Linguistic Characterization of Divisive Topics Online: Case Studies on Contentiousness in Abortion, Climate Change, and Gun Control

Jacob Beel,[1] **Tong Xiang,**[2] **Sandeep Soni,**[1] **Diyi Yang**[1]

[1]Georgia Institute of Technology, [2]Georgetown University
[1]School of Interactive Computing, [2]Department of Computer Science
[1]Atlanta, Georgia 30318, [2]Washington, DC 20057
{ jbeel3@gatech.edu, tx39@georgetown.edu, sandeepsoni@gatech.edu, dyang888@gatech.edu}

Abstract

As public discourse continues to move and grow online, conversations about divisive topics on social media platforms have also increased. These divisive topics prompt both contentious and non-contentious conversations. Although what distinguishes these conversations, often framed as what makes these conversations contentious, is known in broad strokes, much less is known about the linguistic signature of these conversations. Prior work has shown that contentious content and structure can be a predictor for this task, however, most of them have been focused on conversation in general, very specific events, or complex structural analysis. Additionally, many models used in prior work have lacked interpretability, a key factor in online moderation. Our work fills these gaps by focusing on conversations from highly divisive topics (abortion, climate change, and gun control), operationalizing a set of novel linguistic and conversational characteristics and user factors, and incorporating them to build interpretable models. We demonstrate that such characteristics can largely improve the performance of prediction on this task, and also enable nuanced interpretability. Our case studies on these three contentious topics suggest that certain generic linguistic characteristics are highly correlated with contentiousness in conversations while others demonstrate significant contextual influences on specific divisive topics.

Introduction

Divisive topics – from gun control to illegal immigration, from animal cruelty to vaccination in children – have gained another avenue for exposure in online media. These topics evoke opinionated and polarizing conversations and though we broadly understand the different issues that make these topics contentious, we know very little about the structure and dynamics at the level of conversations around these topics. For instance, what makes a conversation about a particular topic contentious? What are the differences between conversations which foster healthy disagreement and discussion about a divisive topic to those which devolve into online shouting matches? What is the extent of this effect in conversations about a particular topic on a particular platform?

Answers to the above questions and a nuanced understanding of these conversations could help in resource allocation on social media platforms: for example, determining the course of a conversation early on in its lifespan could help to guide moderator action and perform a kind of triage for online community moderation. Additionally, the distribution of these conversations on the platform can help in user enculturation: for instance, platforms can guide newcomers so that they are exposed to less contentious conversations and get a gentle introduction to the communities dedicated to discussing divisive topics. Alternatively, measuring the extent of contentious conversations can be used to assess platform health (Matamoros-Fernández 2017): for example, conversations that manifest as constant inter-group conflict or lead to hateful speech or vile ad hominem attacks can be taken as a bad sign, marking platforms as amplifiers and manufacturers of negative discourse; in contrast, discussions that balance between opposing viewpoints or improve information diversity can be viewed as good signals.

Much attention has recently been paid towards predicting conversation quality, typically framed as controversy, from the content of and structure of conversations (Hessel and Lee 2019; Kim 2014). The strength of these works lies in the use of powerful neural networks, such as convolutional neural networks (CNN) and transformers (BERT), to predict controversy in conversations with high accuracy. However, these predictive models are constrained by their relative limitations in highlighting the informative conversation characteristics that distinguish controversial conversations from non-controversial ones. Moreover, their predictive performance has only been tested in generic conversations, making it hard to tease out conversation topics being divisive from participation patterns in discussions leading to controversy. This work fills this gap by focusing exclusively on divisive topics. In this work, we aim to address the above gaps by answering the following research questions: **RQ1** Can we accurately predict contentiousness in conversations via explainable conversation characteristics when controlling for divisive topics? **RQ2** What specific conversation characteristics differentiate non-contentious conversations from contentious ones between certain divisive topics?

