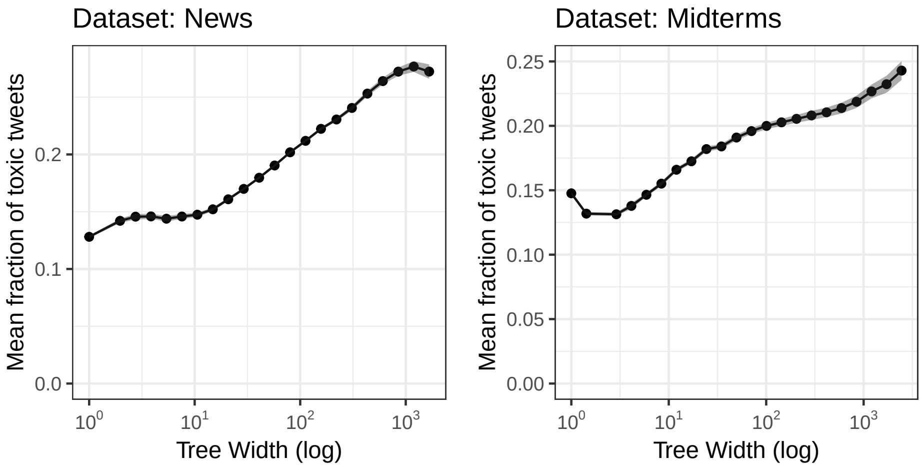


Figure 3.16: Relationship between reply tree depth and the mean fraction of toxic tweets in the conversation. Conversations with wider reply trees tend to be more toxic.

**TREE DEPTH AND WIDTH.** Next, we consider tree depth and width, which we use as summary metrics of the tree structure. We define the tree depth as the depth of its deepest node, and tree width as the maximum number of nodes at any depth in the tree. Using these two metrics, we find that both wider and deeper trees tend to be more toxic, as shown in Figures 3.15 and 3.16. This pattern holds in both the news and the midterms dataset. It is worth noting that both of these metrics are positively correlated with tree size (news: $\rho_{depth} = 0.53$, $\rho_{width} = 0.97$; midterms: $\rho_{depth} = 0.48$, $\rho_{width} = 0.97$) and may be proxies for size. Later in this chapter, we will evaluate their usefulness as features in a predictive task.

**WIENER INDEX.** While tree size, depth, and width summarize important aspects of a tree, we present an additional metric that helps us characterize the internal structure and complexity of a tree. The Wiener index $w(T)$ of a reply tree $T$ is defined as the average distance between all pairs of nodes; that is for $n > 1$ nodes,

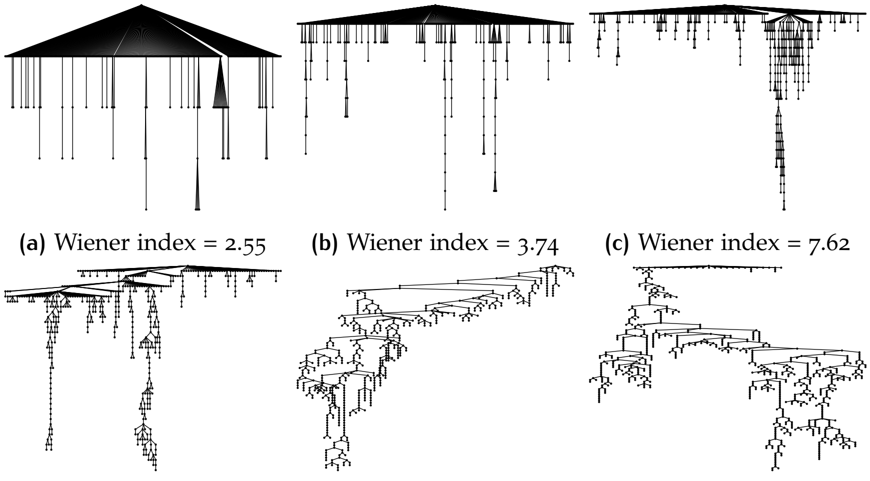
$$w(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij},$$

where $d_{ij}$ denotes the length of the shortest path between nodes $i$ and $j$. The Wiener index was initially proposed in mathematical chemistry to characterize the structure of a molecule [95]. More recently, it has been used to characterize the structure of information diffusion cascades, and in particular to quantify whether information spreads in broadcast or viral (person-to-person) fashion [33].

Figure 3.17 shows six sample reply trees of size between 250 and 1,000 from the midterms dataset with Wiener index raging between 2.55 and 46.1. As it can be observed from the figure, the Wiener index interpolates between two extremes: reply trees in which participants only respond to the original tweet and do not engage with each other (low $w(T)$), and reply trees with a single branch in which participants have many back-and-forth exchanges (high $w(T)$).

In the news dataset, we find that reply trees with a larger Wiener index tend to be more toxic (Figure 3.18A). In the midterms dataset, the mean toxicity of reply trees with varying Wiener index is mostly the same, except for a small fluctuation for trees with a low Wiener index (Figure 3.18B).

A more complicated picture emerges when we plot the relationship between the Wiener index and toxicity for reply trees of different sizes. In Figure 3.18 (C and D), we divide all reply trees into five logarithmically-



(a) Wiener index = 2.55 (b) Wiener index = 3.74 (c) Wiener index = 7.62

(d) Wiener index = 16.8 (e) Wiener index = 34.4 (f) Wiener index = 46.1

**Figure 3.17:** Sample reply trees with different values of the Wiener index.

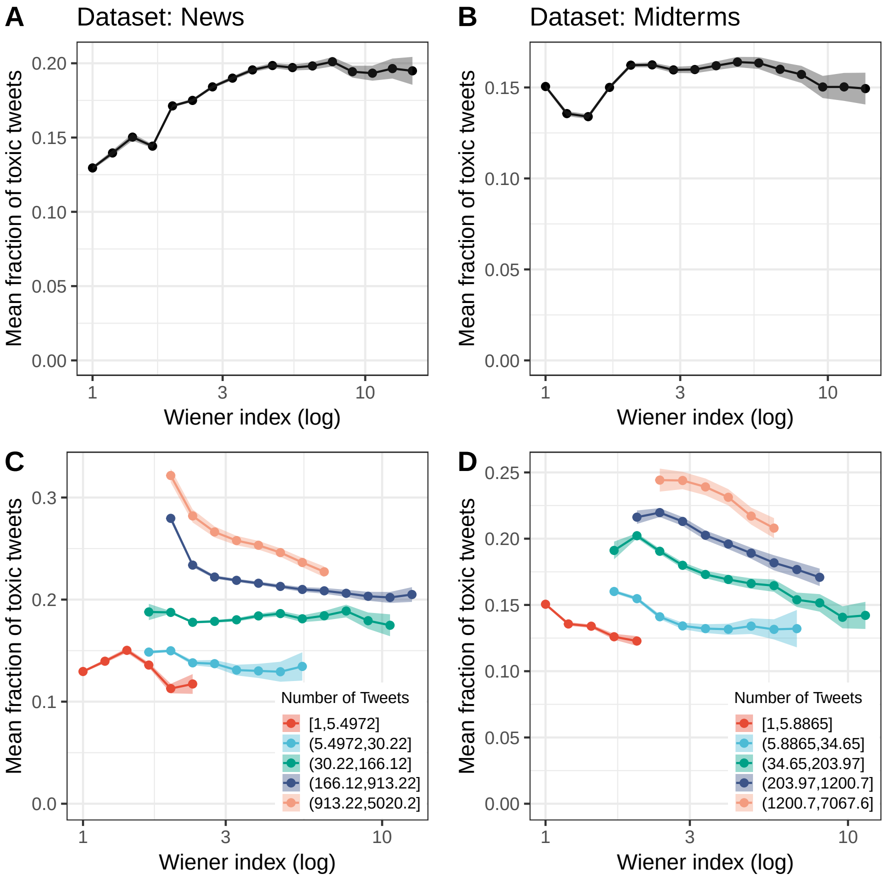


**Figure 3.18:** The relationship between the Wiener index of the reply tree and the mean fraction of toxic tweets, in overall (A and B) and by tree size (C and D).

sized groups according to their size. (We chose the largest number of groups that will leave us with enough data points to compare the relationship between the Wiener index and size.) In the news dataset, we find that for smaller reply trees, the toxicity of the conversations does not vary as a function of the Wiener index; however, for larger reply trees, we find that the toxicity decreases as the Wiener index increases (Figure 3.18C). In the midterms dataset, the fraction of toxic tweets in the conversation decreases as the Wiener index increases for all tree sizes, although the negative correlation is stronger for larger trees (Figure 3.18D).

We use regression analysis to further investigate this phenomenon (Table 3.19). We regress the fraction of toxic tweets as a function of (*i*) number of tweets (i.e., tree size), (*ii*) Wiener index, and (*iii*) both number of tweets and Wiener index. (We make $log_{10}$ transformations of all independent variables.) We find that, when regressed individually, the number of tweets and the Wiener index are both positively correlated with the fraction of toxic tweets (Table 3.19, Models *i* and *ii*). However, we when we regress both variables together, the coefficient for the number of tweets remains positive, but the coefficient for the Wiener index becomes negative (Table 3.19, Model *iii*). This pattern is consistent across the news and the midterms dataset and confirms our previous analysis (Figure 3.18).

The tweets that constitute reply trees make up an important part of the discourse on Twitter. Through the affordances provided by the platform, users can contribute their commentary to a discussion prompted by an initial tweet. No single user shapes the emergent reply tree structure, yet there are consistent relationships between the tree structure and the toxicity of its content.

| | *(i)* | *(ii)* | *(iii)* |
|---|---|---|---|
| (Intercept) | $0.12^{***}$ | $0.13^{***}$ | $0.13^{***}$ |
| | (0.00) | (0.00) | (0.00) |
| $log_{10}$(n tweets) | $0.03^{***}$ | | $0.05^{***}$ |
| | (0.00) | | (0.00) |
| $log_{10}$(wiener index) | | $0.10^{***}$ | $-0.09^{***}$ |
| | | (0.00) | (0.00) |
| RMSE | 0.19 | 0.19 | 0.19 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

**(a)** News

| | *(i)* | *(ii)* | *(iii)* |
|---|---|---|---|
| (Intercept) | $0.14^{***}$ | $0.15^{***}$ | $0.15^{***}$ |
| | (0.00) | (0.00) | (0.00) |
| $log_{10}$(n tweets) | $0.02^{***}$ | | $0.05^{***}$ |
| | (0.00) | | (0.00) |
| $log_{10}$(wiener index) | | $0.02^{***}$ | $-0.14^{***}$ |
| | | (0.00) | (0.00) |
| RMSE | 0.22 | 0.22 | 0.22 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

**(b)** Midterms

**Figure 3.19:** Regression analysis of the relationship between the fraction of toxic tweets, and the Wiener index and size (number of tweets) of the reply trees.

### 3.5.4 Follow Graph Structure

Next, we analyze the relationship between the structure of the follow graph (Figure 3.1d) among the conversation participants and the overall toxicity of the conversations. Similar to the previous section, we define the overall toxicity of the conversation as the fraction of toxic tweets; other definitions such as the mean or the median of the toxicity scores lead to very similar results. It is worth noting that the conversation participants have only a local view of the follow graph, they may recognize their friends or followers, but are unlikely to know how other participants are connected.

**GRAPH SIZE AND DENSITY.** We start by investigating how the size of the graph is related to the overall toxicity. Unsurprisingly, given the results from the previous section, larger follow graphs containing more nodes and more edges tend to be more toxic. However, we find that the density of the connections between the participants also matters (Figure 3.20). The graph's density is defined as the number of edges over the number of pairs of nodes in the graph. We find that, both in the news and the midterms dataset, conversations in which the participants are more densely connected in the follow graph tend to be less toxic. Larger density of connections in the follow graph suggests that the conversation participants are more familiar with each other, which in turn increases the social cost of toxic behavior.

While it is clear that a higher density of connections among the conversation participants correlates negatively with overall toxicity, it is unclear whether the way these connections are distributed in the follow graph impacts toxicity. A follow graph may have high density either because groups of users are very densely connected to each other, or because there are simply many edges uniformly distributed over the graph. In the rest of this section, we use three different graph metrics (number of

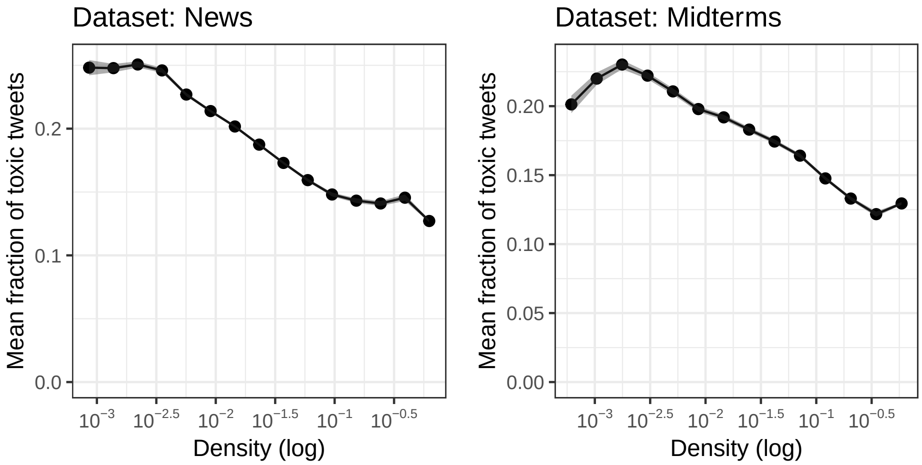


**Figure 3.20:** The relationship between the density of the connections between the users in the Twitter follow graph and the mean fraction of toxic tweets.

connected components, modularity, and overall embeddedness) to answer this question.

**NUMBER OF CONNECTED COMPONENTS.** We start by looking at the relationship between the number of connected components in the follow graph and overall toxicity. A connected component of a graph is a subgraph in which there is a path between any pair of nodes in the subgraph and no path to nodes in the rest of the graph. Here we compute the weakly connected components of the conversation follow graph, i.e., we ignore the direction of the edges when we compute the connected components. The number of connected components has been recently used to quantify the structural diversity of an individual's ego graph and has been shown to explain product adoption decisions made by the ego [91]. In the context of the conversation follow graph, a larger number of connected components suggests that there are many smaller groups of participants who know each other but do not know any other conversation participants. We find that, both in the news and the midterms dataset, the number of connected components is positively correlated with the overall toxicity of the conversation (Figure 3.21).

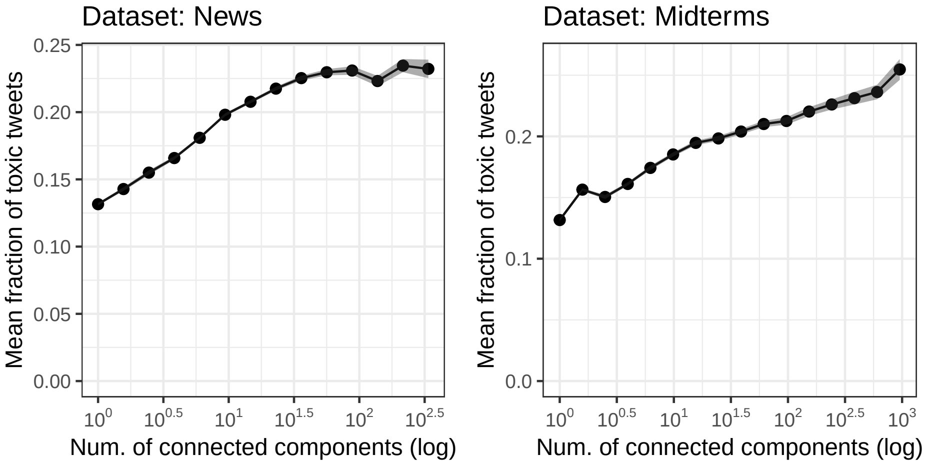


**Figure 3.21:** The relationship between the number of connected components in the follow graph among the conversation participants and the mean fraction of toxic tweets.