To address RQ1 we operationalize a set of linguistically

explainable characteristics including coarse discourse acts, toxicity, sentiment, and occurrence of specific lexicons. In addition, we propose to leverage conversation participants' characteristics such as their gender, location, account age, and prolificity. Many of these features, including discourse acts and user features have not been used in this task previously. We follow prior work (Choi, Jung, and Myaeng 2010; Hessel and Lee 2019) to define quality of conversations. We focus on conversations identified to belong of one of the three divisive topics: abortion, climate change, and gun control and then look to examine the contentiousness of conversations within those divisive topics. These conversation characteristics are then utilized to build machine learning classifiers to predict whether a conversation is contentious. To address RQ2, we visualize the most predictive conversation characteristics, across and within each divisive topic. Our analyses show that certain generic linguistic characteristics are highly correlated with contentiousness in conversations across the board: for instance, negative discourse acts appear significantly more often in contentious conversations across three topics; while others demonstrate significant contextual influences such as the varying importance of user demographic information.

To sum up, the contributions of our work are three-fold. First, we operationalize a set of conversation characteristics such as coarse discourse acts and toxicity, which can improve both the predictive performance and the explainability of contentious conversation detection. Second, we demonstrate that the contentiousness can be detected using our explainable models with reasonable accuracy, even at an early stage of these conversations. Third, our research uncovers both generic and topic specific conversation characteristics associated with discussions about divisive topics online. This work can be leveraged for several different purposes, including further research and aiding moderation strategies of online communities.

Note that this work uses terminology which may be different from prior work due to the structure of this research. For instance, Hessel and Lee refer to conversations as "controversial". In this work, we elect to refer to conversations as "contentious" and to topics as "divisive" in order to ensure that the difference between the nature of the two is as clear as possible. The topics we study are divisive, but will not necessarily result in contentious conversations.

Related Work

Controversy in Online Discussions

There has been extensive computational research about controversy in online discussions. To quantify controversy, Choi, Jung, and Myaeng (2010) used the ratio of positive and negative sentiment words; Rethmeier, Hübner, and Hennig (2018) and Hessel and Lee (2019) define controversial content as that which attracts a mix of positive and negative feedback. Other lines of research used predefined lexicons to collect controversial discussions (Awadallah, Ramanath, and Weikum 2012). Prior work on predicting controversy or distinguishing controversial discussions from non-controversial ones rely heavily on dictionary based approaches (Pennacchiotti and Popescu 2010; Addawood et al. 2017). Such methods benefit from being explainable but have limited predictive power (Dori-Hacohen and Allan 2013). Supervised neural models such as CNNs have also been used recently for predicting controversy (Rethmeier, Hübner, and Hennig 2018). Since it is difficult to get ground-truth annotations of controversy, these approaches typically use distant supervision, such as upvote-downvote ratio, to mark examples as either controversial or non-controversial. The growing popularity of Reddit has attracted research attention to predict and model controversy on its platform (Conover et al. 2011). Hessel and Lee (2019) uses BERT to predict controversial conversations on Reddit. Guimaraes et al. (2019) refines controversy into disputes, disruptions, and discrepancies via user actions and connects those user actions to sentiment and topic analysis to explain controversy specifically within political discussion.

Contentiousness, Divisive Topics and Controversy

Different from all the aforementioned research, we argue that controversy should be studied at least from two perspectives: topics being divisive, and conversations being contentious. The key intuition is that, people can still have deliberative and good discussions regarding a very divisive topic such as gun control. Thus, topics being divisive does not equal to conversations being contentious, and analyzing how conversation characteristics relate to contentious content should control for divisive topics. For the purposes of our work, controversial and non-controversial conversations are referred to as contentious and non-contentious, respectively. This is more accurate to our task at hand, which is to distinguish between two types of discourse which emerge when even discussing divisive topics rather than detect controversy irrespective of topic. Therefore, our work controls for topics in order to explore why certain conversations become controversial and why others do not within and between those topics. We do this by opting to focus on linguistic characterizations of conversations as well as user factors rather than conversation structure and comment timing. In contrast to both Rethmeier, Hübner, and Hennig and Hessel and Lee, we focus on using linguistic characteristics and user factors to build simpler models, rather than end-to-end neural solutions. In doing so we create a more explainable solution, and compare these models against the transformer architecture BERT (Devlin et al. 2018).

Abortion, Climate Change, and Gun Control

To better study conversation characterises and their influences on conversations being contentious, we first need to control for divisive topics, one big confounding factor behind controversy. As a case study, we focus on discussion around three topics – abortion, climate change, and gun control. These contentious topics have long been in the public discourse (Schweizer, Davis, and Thompson 2013) but have gained an additional dimension: views on these topics are now recorded permanently on social media platforms making these platforms attractive venues for research on the social aspects of conversations (Tremayne and Minooie 2013; Wasike 2017; Chor et al. 2019). Online discussions about

these contentious topics tends to be polarizing (Garimella et al. 2017b). For instance, Garimella et al. (2017a) found that debates on contentious topics, such as gun control or abortion, on Twitter become more polarizing as a result of increased public attention that typically follows some external event. Similarly, Demszky et al. (2019) measured polarization based on the content of tweets around issues of gun policies for a series of documented mass shooting events. However, in contrast to these objectives of measuring polarization around these topics, our focus is on characterizing what linguistic features are correlated with controversy. The topic of climate change, though relatively recent compared to gun control and abortion, has also attracted research particularly on how it is discussed online (Olteanu et al. 2015; Anderson and Huntington 2017). Jang and Allan (2018) proposed a method to summarize discussions on contentious topics such as abortion and climate change, to better understand the controversy and these different viewpoints. Our focus in this work is also on explaining controversy in these topics but we diverge in two ways: first, we focus on understanding the linguistic aspects that are predictive of controversy in conversations; second, we use high-capacity neural networks to classify conversations as controversial or non-controversial, allowing us to distinguish between the two for any of the contentious topics that we considered.

Data

Our data consist of filtered and labeled posts and comments from Reddit. Reddit is a content-sharing and discussion social media website where users can post conversations, which other users can comment on. The data consist of most conversations (also known as "posts" or "threads") and comments over the period of time from 2007 to February 2014 (Hessel and Lee 2019), and are drawn from the same source as is used by Hessel and Lee. Deleted conversations are not included, but conversations and comments which were made by accounts that have since been deleted are included. A conversation consists of a title, selftext (the body of an original post), and comments. The title and selftext are input by the original poster (OP) at the time of conversation creation. We label the posts according to the scheme proposed and validated in Hessel and Lee (2019). Under this scheme, a conversation is considered contentious if its upvote ratio is within the bottom quartile of conversations, i.e. it has drawn more mixed user reactions. Only conversations which have an upvote ratio of at least 0.5 are considered.

Although we use the same labelling technique as Hessel and Lee, we control for divisive topics. This allows us explore the manner in which contentious conversations differ within those divisive topics, a question which Hessel and Lee elect not to explore. Additionally, we explore a larger number of subreddits by virtue of our choice to filter by topic. As a robustness check, we also compared this labelling technique to another which is widely used in this task (Choi, Jung, and Myaeng 2010), the ratio of positive to negative sentiment. Across all topics, this resulted in unbalanced data sets, with non-contentious conversations making up over 85% of conversations. We found this definition of controversy to differ significantly, and state of the

art models can not perform significantly better than random chance. These two labelling techniques clearly measure different phenomena, and we urge future work to explore these differences for deeper insights.

To retain only the conversations associated with particularly divisive topics, we began with a larger list of divisive topics available at ProCon.org (http://procon.org) and filtered down to a subset of topics that met our selection criteria. First we aimed to maximize the number of conversations. Second, we aimed to minimize the number of off-topic conversations through a process of filtering, followed by manual inspection, followed by refinement of key words. This process was repeated until the number of off-topic conversations was minimized. Finally, we aimed for topics with a large difference between the upvote ratios of the two label categories. Topics which did not satisfy this criteria, such as the "*U.S. involvement in the Middle East*", "*free or affordable higher education*", "*healthcare*", and "*police brutality*", were excluded. The topics we focus on and the number of conversations n in each topic are: abortion (n=892), climate change (n=1070), and gun control (n=1176). These were selected due to their large differences aside from being political and divisive.