**MODULARITY.** Next, we analyze the relationship between the modularity of the conversation follow graph and the conversation toxicity. Given a partitioning of a graph, modularity measures whether there are more or less edges within the partitions than we would expect at random [72, 73]. It is strictly less than 1, it takes positive values if there are more edges within the partitions than we would expect by chance, and negative values if there are less. We first partition the conversation follow graphs using the Louvain algorithm, and then we compute the modularity of the best partitioning. Louvain is a computationally efficient algorithm that uses greedy optimization to detect partitions with maximum modularity [10]. Partitioning the graph with Louvain is a more flexible way of grouping the users than computing the connected components of the graph, allowing for some edges between users of different groups.

We find that conversations in which the follow graph among the participants has higher modularity tend to be more toxic (Figure 3.22). This pattern holds in both datasets, but it is more pronounced in the midterms dataset. We note that a large fraction of conversation follow graphs have a modularity value of zero, 69.1% in the news and 76.35% in the midterms

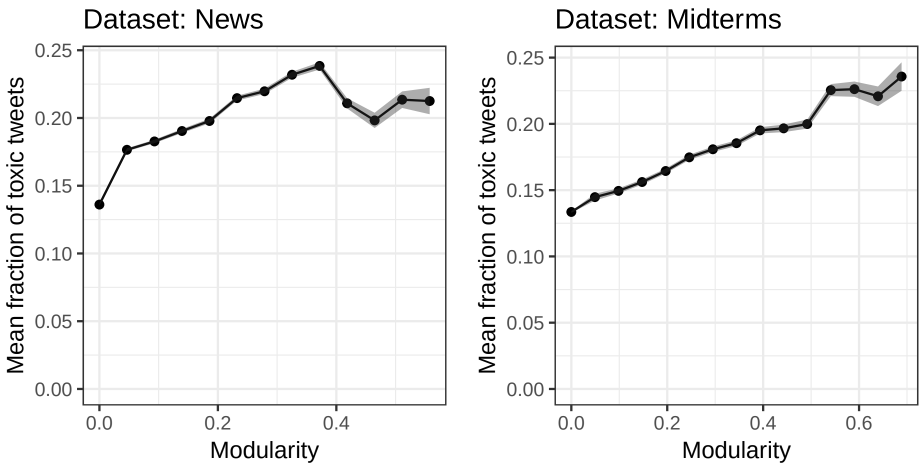


**Figure 3.22:** The relationship between the modularity of the follow graph among the conversation participants after applying the Louvain community detection algorithm and the mean fraction of toxic tweets.

dataset. We suspect that this is due to the fact that many of the follow graphs are sparse.

**EMBEDDEDNESS.** In Section 3.5.2, we analyzed the relationship between embeddedness and toxicity at the dyad level, and we found that replies in which the poster and the replier have many friends in common are less likely to be toxic. Here, we look at the relationship between embeddedness and toxicity at the group level.

Embeddedness allows us to measure the strength of the relationship between the conversation participants, even among those that are not connected to each other in the follow graph. For each conversation, we compute the number of common friends in the follow graph between every pair of conversation participants and calculate the mean and the entropy of the distribution. The mean captures the overall level of embeddedness, and the entropy captures the diversity of the strength of the relationships among the participants.

While we do not observe any relationship between the mean embeddedness and toxicity, we find that conversations with higher embeddedness entropy tend to be more toxic, both in the news and the midterms dataset

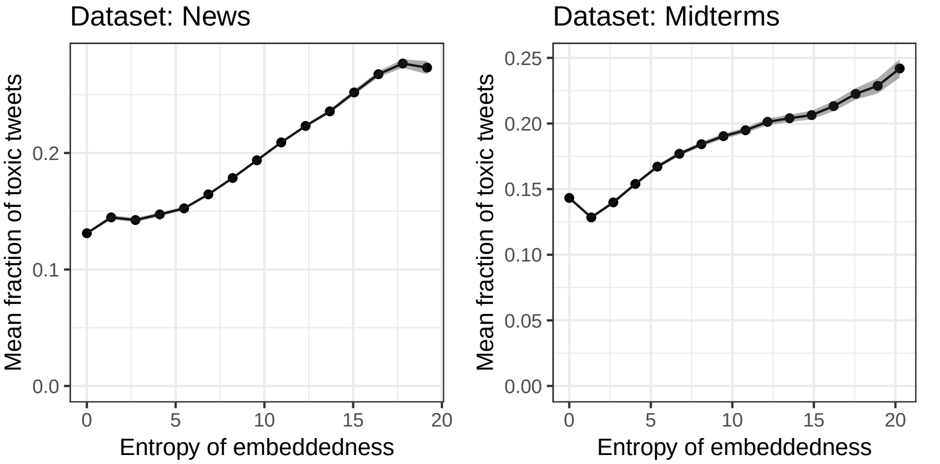


**Figure 3.23:** The relationship between the entropy of the distribution of the embeddedness (i.e., number of common friends) among all the of pairs conversation participants and the mean fraction of toxic tweets.

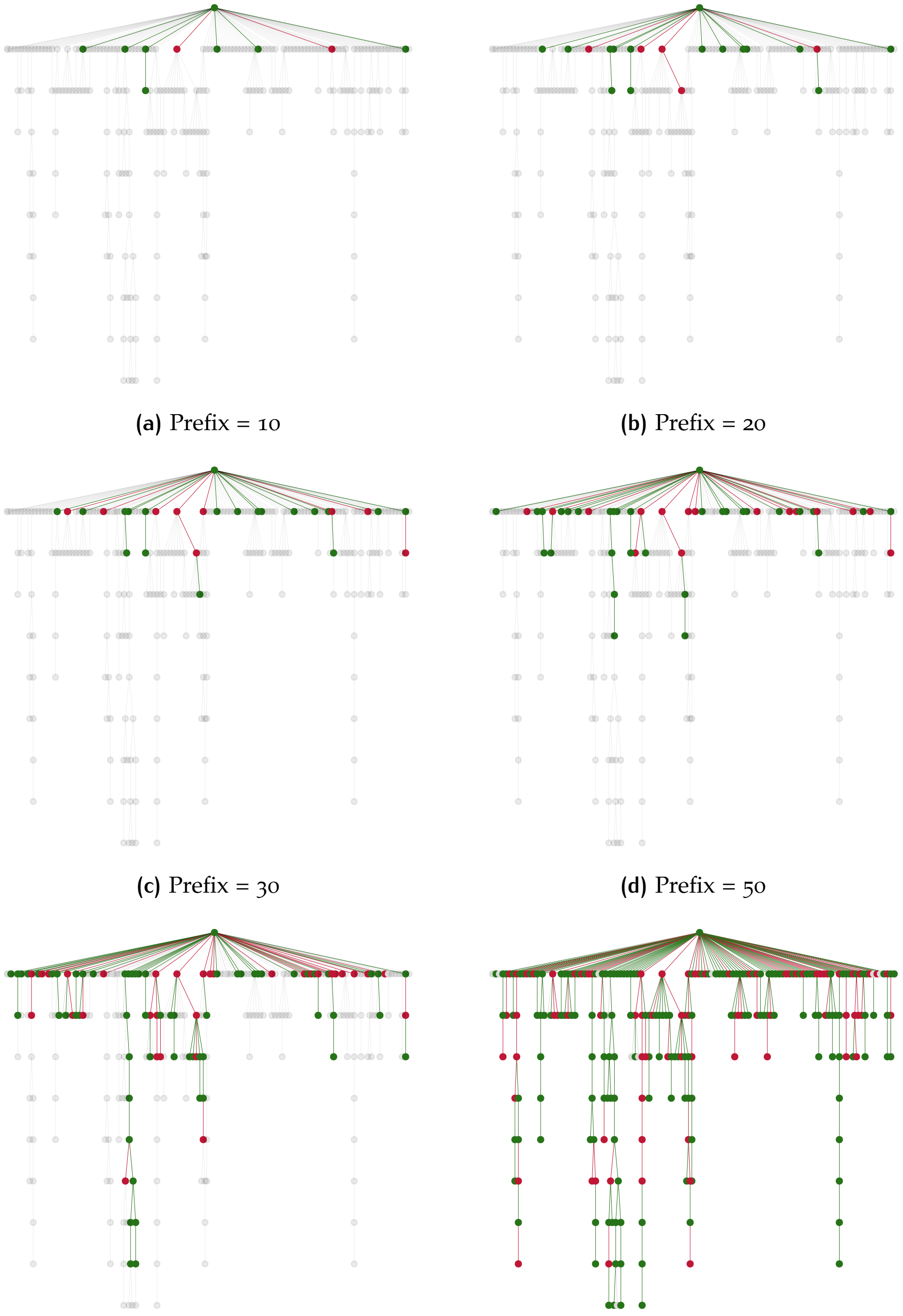
(Figure 3.23). We observe the same relationship between embeddedness entropy and toxicity when we calculate the embeddedness entropy only among pairs of participants that have a follow relationship or have replied to each other. This suggests that even if there are strong ties among the conversation participants, the presence of some weak ties may be enough to lead to higher toxicity.

## 3.6 PREDICTING FUTURE TOXICITY

So far, we analyzed the correlation between toxicity and various structural measures of the conversation, after the conversation has ended. In the next two sections, we consider two prediction tasks that will allow us to measure how useful are the structural properties of the conversation in forecasting how it will unfold in the future. In the first task, we focus on predicting whether the conversation as a whole will become more or less toxic, and in the second task, we focus on predicting the behavior of individual users and whether their next reply will be toxic or not.

In this section, we consider the first task. Given the initial stages of the conversation, e.g., the first ten replies, we are interested in predicting whether the rest of the conversation will turn more or less toxic than expected. To make predictions, we will compute various metrics that characterize the relations among the tweets and the users in the conversation prefix. Figure 3.24 shows a sample conversation and how its reply tree looks at different stages. Our goal is to predict how the shaded regions of the conversation will look, given the highlighted regions.

Beyond allowing us to evaluate which metrics are good indicators of future toxicity, this task also has several important practical applications. First, accurate predictions of future toxicity can be used to decide how much visibility a conversation should be given. For instance, if we suspect that a conversation will turn very toxic, we may decide to downrank the root tweet in users' feeds. These predictions can also be combined with engagement predictions to surface relevant, but nontoxic conversations. Second, early warnings of derailment can be used to prompt the initiator of the conversation to moderate the discussion and prevent it from turning toxic. This is particularly useful for accounts that post frequently, such as news outlets, but do not have the capacity to monitor the conversations. Twitter is currently testing new features that allow users to actively moderate the conversations prompted by their tweets by hiding or prohibiting some replies.



**(a)** Prefix = 10

**(b)** Prefix = 20

**(c)** Prefix = 30

**(d)** Prefix = 50

**(e)** Prefix = 100

**(f)** Complete Reply Tree

**Figure 3.24:** Illustration of the task of predicting future conversation toxicity. The figure shows the state of the reply tree at different stages of the conversation, including the first 10, 20, 30, 50 replies, and the final reply tree. The goal of the task is to predict how the rest of the conversation will unfold (grey nodes/edges), given the conversation prefix (colored nodes/edges).

### 3.6.1 Experimental Setup

**CONTROLLING FOR PREFIX TOXICITY.** A common way to formulate the task for the prediction problem we are interested in is to predict whether the level of toxicity in the conversation suffix will be above or below the median toxicity of all conversations. For instance, this setup has been used to test if it is possible to predict whether a conversation thread will grow [3] or whether an information cascade will grow [16]. However, our scenario is slightly different as the toxicity in the suffix is confounded by the toxicity in the prefix. Even if we fix the size of the prefix, different conversations may contain a different number of toxic tweets in the prefix. Figure 3.25 shows the relationship between the number of toxic tweets in the prefix and the fraction of toxic tweets in the suffix. Unsurprisingly, across all prefix sizes, conversations with more toxicity in the prefix have a higher fraction of toxic tweets in the suffix.

To address this issue, for each prefix size, we first bucket the conversations by the number of toxic tweets in the prefix and then assign the labels depending on whether the fraction of toxic tweets in the suffix is above or below the median of all conversations in the bucket. For example, given the first ten replies of the conversation, four of which are toxic, we aim to predict whether the toxicity in the conversation suffix will be higher than the median toxicity of all conversations that had four toxic tweets within the first ten tweets.

Figure 3.26 shows the distribution of the number of conversations per bucket (i.e., toxic tweets in the prefix) for different prefix sizes. To ensure that there are enough positive and negative examples in each bucket, we only consider buckets with at least 200 conversations. We also exclude conversations smaller than twice the size of the prefix in order to ensure that we have a good estimate of the fraction of toxic tweets in the suffix[7].

[7] We also tested computing the fraction of toxic tweets over the first ten tweets in the suffix; while this increases the number of data points, the results are substantively the same.

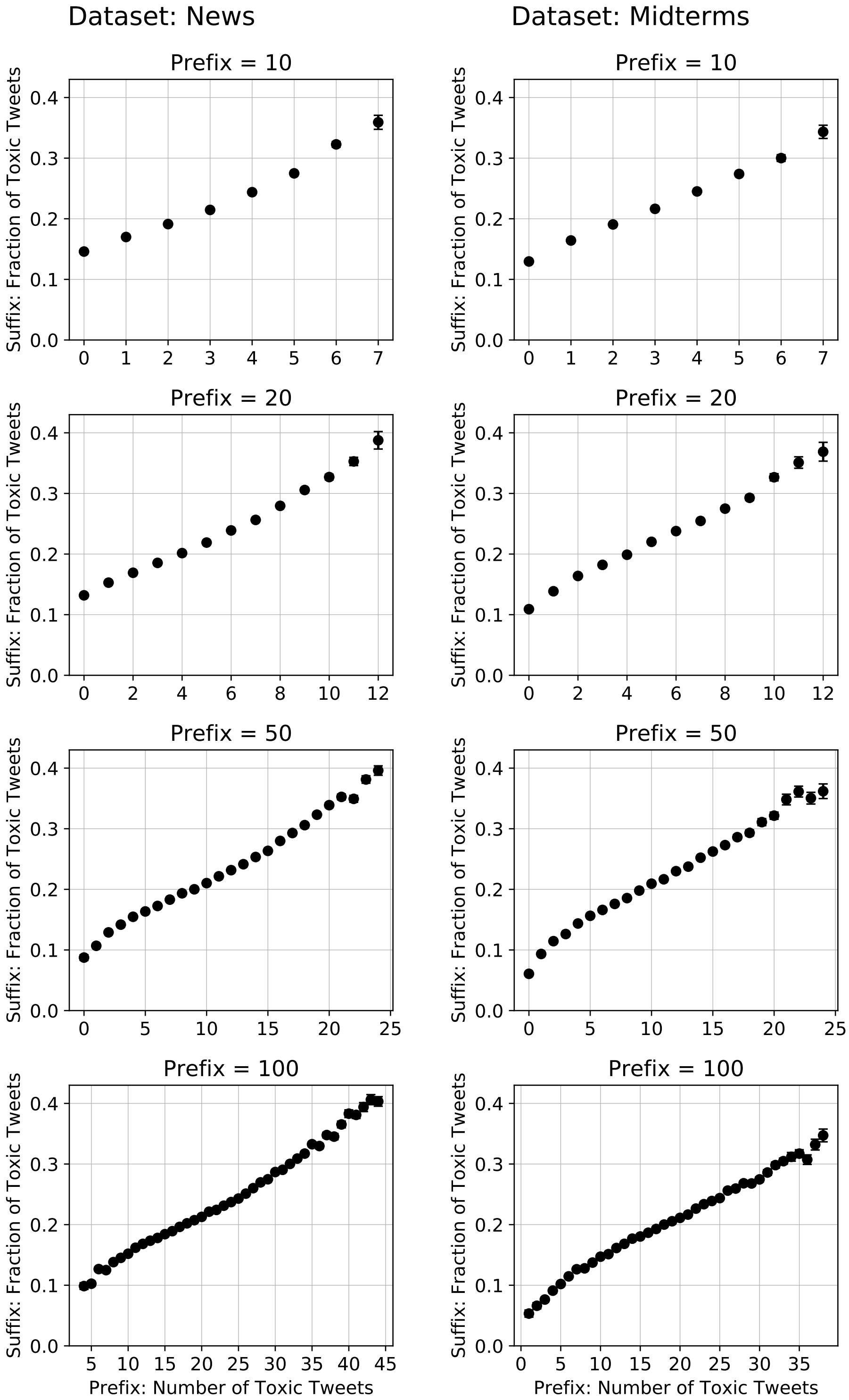


**Figure 3.25:** Relationship between the number of toxic tweets in the prefix and fraction of toxic tweets in the suffix for different prefix sizes. Buckets with less than 50 conversations were excluded.

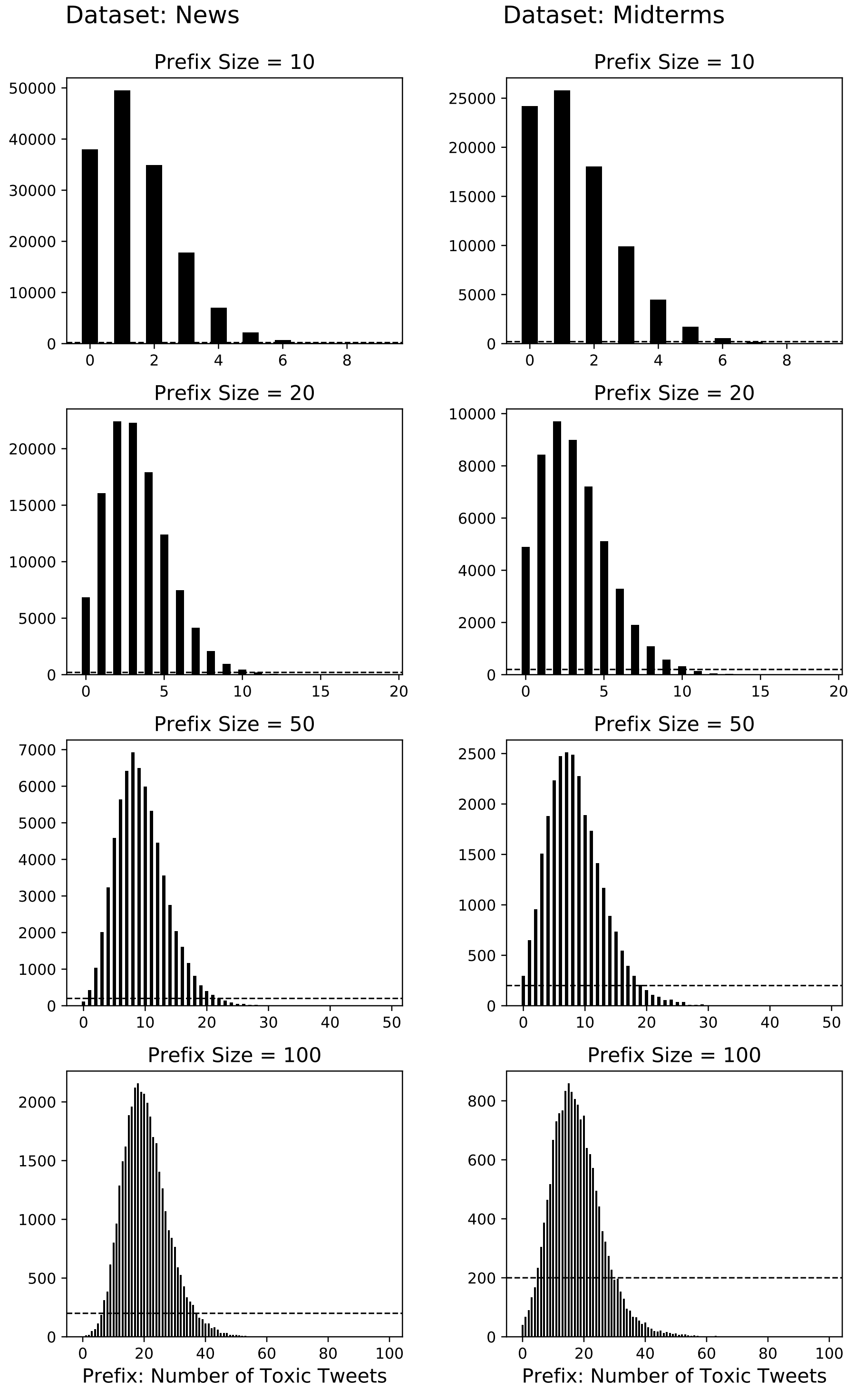


**Figure 3.26:** Distribution of the number of conversations per bucket, i.e., number of toxic tweets in the prefix. The dashed horizontal line shows the threshold of minimum number of conversations per bucket ($n = 200$).

Table 3.4: Number of conversations for each prefix size in the news and the midterms dataset.

| Prefix Size | News | Midterms |
|---|---|---|
| 10 | 148,970 | 84,542 |
| 20 | 112,462 | 51,316 |
| 30 | 91,134 | 38,418 |
| 40 | 76,568 | 31,290 |
| 50 | 65,802 | 26,474 |
| 60 | 57,204 | 22,968 |
| 70 | 50,654 | 20,080 |
| 80 | 44,890 | 17,700 |
| 90 | 40,112 | 16,060 |
| 100 | 35,798 | 14,338 |

This process results in a balanced dataset in which there is no correlation between the labels and the number of toxic tweets in the prefix. Table 3.4 shows the number of data points (i.e., conversations) per prefix size in both the news and the midterms datasets.

**METHODS USED FOR LEARNING.** We tested a variety of linear and non-linear machine learning methods, including Logistic Regression, Linear SVM, Random Forests, and Gradient Boosted Regression Trees (GBRTs). We find that non-linear models perform significantly better (with increases in accuracy ranging between 2% to 5%) and that among them, GBRTs perform best. To simplify the exposition of the results, we only report the performance of the GBRT models.

To evaluate the performance of the models, we used nested cross-validation: in the inner-loop, we perform 5-fold cross-validation to select the best hyper-parameters and refit the model with the best settings, and in the outer-loop, we perform 10-fold cross-validation to measure the performance of the tuned model on unseen data. This procedure leads to unbiased estimates of the expected accuracy of the models after hyper-parameter tuning [14, 92].

We report the mean and 95% confidence intervals of the classification accuracy, the area under the ROC curve (AUC), and the F1 score; all computed across the 10 outer folds.

### 3.6.2 Feature Sets

Next, we describe the features that we use to predict future toxicity. The goal of these features is to characterize the relationships between the tweets and the users in the initial stages of the conversations. To measure the predictive power of the content, we consider summary statistics of the raw toxicity scores of the tweets in the prefix. We also include features that characterize the sequence in which users contribute to the conversation, and the rate the conversation unfolded, both of which have been shown to be predictive of conversation growth [3].

We group the feature into nine feature sets: toxicity, reply tree, follow graph, reply graph, subgraphs, embeddedness, political alignment, arrival sequence, and rate features. In Table 3.5, we show a detailed list of all features.

**TOXICITY FEATURES.** To control for the toxicity in the prefix, we bucketed the conversations by the number of toxic tweets in the prefix. We assigned a binary label, toxic vs. nontoxic, to each of the tweets in the prefix by thresholding their toxicity scores (i.e., $p_{toxic}$). While counting the number of toxic tweets allows us to control for the overall toxicity in the prefix, there still might be some variation in the toxicity scores that we have not accounted for. For example, two conversations may have the same number of toxic tweets in the prefix, but the tweets in one of them may have much higher toxicity scores. This, in turn, may influence the level of toxicity in the suffix.

To test how predictive are the toxicity scores of the tweets in the prefix, we compute a number of features that summarize their distribution (mean, std, min, max, quartiles). We consider these features as a baseline, capturing how predictive of future toxicity is the content of the tweets in the prefix.

**REPLY TREE FEATURES.** We significantly expand the set of features characterizing the structure of the reply trees we considered in Section 3.5.3. In addition to depth, width, and Wiener index, we include features that summarize the distribution of the number of nodes at different depths, the depth of all nodes, the depth of the leaf nodes, and the number of children per node (i.e., branching factor). Since most reply trees have a large number of nodes in the first level of the tree (i.e., direct replies to the root post), we also measure what fraction of nodes are in the first level, how many of them got replies, and how diverse in size are the sub-conversations that they prompt (i.e., the size of the subtrees rooted in them). To capture whether the conversation is dominated by a few users, we also summarize the distribution of the number of tweets per user.

**FOLLOW AND REPLY GRAPH FEATURES.** In addition to the follow graph among the conversation participants, we also consider the reply graph, a user-centric view of the reply-tree in which two users are connected if they have replied to each other. We compute various statistics on both the directed and undirected versions of these graphs. We measure the size and density of the graphs, summarize the degree distributions, and calculate the degree assortativity[8]. We also measure how centralized the graphs are using different centrality measures: betweenness, closeness, eigenvalue centrality, and pagerank. We quantify the level of transitivity in the graph by calculating the local and global clustering coefficients. To measure whether there is a group structure in the graph, we compute the modularity of the best partitions found by Louvain [10], the number of connected components above a certain size, and summary statistics of the $k$-core and $k$-truss of the graph.

[8] Degree assortativity measures whether high-degree users are more or less likely to connect to other high-degree users [71].

Beyond the connections within the conversation, we also compute summary statistics of the number of friends and followers of the conversation participants in the Twitter graph, as well as the level of assortativity, i.e., the tendency for users with a large number of friends/followers to follow or reply to other users with a large number of friends/followers. We also consider the level of assortativity in terms of political alignment, i.e., to what extent users with similar political alignment follow or reply to each other.

**SUBGRAPH FEATURES.** To further characterize the structure of the follow and the reply graphs, we compute the dyadic and the triadic census (i.e., we count the frequency of all possible subgraphs of size two and three) of the follow graph, the reply graph, and the intersection of the two. Previous work has shown that the observed distribution of dyads and triads is useful in classifying whether a conversation is on a controversial topic or not [21]. Here, we test whether it is indicative of future toxicity.

**EMBEDDEDNESS.** To measure the overlap of the social contexts/interests among the conversation participants, we compute a number of embeddedness features. We define three variations of embeddedness between users $i$ and $j$: $a$) number of common friends ($nc_{ij}$), $b$) number of common friends normalized by the smaller friend count among two users ($nc_{ij} / \min(f_i, f_j)$), and $c$) normalized by the total number of unique friends of $i$ and $j$ ($nc_{ij}/(f_i + f_j - nc_{ij}$). We compute summary statistics (mean, variance, entropy, and Gini coefficient) of the distribution of embeddedness among all pairs of users and among pairs of users that have no, one-way, and two-way connections in the reply/follow graph. These features allow us to go beyond the direct connections between the conversation participants and to capture broader, more contextual information about their relationships.

**POLITICAL ALIGNMENT FEATURES.** To measure the overall political alignment of the conversation participants, we compute summary statistics of the distribution of both the numerical alignment scores (mean, std, min, max, quartiles, and interquartile range) and categorical (left vs. right) alignment scores (number of left-leaning users, number of right-leaning users, and entropy).

**ARRIVAL SEQUENCE FEATURES.** The arrival sequence features, proposed in [3], summarize the specific order in which users contribute to the conversation. They consist of two sets of features. The first set is the temporal ids of the user contributing each reply, where the ids are assigned by when the user contributed their first reply. For instance, the sequence of ids: 0, 1, 0, 1, represents a back-and-forth conversation between the first two users. The second set captures the number of unique users at every point of the conversation. Previous work has demonstrated that these features are predictive of the future growth of a conversation [3].

**RATE FEATURES.** These features measure the "speed" at which the initial stage of the conversation unfolded. They capture how much time elapsed between the root tweet and the $i$th reply, how much time elapsed between replies ($i - 1$ and $i$), the average time between replies in overall and between tweets in the first and the second part of the conversation prefix. Rate characteristics have been shown to be indicative of future growth of both conversations [3] and information cascades [16] on Facebook.

Table 3.5: List of features used to build models that, given the initial state of the conversation, predict whether it will become more or less toxic than expected.

| | |
|---|---|
| **Toxicity Features** | |
| $p\_tox_{mean/std/min/max/quartiles}$ | Summary stats of the toxicity scores of the tweets in the conversation prefix |
| **Reply Tree Features** | |
| $depth$ | Depth of the reply tree |
| $width$ | Width of the reply tree |
| $wiener\_index$ | Wiener index, i.e., average distance between all pairs of nodes |
| $depth\_n\_nodes_i$ | Number of nodes at depth $i$ |
| $depth\_n\_nodes_{mean/var/h\text{-}idx/gini/ent}$ | Summary stats (mean, var, h-index, Gini, and entropy) of distribution of number of nodes at every depth |
| $depth\_n\_nodes\_ratio$ | Ratio between the depth and the number of nodes in the tree |
| $nodes\_depths_{mean/var/h\text{-}idx/gini/ent}$ | Summary stats of the depths of all nodes |
| $leaves\_depths_{mean/var/h\text{-}idx/gini/ent}$ | Summary stats of the distribution of leaf node depths |
| $n\_children_{mean/var/hidx}$ | Summary stats of the distribution of number of children |
| $lvl1\_replies_f$ | Fraction of nodes in the first level |
| $lvl1\_replies\_with\_replies_f$ | Fraction of nodes in the first level that received a reply |
| $lvl1\_subtree\_sizes_{gini/entropy}$ | Diversity in the sizes of the subtrees root at nodes in the first level |
| $lvl1\_max\_subtree_{depth\_size\_ratio}$ | Faction between depth and size of the largest subtree rooted in a first level reply |
| $alignment\_corr^{num/cat}$ | Assortativity in political alignment (numerical and categorical) among the users that replied to each other |
| $user\_n\_tweets_{mean/var/h\text{-}idx/gini/ent}$ | Summary stats of the distribution of number of tweets per user |

*(Continued on next page)*

| **Follow / Reply Graph Features** | |
|---|---|
| $n\_nodes$ | Number of nodes |
| $n\_edges^{di/ud}$ | Number of edges in the directed and undirected version of the graph |
| $n\_density^{di/ud}$ | Density of the graph |
| $degrees_{mean/var/f0/gini/h\text{-}idx}$ | Summary stats of the degrees: mean, variance, fraction positive, Gini, h-index |
| $degrees\_corr$ | Degree assortativity |
| $in\_degrees_{mean/var/f0/gini/h\text{-}idx}$ | Summary stats of the in-degrees: mean, variance, fraction positive, Gini, h-index |
| $out\_degrees_{mean/var/f0/gini/h\text{-}idx}$ | Summary stats of the out-degrees: mean, variance, fraction positive, Gini, h-index |
| $out\_in\_degrees\_corr$ | Out and in degree assortativity |
| $dyads_{n/f}^{no/1way/2way}$ | Number / fraction of pairs on nodes with no, 1-way, and 2-way edges |
| $f\_connected\_node\_pairs^{di/ud}$ | Fraction of node pairs connected by a path of any length |
| $cent_x^{di/ud}$ | Centralization in the graph, $x$: betweenness, closeness, eigenvalue, pagerank |
| $algebraic\_connectivity$ | Algebraic Connectivity of the largest Connected Component in the graph |
| $global\_clustering^{di/ud}$ | Global clustering coefficient |
| $local\_clustering^{di/ud}$ | Local clustering coefficient |
| $modularity^{ud}$ | Modularity of the best partitioning found by Louvain |
| $n\_CC1_{f\_nodes}$ | Fraction of nodes in the largest CC |
| $n\_CC > x$ | Number of Connected Components larger than $x = 1, 2, 3, 5, 10$ |
| $k\text{-}core_{n\_nodes/n\_edges/density/n\_CC}$ | Summary stats of the $k$-core of the graph for $k = 1 \dots 5$ |
| $k\text{-}truss_{n\_nodes/n\_edges/density/n\_CC}$ | Summary stats of the $k$-truss of the graph for $k = 1 \dots 5$ |
| $n\_followers_{mean/var/gini/h\text{-}idx}$ | Summary stats of the users' number of followers |
| $n\_friends_{mean/var/gini/h\text{-}idx}$ | Summary stats of the users' number of friends |

*(Continued on next page)*

| | |
|---|---|
| $n\_friends\_n\_followers\_corr$ | Assortativity between number of friends and number of followers |
| $alignment\_corr^{di/ud}_{num/cat}$ | Assortativity of the political alignment among the users, numerical and categorical (L vs. R) |
| $alignment\_corr\_modularity$ | Modularity of the partitions defined by the categorical political alignment |
| **Subgraph Features** | |
| $dyads^{follow/reply/follow\,\cap\,reply}_{n/f}$ | Dyadic and triadic census of the follow, reply, and the intersection of the follow and reply graphs |
| $triads^{follow/reply/follow\,\cap\,reply}_{n/f}$ | |
| **Embeddedness** | |
| $emb^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | Summary stats of the user's number and fraction of common friends with all pairs of conversation participants |
| $emb\_follow\_e0^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | Summary stats (n, mean, variance, entropy, Gini coefficient) of the distribution of the number and the fraction of common friends among all pairs of users that have no ($e0$), one-way ($e1$), or two-way ($e2$) connections in the *follow* graph |
| $emb\_follow\_e1^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | |
| $emb\_follow\_e2^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | |
| $emb\_reply\_e0^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | Summary stats (n, mean, variance, entropy, Gini coefficient) of the distribution of the number and the fraction of common friends among all pairs of users that have no ($e0$), one-way ($e1$), or two-way ($e2$) connections in the *reply* graph |
| $emb\_reply\_e1^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | |
| $emb\_reply\_e2^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | |
| **Political Alignment Features** | |
| $alg_{mean/std/min/max/quartiles/iqr}$ | Summary stats of the (numerical) political alignment of the users |
| $alg_{n\_left/n\_right/entropy}$ | Number of left and right leaning users and entropy of the (categorical) alignment distribution |

*(Continued on next page)*

| | |
|---|---|
| **Arrival Sequence Features** | |
| $user\_id_i$ | The temporal id (assigned sequentially to the every new replier) of the user posting the *i*th reply |
| $unique\_users_i$ | Number of unique users up to the *i*th reply |
| **Rate Features** | |
| $time_i$ | Time elapsed between the root tweet and the *i*th reply |
| $time\_d_i$ | Time elapsed between reply $i-1$ and $i$ |
| $time\_d^{mean}$ | Mean time between replies |
| $time\_d^{mean}_{1...k/2}$ | Mean time between replies in the first half of the conversation prefix |
| $time\_d^{mean}_{k/2...k}$ | Mean time between replies in the second half of the conversation prefix |

### 3.6.3 Results

Next, we evaluate our method's performance in predicting future toxicity given different sizes of the conversation prefix. We report the classification accuracy (ACC), the area under the ROC curve (AUC), and the F1-score, across the 10 cross-validation folds for both the news and the midterms dataset. Since the classification task is balanced, random guessing would results in a performance of 0.5.

We find that, in both datasets, our method achieves the best performance when combining all feature sets, with classification accuracy ranging between 0.61 and 0.64 in the news and between 0.61 and 0.63 in the midterms dataset. While each feature set is individually significantly better than predicting at random, it is the reply graph and embeddedness feature sets that perform best across different prefix sizes in the news dataset and the reply graph feature set in the midterms dataset.

To measure the contribution of the toxicity features to the performance of the full model, we train a classifier with all but the toxicity features

(Figures 3.27 & 3.28, All / Toxicity). Unlike the other feature sets, the toxicity features are based on the content of the tweets in the conversation prefix. We find that, when combined together, the structural features perform significantly better than the toxicity features. However, adding the toxicity features to the structural features (Figures 3.27 & 3.28, All) significantly improves the overall performance of the model. This implies that the structural features capture distinct predictive characteristics of the conversations that are not captured by the content features.

We also note that while there is some variation in the performance of the different structural feature sets, combining them together (All / Toxicity), we obtain a significantly better performance than using any feature set individually. This pattern holds in both datasets. This suggests that each feature set captures a different and complementary aspect of the conversational structure.

Intuitively, we would expect the performance of the classifiers to increase as we observe more of the conversation. However, we find the opposite to be true, especially in the news dataset. We offer two possible explanations for this phenomenon. First, as we increase the prefix size, the number of conversations that are big enough to be considered decreases significantly. For instance, there are 149k conversations in the news prefix 10 dataset and 36k data points in the news prefix 100 dataset. Less training data makes generalization harder and often results in lower prediction performance on unseen data. Second, since we define the classification labels by bucketing the conversations by the number of toxic tweets in the prefix, the prediction problem itself becomes harder and more nuanced as we increase the prefix. There are significantly more prefix toxicity buckets as we increase the size of the prefix (Figure 3.26). In fact, due to the differences in the distribution of conversation sizes between the two datasets, for larger prefix sizes there are fewer prefix toxicity buckets that meet the minimum threshold in the midterms than in the news dataset (Figure 3.26, prefix = 10). This may explain the steeper decrease in performance for larger prefix sizes in the news dataset.

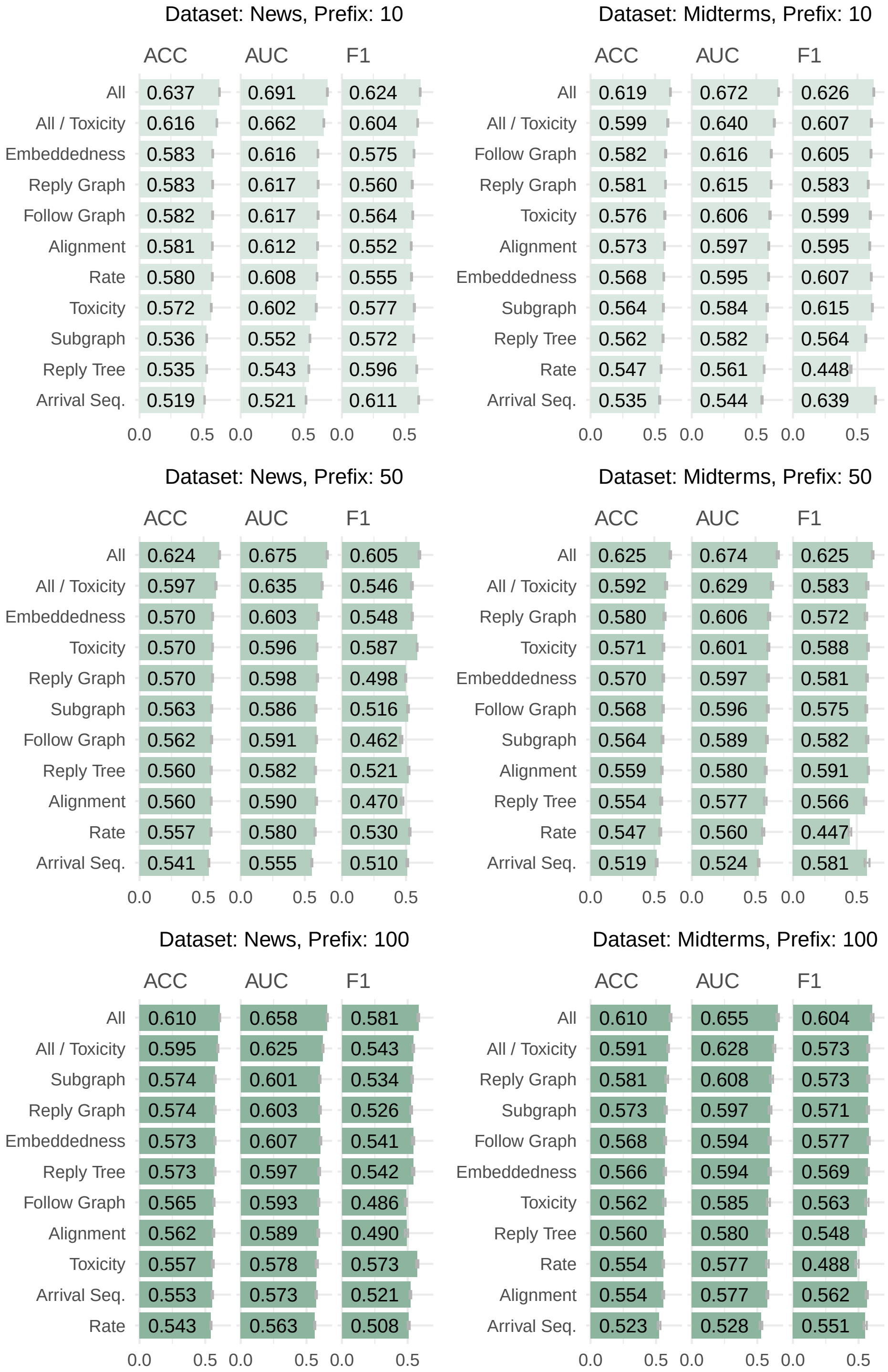


**Figure 3.27:** Classification performance of predicting future toxicity in the conversation given the initial 10, 50, and 100 replies.

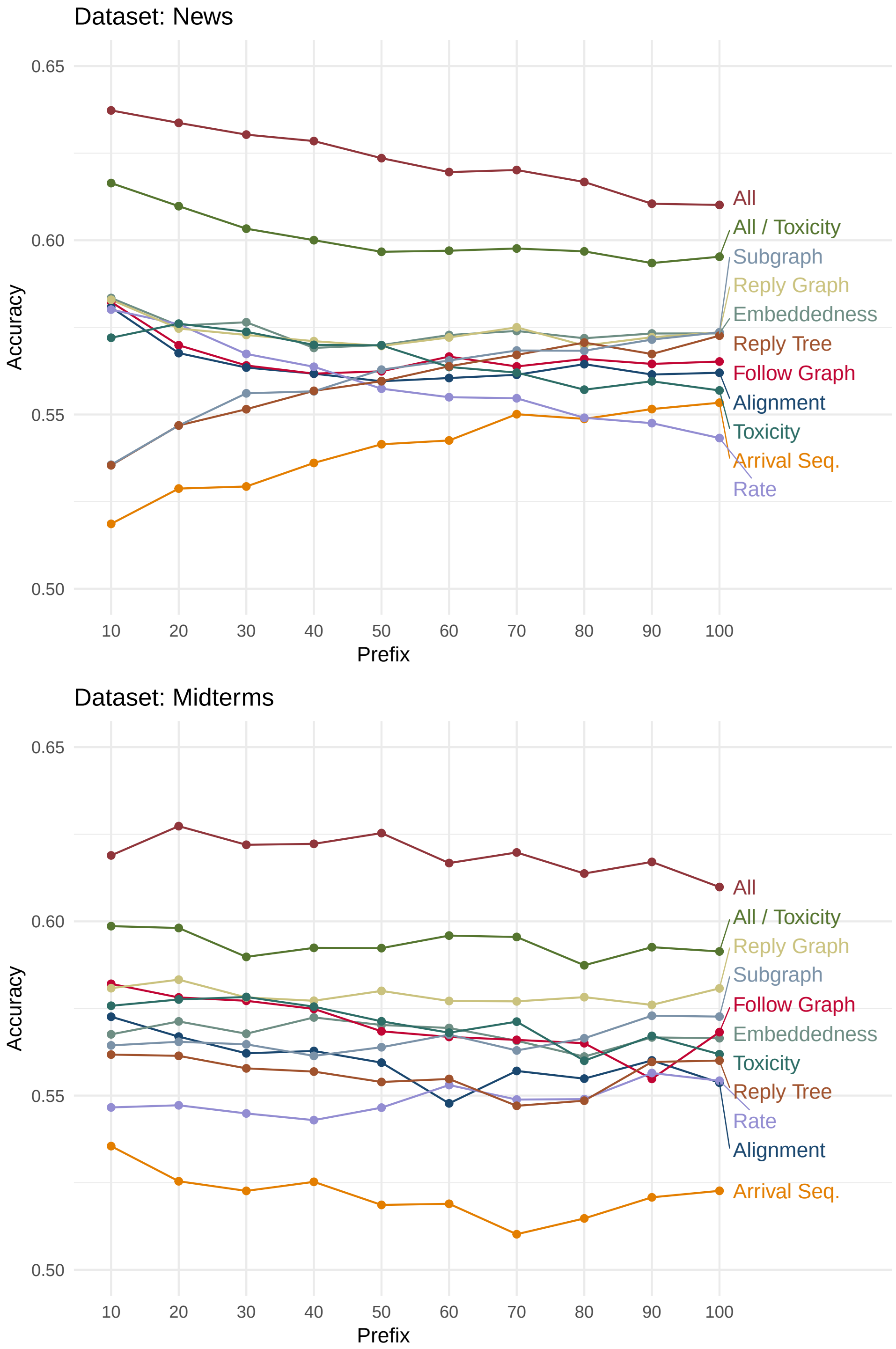


**Figure 3.28:** Classification accuracy of predicting future conversation toxicity given different prefix sizes, *prefix* = {10, 20, 30, 40, 50, 60, 70, 80, 90, 100}.

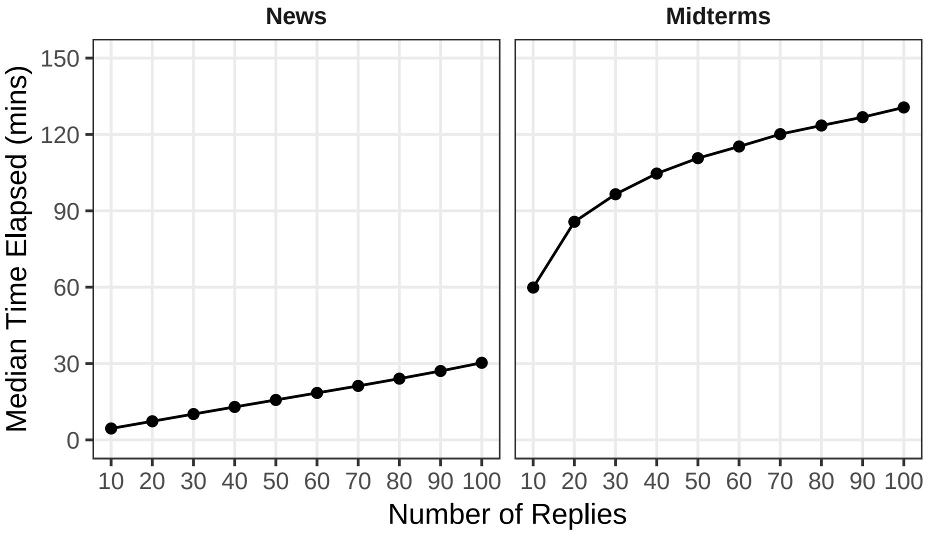


**Figure 3.29:** The median time (in minutes) it takes for a conversation to reach a certain size.

To put the prediction performance in perspective and understand how early we can give a warning that the conversation may derail, we compute the median time it takes for a conversation to reach a certain size (Figure 3.29). The conversations in the news dataset grow much faster than those in the midterms dataset. This is not surprising given the much higher follower counts of the news outlets. In the news dataset, half of the conversations have 10 replies within the first 5 minutes and reach a size of 100 within 30 minutes. In the midterms dataset, half of the conversations reach size of 10 within an hour and size of 100 within 130 minutes. This suggests that we can give a reasonably accurate warning that the conversation may become toxic as early as 5 minutes after the root tweet was posted in the news dataset and within one hour in the midterms dataset.

## 3.7 NEXT REPLY PREDICTIONS

The goal of the first prediction task was to predict how the conversation, as a whole, will unfold in the future by characterizing how the participants are connected to and interact with each other during the initial stages of the conversation. In the second prediction problem, we focus on forecasting the behavior of individual users.

In particular, we aim to predict whether the next reply by a specific user will be toxic, given the conversation so far and the user's relationship to other conversation participants, including the user that they are replying to. This prediction problem is inspired by the practical need to rank the different branches of a conversation to present them to the end-user in a linear order. While Twitter conversations have a tree structure (Figure 3.1b), Twitter's user interface displays the replies in a linear order, which requires one to decide how to order the different branches of the conversation tree. If we can estimate how likely the user is to post a toxic reply to each of the conversation branches, then we can display the branches for which the user is least likely to post a toxic reply first. This will make it less likely for the user to reach parts of the conversation that may prompt them to post a toxic reply.

It is worth noting that unlike the previous prediction problem where we did not know who will contribute to the conversation next, here we assume that we know the identity of the user who will reply next, but we do not know whether their reply will be toxic or not. This setup matches exactly the scenario that we would face in a production system: when a specific user opens a tweet, we need to decide how to rank the reply branches of the conversation such that, if they post a reply, they are more likely to post a nontoxic one. Moreover, this setup creates an opportunity for building personalized models that rank the branches of the conversation based on the identity of the user viewing the conversation.

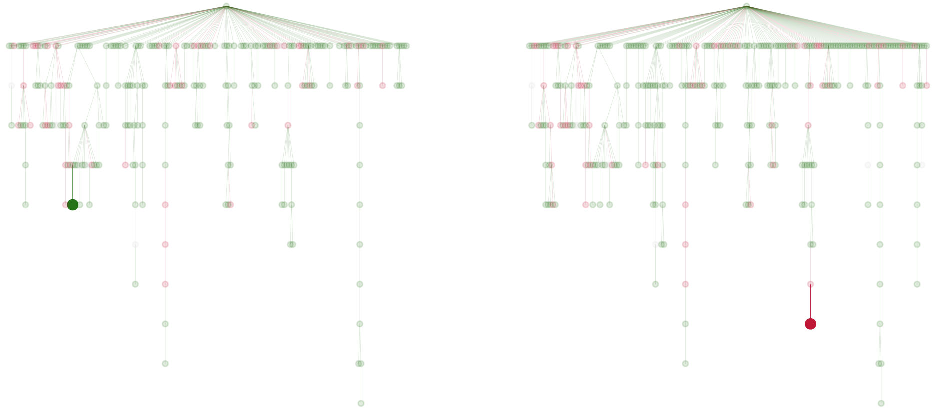
### 3.7.1 Experimental Setup

**CONTROLLING FOR CONTENT.** The content of the root tweet may, to a large extent, drive the structure and the toxicity of the conversation. For instance, tweets by news outlets that cover divisive topics or tweets by midterm candidates sharing their policies on contested issues may be more likely to spur toxic conversations. Moreover, the content discussed across different communities (e.g., audiences of different news outlets) may vary significantly. This, however, does not imply that we should limit the conversations on contested topics altogether. All these considerations motivate the need for an experimental setup that allows us to evaluate the predictive power of the metrics that we propose, but factors out the influence of the content.

To achieve this, we control for the content by using a paired prediction scheme: for each conversation, we sample a pair of a toxic and a nontoxic tweet and aim to predict which one of the two tweets is more likely to be toxic (Figure 3.30). Each pair of tweets is one instance of the prediction task. To represent a pair, we take the difference of the features of the individual tweets and define the label as positive if the first tweet was toxic and negative otherwise. To ensure a balance between the positive and negative class, we construct the pairs such that in exactly half of them, the first tweet is toxic. To avoid overrepresenting any one conversation in the dataset, we sample at most one pair per conversation[9]. While sampling tweets, we exclude self-replies and direct replies to the root as we are interested in identifying indicators of toxicity among the conversation participants. We also exclude tweets whose toxicity scores were close to the threshold and consider only tweets for which the Perspective API prediction ($p_{toxic}$) was below 0.25 or above 0.75.

This paired prediction scheme has been used to control for content in several previous studies [50, 90, 98]. While controlling for the content

[9] Some conversations did not have any pairs of tweets that fit our criteria.



(a) Nontoxic tweet

(b) Toxic tweet

Figure 3.30: Illustration of the next reply paired prediction scheme. Each image shows the state of the conversation right before the sampled tweets were posted, and the highlighted nodes represent the two randomly sampled tweets.

makes the prediction problem more difficult, it allows us to measure the predictive power of the structural representation of the conversations.

**METHODS USED FOR LEARNING.** Similar to the previous prediction task, we experimented with different linear and non-linear models and found that Gradient Boosted Regression Trees (GBRTs) perform best. We follow the same nested cross-validation setup, selecting the best hyper-parameters using 5-fold cross-validation in the inner loop, and measuring the performance on unseen data using 10-fold cross-validation in the outer loop. We tuned only one hyper-parameter, the number of GBRT estimators, choosing one of the following values: $n_{estimators} \in \{10, 25, 50, 100, 500, 1000, 2000, 3000, 5000, 10000\}$. As before, to measure the classification performance, we compute accuracy, area under the ROC curve (AUC), and F1 score.

### 3.7.2 Feature Sets

We proceed by describing the features that we use to predict whether the next reply will be toxic. The goal of these features is to capture the structural relationship between the next tweet and tweets in the conversation so far, and between the user and the current conversation participants. We significantly expand on the individual level and dyad-level properties described in Section 3.5. We group the features into ten feature sets: features capturing the current state of the conversation, properties of the user-parent and user-root dyadic relationship, the tweet's position in the reply tree, features describing the user's position in the follow and reply graphs, the level of embeddedness between the user and other conversation participants, the political alignment of the user relative to other users, and general user features. Below we describe the most important features of each group, and in Table 3.6, we present a detailed list of all features.

**CONVERSATION STATE FEATURES.** In the analysis in Section 3.6.1, we saw that the level of toxicity in the initial portion of the conversation is highly correlated with how the rest of the conversation will unfold. Thus, it is reasonable to assume that the initial level of toxicity is also correlated with the probability that the next reply will be toxic. We record the number of toxic tweets in the conversation so far, but also the number of toxic tweets posted by the user or posted in reply to previous tweets by the user.

**DYADIC RELATIONSHIP FEATURES.** In Section 3.5.2, we found that the dyadic relationship between the user and the parent (i.e., the user they are replying to) is highly correlated with the probability of a toxic reply. In addition to the user-parent relationship, here we also describe the user-root relationship. We use a number of features that characterize different aspects of these relationships.

We capture the relationship of the two users in both the follow and reply graphs[10] as well as the difference of their centrality score according to a number of different measures (degree, betweenness, closeness, eigenvalue, and pagerank). We also record the number of interactions between the two

[10] We use both the directed and undirected versions of these graphs.

users in the conversation so far, and how many of them were toxic. Lastly, we measure whether their political alignments are similar.

**REPLY TREE FEATURES.** In Section 3.5.3, we found that the size, depth, and width of the reply tree are correlated with toxicity. Here, we slightly adapt these features to characterize the position of the tweet in the reply tree. We record the depth, number of siblings (i.e., number of other replies to the same tweet), and the size of the subtree that the tweet belongs to.

**FOLLOW AND REPLY GRAPH FEATURES.** In Section 3.5.4, we found a strong correlation between the structure of the participants' follow graph and the overall toxicity of the conversation. Here, in addition to the follow graph, we also consider the reply graph, which is a user-centric view of the reply tree. We characterize the user's position in the two graphs by measuring the user's centrality (degree, betweenness, closeness, eigenvalue, and pagerank) and the size of the connected component and Louvain partition that the user belongs to. We also record the number of edges to toxic and nontoxic users. To capture how the strength of the relationship between the user and other conversation participants varies depending on their connection in the graph, we break down the user's embeddedness summary statistics by edge type (below we give more details about the embeddedness metrics).

**OVERALL EMBEDDEDNESS FEATURES.** To capture the strength of the relationship between the user and the other conversation participants, we compute a number of summary statistics of the embeddedness between the user and others. We define three variations of embeddedness between users $i$ and $j$: $a$) number of common friends ($nc_{ij}$), $b$) number of common friends normalized by the smaller friend count among two users ($nc_{ij} / \min(f_i, f_j)$), and $c$) normalized by the total number of unique friends of $i$ and $j$ ($nc_{ij}/(f_i + f_j - nc_{ij})$). We characterize the user's overall embeddedness by computing the mean, variance, entropy, and Gini coefficient of the distribution of embeddedness with all other users.

**TOXIC EMBEDDEDNESS FEATURES.** Similar to the overall embeddedness features, we also measure the strength of the user's relationship with toxic vs. nontoxic users by computing the same summary statistics of the distribution of user's embeddedness with users from each group. Here, we consider a user as toxic if they contributed at least one toxic tweet to the conversation.

**POLITICAL ALIGNMENT FEATURES.** To capture how politically aligned the user is with other conversation participants, we compute the mean difference between the user's numerical alignment score and the alignment score of all other users as well as the fraction of users who have the same categorical alignment (left vs. right, categories obtained by thresholding the numerical alignment).

**USER INFORMATION FEATURES.** We record the number of friends and followers of the user in the follow graph at the time when the conversation occurred.

Table 3.6: List of features used to build models that predict whether the next reply will be toxic.

| | |
|---|---|
| **Conversation State Features** | |
| $replies_{n/n\_tox/f\_tox}$ | Total number of replies, number and fraction of toxic replies in the conversation |
| $from\_replies_{n/n\_tox/f\_tox}$ | Total number of replies, number of toxic and fraction of toxic replies *posted by* the focal user |
| $to\_replies_{n/n\_tox/f\_tox}$ | Total number of replies, number of toxic and fraction of toxic *replies to* the the focal user |
| **User-Parent / User-Root Dyadic Features**<br>($*$: user-parent only) | |
| $tweet\_tox$ | Whether the parent / root tweet is toxic |
| $follow\_edge\_type$ | One of four types: O←O, O→O, O=O, O O |
| $embeddedness$ | Number and fraction of common friends |
| $n\_friends\_d$ | Difference in friend counts |
| $n\_followers\_d$ | Difference in follower counts |
| $follow\_di\_d\_cent_x$ $*$ | Difference in centrality scores in the *directed follow graph*, $x$: in-degree, out-degree, betweenness, closeness, eigenvalue, pagerank |
| $follow\_ud\_d\_cent_x$ $*$ | Difference in centrality scores in the *undirected follow graph*, $x$: degree, betweenness, closeness, eigenvalue, pagerank |
| $follow\_ud\_same_{CC/LP}$ $*$ | Whether the users are in the same Connected Component / Louvain partition |
| $follow\_ud\_d\_size_{CC/LP}$ $*$ | Difference in size of the Connected Components / Louvain partitions of the two users |
| $reply\_di\_d\_cent_x$ $*$ | Difference in centrality scores in the *directed reply graph*, $x$: in-degree, out-degree, betweenness, closeness, eigenvalue, page-rank |
| $reply\_ud\_d\_cent_x$ $*$ | Difference in centrality scores in the *undirected reply graph*, $x$: degree, betweenness, closeness, eigenvalue, pagerank |
| $reply\_ud\_same_{CC/LP}$ $*$ | Whether the users are in the same Connected Component / Louvain partition |
| $reply\_ud\_d\_size_{CC/LP}$ $*$ | Difference in size of the Connected Components / Louvain partitions of the two users |
| $replies_{f\_tox}$ | Fraction of toxic replies between the user and the parent / root |

*(Continued on next page)*

| | |
|---|---|
| $replies^{child\rightarrow parent/root}_{n,n\_tox,f\_tox}$ | Number of replies, number of toxic and fraction of toxic replies from the child to the parent / root |
| $replies^{parent/root\rightarrow child}_{n,n\_tox,f\_tox}$ | Number of replies, number of toxic and fraction of toxic replies from the parent / root to the child |
| $alg\_num\_d$ | Difference in numerical political alignment |
| $alg\_cat\_same$ | Whether their categorical political alignment (left vs. right) is the same |
| **Follow / Reply Graph Features** | |
| $di\_cent_x$ | Centrality score of the user in the *directed* version of the graph, $x$: in-degree, out-degree, betweenness, closeness, eigenvalue, page-rank |
| $di\_edges^{in/out/2way}_{n/n\_tox/f\_tox}$ | Number / fraction of in, out, and two-way edges between the user and other toxic and nontoxic users in the *directed* graph |
| $ud\_cent_x$ | Centrality score of the user in the *undirected* version of the graph, $x$: degree, betweenness, closeness, eigenvalue, page-rank |
| $ud\_edges_{n/n\_tox/f\_tox}$ | Number / fraction of edges between the user and other toxic and nontoxic users in the *undirected* version of the graph |
| $ud\_CC_{n/f}$ | Number and fraction of other nodes in the same Connected Component as the user |
| $ud\_LP_{n/f}$ | Number and fraction of other nodes in the same Louvain partition as the user |
| $emb\_no^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | Summary statistics (n, mean, variance, entropy, Gini coefficient) of the distribution of the number and the fraction of common friends with other users who are not connected to the user (*emb_no*), the user is connected to (*emb_in*), and are connected to the user (*emb_out*) |
| $emb\_in^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | |
| $emb\_out^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | |
| **Reply Tree Features** | |
| $n\_siblings$ | Number of siblings in the reply tree |
| $depth$ | Depth of the tweet |
| $subtree_{size,f\_size}$ | Size of the subtree the reply belongs to and fraction of tweets in the subtree |
| $depth/subtree\_size$ | Ratio between the depth of the tweet and the subtree size |

*(Continued on next page)*

| | |
|---|---|
| **Overall Embeddedness** | |
| $emb^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | Summary statistics of the user's number and fraction of common friends with all other conversation participants |
| **Toxic Embeddedness** | |
| $emb\_tox^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | Summary statistics of the user's number and fraction of common friends with users with *at least one toxic tweet* |
| $emb\_nontox^{n/mean/var/ent/gini}_{n/f\_min/f\_union}$ | Summary statistics of the user's number and fraction of common friends with users with *no toxic tweets* |
| **Political Alignment Features** | |
| $alignment\_delta_{avg}$ | Average difference between the user's (numerical) alignment score and all other users |
| $alignment\_f\_same$ | Fraction of other users with the same (categorical) alignment as the user |
| **User Information Features** | |
| $n\_friends$ | Number of friends in the follow graph |
| $n\_followers$ | Number of followers in the follow graph |
| $friends\_followers\_ratio$ | Ratio between the number of friends and followers |

### 3.7.3 Results

Next, we evaluate our method's performance in predicting the toxicity of the next reply under the paired prediction scheme defined above. We sample 96,520 pairs of tweets from the news and 50,143 pairs of tweets from the midterms dataset, where each pair is sampled from a different conversation. We report the classification accuracy, area under the ROC curve (AUC), and F1 score for each dataset over 10 cross-validation folds. Since there is an equal number of positive and negative examples in each class, random guessing would result in a performance of 0.5.

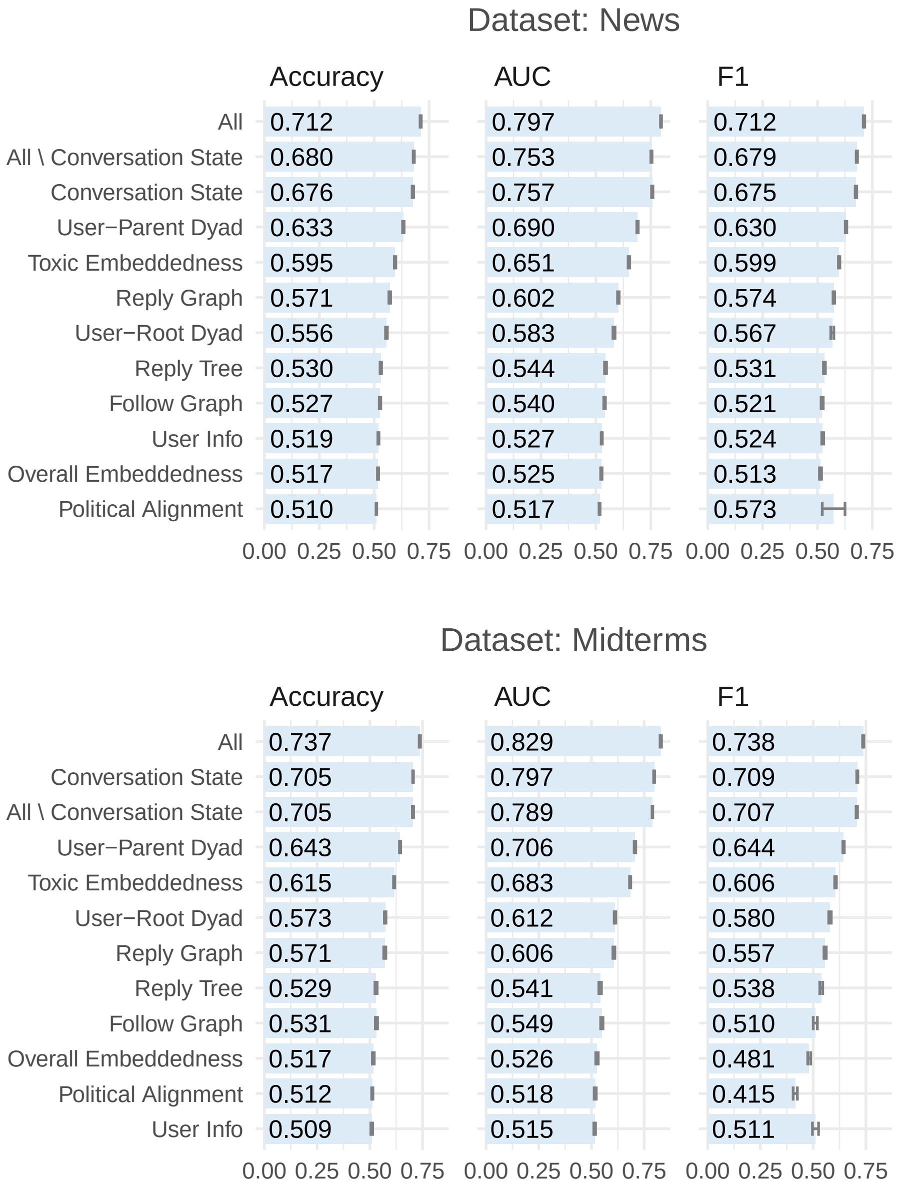


**Figure 3.31:** Classification performance of predicting whether the next reply will be toxic in a paired prediction scheme. The error bars represent 95% confidence intervals computed over the 10 cross-validation folds.

We find that our method achieves surprisingly strong performance in both datasets: accuracy of 0.712 and AUC of 0.797 in the news dataset, and accuracy of 0.737 and AUC of 0.829 in the midterms dataset (Figure 3.31). When we consider the performance of the individual feature sets, we find that features capturing the state of the conversation perform best. To understand whether we could do well without the conversation state features, we trained a classifier which excluded them. We find that while the classification accuracy drops by 0.032 in both datasets, we are still able to obtain reasonable performance even without these features. We achieve similar performance when we use just the conversation state features and just the structural features (i.e., Figure 3.31: All / Conversation State). Moreover, we find that, in both datasets, combining the two (i.e., using all features) significantly improves the classification performance. This suggests that the conversation state features and the structural features capture different and complementary aspects of the conversation that are predictive of whether the next reply will be toxic.

We find that both the absolute and relative performance of the individual features sets is similar in both datasets. This is perhaps due to the fact that our experimental setup was designed to control for the content of the root tweets and suggests that the proposed features could generalize beyond political conversations. We also observe that most feature sets perform significantly better than random, which suggests that our predictions do not rely on any individual feature set and demonstrates that the predictions are robust.

### 3.7.4 Remarks

We designed the prediction problem with the assumption that the user will contribute to the conversation, i.e., we consider the probability of a toxic reply given a reply. In other words, we do not consider counterfactual cases in which the user would have replied but did not. Nevertheless, we think that this assumption is realistic and useful in the context of automatically moderating online conversations that we consider. The usual caveat of prediction models holds here too: the models capture associations rather than causal relationships.

Our models can be used to rank the branches of the conversation such that they are least likely to prompt the user to post a toxic reply. One may worry that this will reduce the number of replies in the conversation, which might be an objective that the platform designers aim for. There is a trivial way to reduce the number of toxic replies by simply not allowing any replies to be posted. In fact, currently, Twitter is testing a feature that allows the user who posted the tweet to allow replies by only users who were @mentioned, which if the tweet does not contain @mentions, it means no replies are allowed. In practice, the models that we propose here may be used in tandem with other models that predict engagement, and together aim to maximize the number of replies but minimize the number of toxic ones.

## 3.8 CONCLUSION

In this chapter, we focused on the structural view of political conversations on Twitter and the relationship to toxicity. We examined 1.18M conversations rooted in tweets that are posted by or mention the Twitter accounts of major news outlets and 2018 midterm election candidates.

To understand the link between structure and toxicity, we analyzed the conversations at three levels: individual, dyad, and group level.

At the individual level, we found that toxicity is not concentrated among a small number of highly toxic users, but it is rather distributed over many low to moderately toxic users (Section 3.5.1). Highly toxic users tend to be more active: beyond posting more toxic replies, toxic users tend to post more replies in general.

At the dyad level, we found that toxic posts are more likely to attract toxic replies than nontoxic posts (Section 3.5.2). Given a toxic post, a toxic reply is more likely to come from a stranger, i.e., a user who does not have any follow relationship with the poster. Given a nontoxic post, a toxic reply is more likely to be posted by a stranger or someone who follows the poster but is not followed back. Users who are more embedded in the social graph, i.e., have more common friends, are less likely to be toxic to each other. This is especially pronounced among users who do not follow each other.

At the group level, we found a strong correlation between the overall structure of the conversation and the overall toxicity of the conversation (Sections 3.5.3 and 3.5.4). Toxic conversations tend to have larger, wider, and deeper reply trees. Conversations with more participants tend to be more toxic; however, when the participants are more densely connected in the follow graph, the conversations tend to be less toxic.

To test the predictive power of the structural characteristics of the conversations, we also considered two prediction tasks. In the first task, we aimed to predict whether the conversation will become more or less toxic than expected, given the initial stages of the conversation (Section 3.6). We found that we can predict how the conversation will unfold with an accuracy of up to 0.62 given only the first ten replies, using only the structural features and after controlling for the toxicity in the initial ten replies. In the second task, we predicted whether the next reply, posted by a specific user, will be toxic (Section 3.7). We found that we can predict the toxicity of the next reply with an accuracy of up to 0.74, even after controlling for the content of the tweets that prompted the conversation.

These findings advance our understanding of the social conditions that are more likely to lead to toxic behavior online. They also have direct practical applications: the models proposed in the two prediction tasks can be readily used to curate conversations algorithmically and to provide early warnings of conversation derailment. The predictions of future toxicity based on the initial conversation tweets can be used to rank or highlight conversations at their early stages (e.g., in users' feeds) or to alert the user who initiated the conversation that it may derail and prompt them to intervene. The predictions of whether the next reply posted by a specific user will be toxic can be used to order the different threads of the conversation such that the user is least likely to post a toxic reply.

While our analysis is based on Twitter data, we suspect that many of our findings would apply to other online, conversational platforms with similar network relationships between users and the ability to engage in semi-public conversations.

# 4 THE RELATIONSHIP BETWEEN POLARIZATION AND TOXICITY

So far, we studied two different phenomena related to tweets posted by political accounts: in Chapter 2 we focused on the political diversity of the users who shared the tweets, and in Chapter 3 we focused on the toxicity of the conversations prompted by them. The next logical step is to investigate if and how these two phenomena—polarization and toxicity—relate to one another. In this chapter, we analyze: (*i*) the relationship between the political diversity of the users who shared a tweet and the overall toxicity of the conversation prompted by the tweet, and (*ii*) the relationship between the political diversity of the conversation participants and the overall toxicity of the conversation.

## 4.1 CONVERSATION TOXICITY AND RETWEETERS' POLITICAL DIVERSITY

We start by analyzing the relationship between the political diversity of the retweeters and the overall toxicity of the conversation prompted by the tweet. We focus on tweets posted by the news outlets between May 2018 and May 2019, which corresponds to the overlapping time period of the data we collected for our polarization (Section 2.3) and toxicity (Section 3.3) analysis. We selected only tweets with at least three retweets and at least two replies so that we could estimate the political diversity of the audience and the overall toxicity of the conversation. In total, we considered 160k tweets. To compute the political diversity of the audience, we use the measure defined in Section 2.5.1, i.e., we compute the entropy

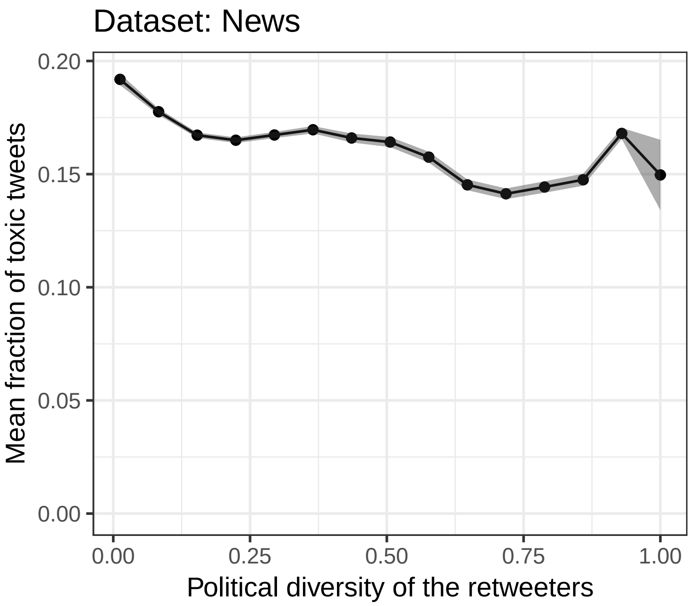


**Figure 4.1:** The relationship between the political diversity of the retweeters and the mean fraction of toxic tweets in the conversation for tweets in the news dataset.

of the distribution of left- vs. right-leaning retweeters, and to measure the overall toxicity of the conversation we calculate the fraction of toxic tweets in the conversation.

We find a weak negative correlation between the political diversity of the retweeters and the overall toxicity of the conversation, i.e., tweets shared by a set of more politically diverse users tend to be slightly less toxic (Figure 4.1). The overall correlation between the two variables is $\rho = -0.08$ ($p < 10^{-15}$). When we compute the correlation for each outlet individually, we find a positive correlation in some cases and negative in others; however, in all cases, the coefficient is close to zero ($\rho$ between -0.13 and 0.07). In summary, we do not find a strong correlation between the political diversity of the users that shared the tweet and the overall toxicity of the conversation prompted by the tweet.

## 4.2 CONVERSATION TOXICITY AND PARTICIPANTS' POLITICAL DIVERSITY

Next, we focus just on the conversations prompted by the tweets and, in particular, the relationship between the political leaning of the conversation participants and the toxicity of the conversation. We consider both the news (510k conversations) and the midterms (676k conversations) datasets. We provide more details about the datasets in Section 3.3.

### 4.2.1 Conversation Toxicity vs. Political Diversity of the Participants

We analyze how the conversation toxicity varies as a function of the overall political diversity of the conversation participants. As before, we measure the political diversity of the conversation participants by computing the entropy of the distribution of left- vs. right-leaning users. We find that conversations in which the participants are more politically diverse tend to be less toxic (Figure 4.2). This pattern is more pronounced in the news dataset where the mean fraction of toxic tweets drops significantly between conversations with no diversity (i.e., entropy = 0) and conversations with perfect audience diversity (i.e., entropy = 1). The pattern is slightly different in the midterms dataset with the mean fraction of toxicity dropping for conversations with political diversity between 0 and 0.6, and slightly increasing between 0.6 and 1.

### 4.2.2 Conversation Toxicity vs. Political Alignment Assortativity

The entropy of the distribution of the conversation participants' political leanings measures the overall diversity of the conversation participants. However, it does not take into account the structure of the conversation. For instance, the conversation participants might be politically diverse, but they may interact only with users who have similar political alignment, e.g.,

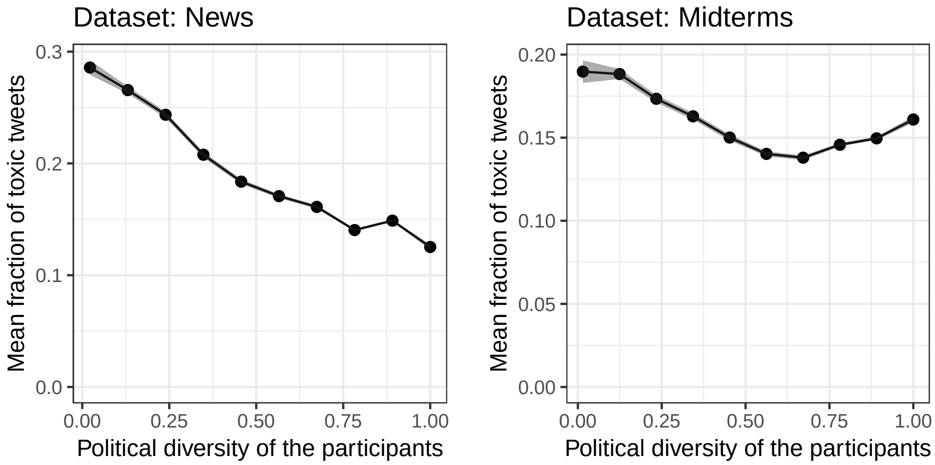


Figure 4.2: The relationship between the political diversity of the conversation participants and the overall toxicity of the conversation.

left-leaning users interacting with each other in one part of the reply tree and right-leaning users in another. To quantify more precisely how users with different political alignments interact with each other, we compute the assortativity coefficient of the users in the reply graph. As a reminder, the reply graph is a user-centric view of the conversation in which the edges represent "replied to" relationships. The assortativity coefficient measures to what extent users interact with other users who have similar political alignment than we would expect by chance. It varies between -1: when users interact only with other users who have different political alignment, and 1: when users interact only with other users who have similar political alignment.

We find that conversations with an assortativity coefficient close to zero tend to be more toxic than conversations with an assortativity coefficient closer to the extremes, i.e., -1 or 1 (Figure 4.3). This pattern holds in both datasets but is more pronounced in the news dataset. We note that in both datasets, a large fraction of conversations have an assortativity coefficient close to zero.

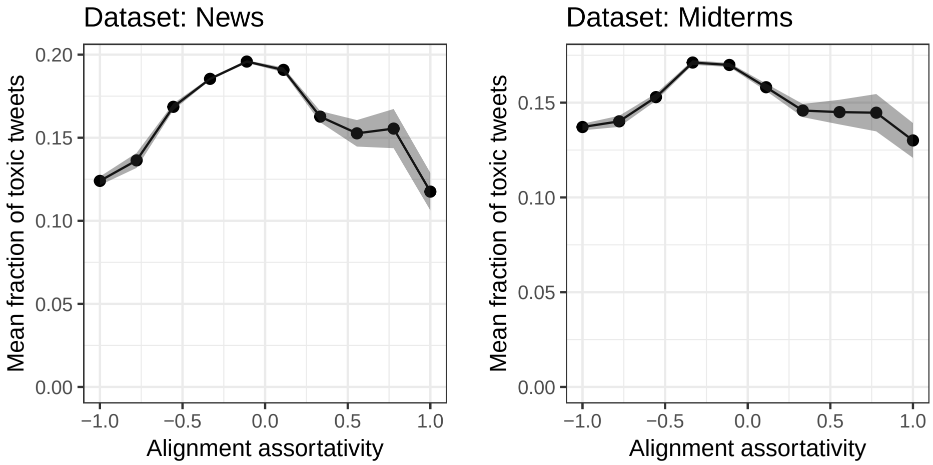


**Figure 4.3:** The relationship between the assortativity of the participants' political alignments in the reply graph and the overall toxicity of the conversation.

### 4.2.3 Toxicity and Political Alignment in Dyadic Interactions

Next, we study the relationship between political alignment and toxicity at the dyad level. We compare the probability of a toxic reply conditioned on whether the poster and the replier have the same or different political leanings and whether the parent post is toxic or not. Similar to Section 3.5.2, we do not consider direct replies to the root tweet and self-replies.

As we found in our previous analysis (Section 3.5.2) toxic posts are more likely to receive toxic replies. Moreover, toxic replies are more likely to come from users who have a different political leaning than the poster (Figure 4.4). Given a toxic post, the probability of a toxic reply by a user with a political leaning different from the poster is 6.8% higher in the news and 7.2% higher in the midterms dataset. Similarly, given a non-toxic post, the probability of a toxic reply from a user with different political leaning is 5% higher in the news and 6.6% higher in the midterms dataset. This finding is somewhat at odds with our previous two results, which suggest that higher political diversity at the group level is associated with lower toxicity. This discrepancy can be explained by the fact that in our

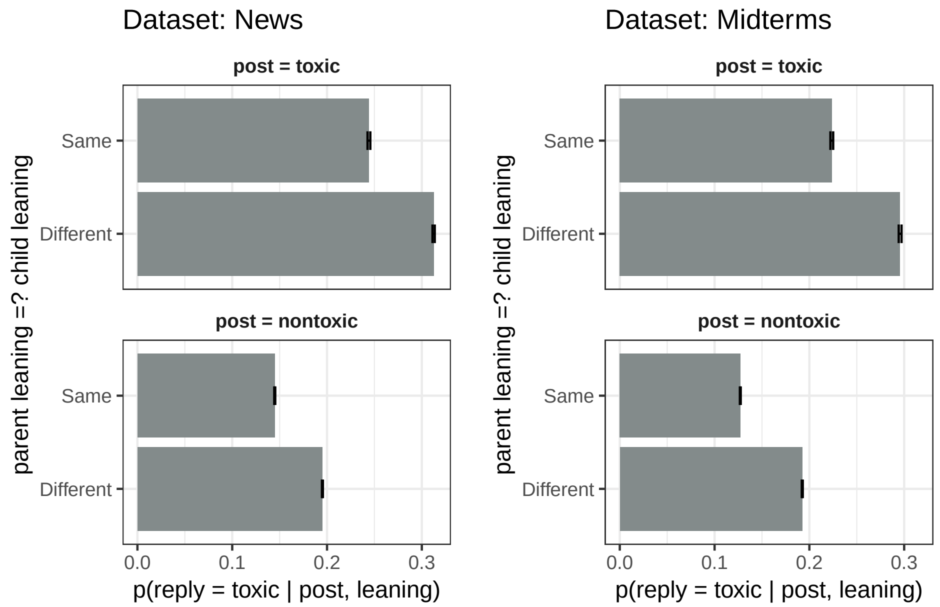


Figure 4.4: The probability of a toxic reply conditioned on whether the poster and the replier have the same or different political leaning and whether the post is toxic or not.

dyadic analysis, we did not consider direct replies to the root, which in most conversations constitute a large fraction of all replies.

### 4.2.4 Predictive Power of the Polarization Features

In Chapter 3, we considered two prediction tasks: (*i*) predicting how the rest of the conversation will unfold given the conversation prefix (Section 3.6), and (*ii*) predicting whether the next reply posted by a specific user will be toxic given their relationship with the participants in the conversation so far (Section 3.7). In both prediction tasks, we included feature sets that capture the political alignment/leaning of the conversation participants.

In the first prediction task, predicting future toxicity, we computed summary statistics of the distribution of both the political alignment scores (mean, std, min, max, quartiles, and interquartile range) and the political leaning (number of left-leaning users, number of right-leaning users, and entropy) of the users in the conversation prefix. The prediction accuracy of these features alone was significantly better than random, ranging between 0.56 and 0.58 in the news dataset, and between 0.55 and 0.57 in the midterms dataset (Section 3.6.3). This suggests that the political alignment of the conversation participants is predictive of how the rest of the conversation will unfold.

In the second prediction task, predicting whether the next reply posted by a specific user will be toxic, we computed the mean difference between the user's political alignment and the alignment scores of all other users, and the fraction of other users with the same political leaning as the user. We found that these features alone perform only slightly better than random with an accuracy of 0.51 in the news and 0.52 in the midterms dataset (Section 3.7.3). However, we note that the dyadic feature set also captures information related to the users' political alignment, such as whether the poster and the replier have the same political leaning, which is predictive of toxicity.

# 5 CONCLUSION AND FUTURE WORK

## 5.1 CONCLUSION

In this thesis, we conducted focused studies of two online phenomena: political polarization and toxic behavior. In both studies, we took a computational approach, applying techniques from natural language processing and social network analysis to model and understand new aspects of these phenomena.

In the first part of this thesis, we focused on the role that media outlets play in political polarization on social media. We studied how the language they use to promote their content influences the political diversity of their audience. We tracked the tweets posted by five major news outlets in the US over three years and measured the political diversity of the users that retweeted them. Using this data, we trained machine learning models that, given the tweet text, predict the political diversity of the audience. We then integrated these models into a web application that helps journalists craft tweets that are engaging to politically diverse audiences by iterating on the tweet text based on the model predictions. To test the effectiveness of our approach in a real-world scenario, we partnered with the PBS documentary series Frontline and ran a series of experiments on Twitter's advertising platform. In each experiment, we used our tool to select one treatment tweet—predicted to be engaging to a more politically diverse audience, and one control tweet—predicted to be engaging to a less politically diverse audience, and ran advertising campaigns to test whether the model predictions will materialize. We found that in five out of the seven advertising experiments, the treatment tweets were indeed engaging to a more politically diverse audience, matching the predictions of our model. These experiments illustrate that we can not only predict the political diversity of the tweets accurately but also use the models to select tweets that are engaging to a more politically diverse audience.

In the second part of this thesis, we studied the relationship between the structure and the toxicity of political conversations on Twitter. We collected data on conversations prompted by tweets posted by five news outlets and candidates who ran for office during the 2018 midterm elections in the US. To analyze the structure of the conversations, we constructed three different views of each conversation: (*i*) reply tree, capturing which tweet was posted in reply to which other tweet, (*ii*) reply graph, encoding which users replied to each other, (*iii*) follow graph, capturing which users have a follow relationship in the Twitter social graph. We analyzed the conversations at the individual, dyad, and group levels. At the individual level, we found that toxicity is not concentrated among a few highly-toxic users, but it is rather dispersed across many low and moderately toxic users. At the dyad level, we found that toxic posts are more likely to receive toxic replies and that toxic replies are more likely to come from users who do not have a social connection with the poster. At the group level, we found that toxic conversations tend to be larger, have wider and deeper reply trees, but less dense follow graphs. To test the utility of the structural features of the conversations in forecasting toxicity, we considered two prediction tasks. In the first task, we predicted whether the conversation, as a whole, will become more or less toxic than expected, given the initial stages of the conversation. In the second task, we predicted whether the next reply, posted by a specific user, will be toxic, given the current state of the conversation and the user's relationship with the other conversation participants. In both tasks, we demonstrated that the structural features are highly predictive of future toxicity. We also showed that combining the structural features with content features leads to even more accurate predictions, suggesting that the two feature sets capture different but complementary aspects of the conversations.

## 5.2 FUTURE WORK

In this section, we outline potential future work that can be undertaken to both broaden and deepen the research presented in this thesis.

**USING MORE GRANULAR CATEGORIES.** In both our analysis of political polarization and conversation toxicity, we relied on coarse categorizations of political alignment and tweet toxicity.

To analyze the political diversity of the audience that engages with tweets posted by media outlets, we classified users in two broad categories: users with left- and users with right-leaning media sharing patterns (Section 2.3.2). We opted for this classification as it led to an intuitive definition of diversity, i.e., the entropy of the distribution of left- vs. right-leaning users. One direction for future work is to explore different definitions of audience diversity by considering more granular categories of users' political alignment. For instance, one can classify users as left-leaning, moderate, and right-leaning; or even more granular categories: far-left, left, moderate, right, far-right. The main challenges of adopting these categories are to determine: (*i*) how to classify users in such fine-grained classes accurately, and (*ii*) how to define audience diversity based on these categories in a way that is intuitive and easy to explain.

To study the relationship between structure and toxicity in conversations, we relied on a broad definition of toxic behavior (Section 3.4). We adopted the definition used by the Perspective API [97], which considers a comment to be toxic if it is "a rude, disrespectful, or unreasonable comment that may make you leave a discussion." Recent work has studied the differences between several, more specific antisocial behaviors, such as offensive, abusive, aggressive, and cyberbullying behavior [30]. One avenue for future research is to explore how other, more specific antisocial behaviors (e.g., abusive behavior) are related to the conversations' structure.

**CHARACTERIZING BRIDGING LANGUAGE.** In Chapter 2, we collected a large number of tweets posted by media outlets and measured the political diversity of the users that retweeted them. Using this data, we built models that, given the tweet text, predict the political diversity of the audience. While these models are useful in making predictions about new tweets, they do not help us understand which characteristics of the text are associated with a higher political diversity of the audience. As we discussed in Section 2.2, previous studies have extensively investigated the link between different text characteristics and popularity, finding that posts that use emotional language [9], are shorter [32], and easier to read [90] tend to be more popular. One avenue for future research is to identify the text characteristics that best differentiate between tweets that are engaging to a politically diverse audience and those that are engaging to only left- or only right-leaning users. One rigorous way of identifying such characteristics is by studying the differences between tweets posted by the same media outlet, sharing the same article (i.e., URL), but with a different text. This would allow us to separate the influence of the tweet text characteristics from the composition of the outlet's audience and the content of the article. Once we have identified the differentiating characteristics, we can use them (*i*) to provide journalists high-level guidance on how to write more bridging tweets, or (*ii*) to supplement the model predictions by indicating whether the tweet drafts have any of the characteristics that are associated with higher audience diversity (similar to the idea of providing explanations based on Concept Activation Vectors [53]).

**RUNNING RANDOMIZED ADVERTISING EXPERIMENTS.** In Chapter 2, we ran a series of advertising experiments on Twitter to test whether we can effectively select tweets engaging to a politically diverse audience using the models and tools that we developed. Running advertising campaigns on Twitter gave us a unique opportunity to conduct realistic experiments and measure how thousands of users respond to the selected tweets

(Section 2.7). However, one limitation of these experiments is that they are not randomized experiments or A/B tests. While we used various features of Twitter's advertising platform to design experiments that resemble A/B tests as closely as possible, we were unable to remove the influence of the advertising engine, which decides which subset of the targeted users will be exposed to the advertisements. Due to algorithmic predictions or market forces, the advertising engine may show the test tweets to users who are more likely to engage with them, instead of a random subset of the targeted users. As a result, we cannot rule out the possibility that the higher audience diversity of the treatment tweets is not due to differences in the tweet content, but due to differences in the delivery of the advertisements.

One of the key future directions for this work is to repeat the experiments using a randomized assignment administered by the advertising engine when such a feature is available on Twitter's advertising platform. Other platforms, e.g., Facebook[1], already offer such capability, and we hope that a similar feature will soon be available on Twitter. Beyond our research question, such capability will create a new opportunity for researchers outside of the company to run realistic and methodologically sound experiments.

**PREDICTING TOXICITY WITH GRAPH NEURAL NETWORKS.** In Chapter 3, we took a more traditional machine learning approach to modeling the relationship between the structure and toxicity of conversations: we computed many features that characterize various aspects of the conversation structure (e.g., properties of the reply tree or the users' follow graph) and applied different learning algorithms. This allowed us to measure and compare the predictive power of the different structural representations of the conversation, e.g., the reply tree vs. the follow graph. In contrast, the main idea behind deep learning methods is to learn meaningful representations of the data, rather than using hand-crafted features. Graph neural

[1] https://www.facebook.com/business/help/1738164643098669

networks have recently spurred much excitement in the machine learning community as they can be effectively used to learn representations of graphs, which cannot be trivially represented in a Euclidean space [39, 96, 100]. They have been successfully applied to a wide range of tasks ranging from drug discovery [86] to fake news detection [68]. One promising future direction is to apply graph neural networks on the conversation reply and follow graphs to predict future toxicity. This would allow us to learn conversation representations that are predictive of toxicity. The main question then is whether these new representations will lead to significantly more accurate predictions of future toxicity to justify the lack of interpretability of the model and its predictions.

## 5.3 CLOSING REMARKS

After the initial euphoria about the potential of the web and social media, we are now slowly starting to realize that these technologies are not a panacea. In fact, many feel that they are responsible for creating new and amplifying existing social problems. However, from a historical perspective, the web and social media are still nascent technologies, and we are just starting to understand how they affect our society. Eventually, the long-term effects of these technologies will depend on our ability to understand their drawbacks and to find ways to improve them. By studying specific aspects of two important online phenomena, polarization and toxicity, this thesis hopes to be a small step in that direction.

# BIBLIOGRAPHY

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. "A simple but tough-to-beat baseline for sentence embeddings." In: *Proceedings of the International Conference on Learning Representations*. 2017.

[2] Yoav Artzi, Patrick Pantel, and Michael Gamon. "Predicting responses to microblog posts." In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. ACL. 2012, pp. 602–606.

[3] Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. "Characterizing and curating conversation threads: expansion, focus, volume, re-entry." In: *Proceedings of the International Conference on Web Search and Data Mining*. ACM. 2013, pp. 13–22.

[4] Paul Baker. "Moral panic and alternative identity construction in Usenet." In: *Journal of Computer-Mediated Communication* 7.1 (2001).

[5] Etyan Bakshy, Solomon Messing, and Lada Adamic. *Replication data for: Exposure to ideologically diverse news and opinion on Facebook*. Version V1. 2019. DOI: 10.7910/DVN/AAI7VA. URL: https://doi.org/10.7910/DVN/AAI7VA.

[6] Eytan Bakshy, Solomon Messing, and Lada A Adamic. "Exposure to ideologically diverse news and opinion on Facebook." In: *Science* 348.6239 (2015), pp. 1130–1132.

[7] Pablo Barberá. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." In: *Political Analysis* 23.1 (2015), pp. 76–91.

[8] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. “Tweeting from left to right: Is online political communication more than an echo chamber?” In: *Psychological Science* 26.10 (2015), pp. 1531–1542.

[9] Jonah Berger and Katherine L Milkman. “What makes online content viral?” In: *Journal of Marketing Research* 49.2 (2012), pp. 192–205.

[10] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. “Fast unfolding of communities in large networks.” In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008.

[11] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. “Enriching word vectors with subword information.” In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[12] Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. “Trolls just want to have fun.” In: *Personality and Individual Differences* 67 (2014), pp. 97–102.

[13] Ceren Budak, Sharad Goel, and Justin M Rao. “Fair and balanced? Quantifying media bias through crowdsourced content analysis.” In: *Public Opinion Quarterly* (2016).

[14] Gavin C Cawley and Nicola LC Talbot. “On over-fitting in model selection and subsequent selection bias in performance evaluation.” In: *Journal of Machine Learning Research* 11.Jul (2010), pp. 2079–2107.

[15] Jianpeng Cheng, Li Dong, and Mirella Lapata. “Long short-term memory-networks for machine reading.” In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2016, pp. 551–561.

[16] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. “Can cascades be predicted?” In: *Proceedings of*

*the International Conference on World wide web*. ACM. 2014, pp. 925–936.

[17] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. “Anyone can become a troll: Causes of trolling behavior in online discussions.” In: *Proceedings of the Conference on Computer Supported Cooperative Work and Social Computing*. ACM. 2017, pp. 1217–1230.

[18] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. “Learning phrase representations using RNN encoder–decoder for statistical machine translation.” In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2014, pp. 1724–1734.

[19] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. “Empirical evaluation of gated recurrent neural networks on sequence modeling.” In: *Conference on Advances in Neural Information Processing Systems, Workshop on Deep Learning*. 2014.

[20] James S Coleman. “Social capital in the creation of human capital.” In: *American Journal of Sociology* 94 (1988), S95–S120.

[21] Mauro Coletto, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. “Automatic controversy detection in social media: A content-independent motif-based approach.” In: *Online Social Networks and Media* 3 (2017), pp. 22–31.

[22] MIT Election Data and Science Lab. *U.S. President 1976–2016*. Version V5. 2017. DOI: 10.7910/DVN/42MVDX. URL: https://doi.org/10.7910/DVN/42MVDX.

[23] MIT Election Data and Science Lab. *U.S. Senate 1976–2018*. Version V4. 2017. DOI: 10.7910/DVN/PEJ5QU. URL: https://doi.org/10.7910/DVN/PEJ5QU.

[24] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. "Analyzing polarization in social media: Method and application to tweets on 21 mass shootings." In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. ACL. 2019.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding." In: *Proceedings of the North American Chapter of the Association for Computational Linguistics*. ACL. 2018.

[26] Julio Cesar Soares Dos Rieis, Fabrício Benevenuto de Souza, Pedro Olmo S Vaz de Melo, Raquel Oliveira Prates, Haewoon Kwak, and Jisun An. "Breaking the news: First impressions matter on online news." In: *Proceedings of the International Conference on Web and Social Media*. AAAI. 2015.

[27] Maeve Duggan. "Online harassment 2017." In: *Pew Research Center* (2017).

[28] Dean Eckles, Brett R Gordon, and Garrett A Johnson. "Field studies of psychologically targeted ads face threats to internal validity." In: *Proceedings of the National Academy of Sciences* 115.23 (2018), E5254–E5255.

[29] Morris P Fiorina, Samuel J Abrams, and Jeremy C Pope. "Culture war." In: *The myth of a polarized America* (2005).

[30] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. "Large scale crowdsourcing and characterization of twitter abusive behavior." In: *Proceedings of the International Conference on Web and Social Media*. AAAI. 2018.

[31] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[32] Kristina Gligorić, Ashton Anderson, and Robert West. "Causal effects of brevity on style and success in social media." In: *Proceedings of the Conference on Computer Supported Cooperative Work and Social Computing*. ACM. 2019, p. 45.

[33] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. "The structural virality of online diffusion." In: *Management Science* (2015).

[34] Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. "Statistical analysis of the social network and discussion threads in slashdot." In: *Proceedings of the International Conference on World Wide Web*. ACM. 2008, pp. 645–654.

[35] Sandra Gonzalez-Bailon, Andreas Kaltenbrunner, and Rafael E Banchs. "The structure of political discussion networks: A model for the analysis of online deliberation." In: *Journal of Information Technology* (2010).

[36] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[37] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. "All you need is "love": Evading hate-speech detection." In: *Proceedings of the ACM Workshop on Artificial Intelligence and Security* (2018).

[38] Marco Guerini, Carlo Strapparava, and Gozde Ozbal. "Exploring text virality in social networks." In: *Proceedings of the 2011 International Conference on Weblogs and Social Media*. AAAI. 2011.

[39] William L Hamilton, Rex Ying, and Jure Leskovec. "Representation learning on graphs: Methods and applications." In: *IEEE Data Engineering Bulletin* (2017).

[40] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. "Searching for safety online: Managing "trolling" in a feminist forum." In: *The Information Society* 18.5 (2002), pp. 371–384.

[41] Jack Hessel and Lillian Lee. “Something’s brewing! Early prediction of controversy-causing posts from discussion features.” In: *Proceedings of the North American Chapter of the Association for Computational Linguistics*. ACL. 2019.

[42] Marc J Hetherington and Jonathan D Weiler. *Authoritarianism and polarization in American politics*. Cambridge University Press, 2009.

[43] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory.” In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[44] Liangjie Hong, Ovidiu Dan, and Brian D Davison. “Predicting popular messages in twitter.” In: *Proceedings of the International Conference Companion on World Wide Web*. ACM. 2011, pp. 57–58.

[45] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. “Affect, not ideologya social identity perspective on polarization.” In: *Public opinion quarterly* 76.3 (2012), pp. 405–431.

[46] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. “The origins and consequences of affective polarization in the United States.” In: *Annual Review of Political Science* 22 (2019), pp. 129–146.

[47] Sarthak Jain and Byron C Wallace. “Attention is not explanation.” In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistic*. ACL. 2019, pp. 3543–3556.

[48] Maximilian Jenders, Gjergji Kasneci, and Felix Naumann. “Analyzing and predicting viral tweets.” In: *Proceedings of the International Conference on World Wide Web*. ACM. 2013, pp. 657–664.

[49] Jeff Johnson, Matthijs Douze, and Hervé Jégou. “Billion-scale similarity search with GPUs.” In: *IEEE Transactions on Big Data* (2019).

[50] Cristian Danescu-Niculescu-Mizil Jonathan P. Chang. “Trouble on the horizon: Forecasting the derailment of online conversations as they develop.” In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2019.

[51] Maurice G Kendall. “A new measure of rank correlation.” In: *Biometrika* 30.1/2 (1938), pp. 81–93.

[52] Maurice G Kendall. “The treatment of ties in ranking problems.” In: *Biometrika* (1945), pp. 239–251.

[53] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors.” In: *Proceedings of the International Conference on Machine Learning*. ACM. 2018, pp. 2668–2677.

[54] D Kingma and J Ba. “Adam: A method for stochastic optimization.” In: *Proceedings of the International Conference on Learning Representations*. 2015.

[55] Klaus Krippendorff. “Computing Krippendorff’s alpha-reliability.” In: *Working Paper* (2011).

[56] Taku Kudo. “Subword regularization: Improving neural network translation models with multiple subword candidates.” In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2018, pp. 66–75.

[57] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. “What’s in a name? Understanding the interplay between titles, content, and communities in social media.” In: *Proceedings of the International Conference on Weblogs and Social Media*. AAAI. 2013.

[58] Geoffrey C Layman, Thomas M Carsey, and Juliana Menasce Horowitz. “Party polarization in American politics: Characteristics, causes, and consequences.” In: *Annual Review of Political Science* 9 (2006), pp. 83–110.

[59] Winston Lin, Donald P Green, and Alexander Coppock. *Standard operating procedures for Don Green’s lab at Columbia*. 2016.

[60] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. "A structured self-attentive sentence embedding." In: *Proceedings of the International Conference on Learning Representations*. 2017.

[61] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. USA: Cambridge University Press, 2008. ISBN: 0521865719.

[62] Travis Martin, Jake M Hofman, Amit Sharma, Ashton Anderson, and Duncan J Watts. "Exploring limits to prediction in complex social systems." In: *Proceedings of the International Conference on World Wide Web*. ACM. 2016, pp. 683–694.

[63] Katerina Eva Matsa and Elisa Shearer. "News use across social media platforms 2018." In: *The Pew Research Center* 10 (2018).

[64] Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. "Psychological targeting as an effective approach to digital mass persuasion." In: *Proceedings of the National Academy of Sciences* 114.48 (2017), pp. 12714–12719.

[65] Nolan McCarty, Keith T Poole, and Howard Rosenthal. "The hunt for party discipline in congress." In: *American Political Science Review* 95.3 (2001), pp. 673–687.

[66] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks." In: *Annual Review of Sociology* 27.1 (2001), pp. 415–444.

[67] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In: *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 2013, pp. 3111–3119.

[68] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. "Fake news detection on social media using geometric deep learning." In: *arXiv preprint arXiv:1902.06673* (2019).

[69] Mohsen Mosleh, Gordon Pennycook, and David G Rand. “Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter.” In: *PLOS ONE* 15.2 (2020).

[70] Vinod Nair and Geoffrey E Hinton. “Rectified linear units improve restricted boltzmann machines.” In: *Proceedings of the International Conference on Machine Learning*. ACM. 2010.

[71] Mark EJ Newman. “Mixing patterns in networks.” In: *Physical Review E* 67.2 (2003), p. 026126.

[72] Mark EJ Newman and Michelle Girvan. “Finding and evaluating community structure in networks.” In: *Physical review E* 69.2 (2004), p. 026113.

[73] Mark Newman. *Networks*. Oxford university press, 2010.

[74] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. “A decomposable attention model for natural language inference.” In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2016, pp. 2249–2255.

[75] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[76] Romain Paulus, Caiming Xiong, and Richard Socher. “A deep reinforced model for abstractive summarization.” In: *Proceedings of the International Conference on Learning Representations*. 2018.

[77] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation.” In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2014, pp. 1532–1543.

[78] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. “RT to win! Predicting message propagation in Twitter.” In: *Proceedings of the International Conference on Weblogs and Social Media*. AAAI. 2011.

[79] Adrian Raine. “Annotation: The role of prefrontal deficits, low autonomic arousal, and early health factors in the development of antisocial and aggressive behavior in children.” In: *Journal of Child Psychology and Psychiatry* 43.4 (2002), pp. 417–434.

[80] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. “A Primer in BERTology: What we know about how BERT works.” In: *ArXiv* abs/2002.12327 (2020).

[81] Abraham Savitzky and Marcel JE Golay. “Smoothing and differentiation of data by simplified least squares procedures.” In: *Analytical Chemistry* 36.8 (1964), pp. 1627–1639.

[82] R. W. Schafer. “What is a Savitzky-Golay filter?” In: *IEEE Signal Processing Magazine* 28.4 (2011), pp. 111–117.

[83] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. “Building end-to-end dialogue systems using generative hierarchical neural network models.” In: *Proceedings of the Conference on Artificial Intelligence*. AAAI. 2016.

[84] Pnina Shachaf and Noriko Hara. “Beyond vandalism: Wikipedia trolls.” In: *Journal of Information Science* 36.3 (2010), pp. 357–370.

[85] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. “A hierarchical recurrent encoder-decoder for generative context-aware query suggestion.” In: *Proceedings of the International Conference on Information and Knowledge Management*. ACM. 2015, pp. 553–562.

[86] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackerman, et al. “A deep learning approach to antibiotic discovery.” In: *Cell* 180.4 (2020), pp. 688–702.

[87] John Suler. “The online disinhibition effect.” In: *Cyberpsychology & Behavior* 7.3 (2004), pp. 321–326.

[88] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." In: *Proceedings of the International Conference on Machine Learning*. ACM. 2017, pp. 3319–3328.

[89] Cass R. Sunstein. *Republic.Com 2.0*. Princeton University Press, 2007.

[90] Chenhao Tan, Lillian Lee, and Bo Pang. "The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 2014, pp. 175–185.

[91] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. "Structural diversity in social contagion." In: *Proceedings of the National Academy of Sciences* 109.16 (2012), pp. 5962–5966.

[92] Sudhir Varma and Richard Simon. "Bias in error estimation when using cross-validation for model selection." In: *BMC Bioinformatics* 7.1 (2006), p. 91.

[93] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In: *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 2017, pp. 5998–6008.

[94] Sarah Wiegreffe and Yuval Pinter. "Attention is not not explanation." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*. 2019, pp. 11–20.

[95] Harry Wiener. "Structural determination of paraffin boiling points." In: *Journal of the American Chemical Society* (1947).

[96] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. "A comprehensive survey on graph neural networks." In: *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[97] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. “Ex machina: Personal attacks seen at scale.” In: *Proceedings of the International Conference on World Wide Web*. ACM. 2017.

[98] Justine Zhang, Cristian Danescu-Niculescu-Mizil, Christina Sauper, and Sean J Taylor. “Characterizing online public discussions through patterns of participant interactions.” In: *Proceedings of the Conference on Computer-Supported Cooperative Work and Social Computing*. ACM, 2018.

[99] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. “Conversations gone awry: Detecting early signs of conversational failure.” In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL. 2018.

[100] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. “Graph neural networks: A review of methods and applications.” In: *arXiv preprint arXiv:1812.08434* (2018).