These discussions (conversations and comments) were identified using keywords selected from the ProCon.org pages for these topics: abortion (e.g., *"pro-life", "pro-choice", "fetus", "abortion"*), climate change (e.g., *"global warming", "climate change", "greenhouse gas", "fossil fuel"*), and gun control (e.g., *"ar-15", "national rifle association", "assault rifle", "second amendment"*). The three selected topics had the highest number of conversations of the topics we considered and a wide enough range of upvote ratio to indicate the presence of different classes of conversations. These topics also minimized the number of false positives, conversations that matched one or more keywords but did not actually pertain to the topic. Additionally, these topics presented ways in which they may differ. For instance, we hypothesized that abortion would be a topic where a non-minority presence of male users in a conversation would lead to contentiousness, whereas a non-minority presence of female users in a conversation would not.

Conversation Characteristics

We operationalize a set of conversation characteristics that include both linguistic and non-linguistic features needed by subsequent models and explainable analyses.

Linguistic Operationalization

Discourse Acts and Discussion Patterns contentiousness in a discussion is reflected by the expression of opposing views or disagreements. To better understand those discourse patterns in contentious conversations, we leverage coarse discourse acts (Zhang, Culbertson, and Paritosh 2017) to extract different types of patterns in users' discussions. These discourse acts consist of the following 10 acts: *question*, *answer*, *announcement*, *agree*, *appreciate*, *disagree*, *negative*, *elaborate*, *humor*, and *other*. We used BERT and fine-tuned it to predict these discourse acts for

each post and each comment using the Reddit annotated data from Zhang, Culbertson, and Paritosh (2017), with 65% 10-class macro accuracy at predicting these discourse acts. The entire conversation conversation can then be represented in terms of these discourse acts as unigram and bigram vectors.

Toxicity Previous work has shown that toxicity is linked with contentious conversations on social media (Zhang et al. 2018; Mejova et al. 2014). Prior work has leveraged it as a signal for unhealthy interactions in online conversations (Raman et al. 2020). We design toxicity features in order to evaluate the rudeness, disrespectfulness, or unreasonableness of posts within a conversation. We use Google's Perspective API[1] to extract the toxicity of text. The API returns a value between 0 and 1 for a given input, where 0 indicates complete inoffensiveness and 1 indicates absolute offensiveness. We construct three features with the aforementioned toxicity values: *title toxicity* to represent the toxicity of the title, *post toxicity* to represent the toxicity of the post, and *max. comment toxicity* to represent the most toxic comments in the conversation conversation.

Sentiment Sentiment is also highly related to contentiousness (Mishne, Glance, and others 2006; Pennacchiotti and Popescu 2010; Mejova et al. 2014). We use Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto and Gilbert 2014) to measure the sentiment of posts and their comments. We use the built-in VADER lexicons from `NLTK` (Bird and Loper 2004) to obtain a sentiment score for any given text. We measure the level of positivity and negativity in the post (*max. sentiment*, *min. post sentiment*), the maximum level of positivity in the comments (*max. comment sentiment*), the average amount of positivity in the comments (*avg. comment positive sentiment*), as well as the maximum and average amount of negativity in the comments (*min. comment sentiment*, and *avg. comment negative sentiment*) and use them as separate features for prediction.

LIWC The linguistic styles and word choices that speakers choose to convey their intent can reflect how they conceptualize meaning (Trott et al. 2020). Prior studies suggested that information gained from examining these lexical patterns can be useful in contentiousness prediction (Addawood et al. 2017; Addawood and Bashir 2016). To understand how linguistic styles affect contentious conversations, we employ the Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, and Booth 2001; Pennebaker et al. 2007). We construct a set of features to help capture lexical evidence using its 64 dimensions, including aspects such as attentional focus, social relationships, thinking styles, and others (Tausczik and Pennebaker 2010), by extracting a count vector of occurrences of words from the LIWC lexicon.

Text Representations We choose to represent the meaning of the text of each conversation conversation in two vector representations: TF-IDF and contextual representation from a finetuned BERT model. The TF-IDF representation is the sum of the tf-idf vector for the post and the average tf-idf vector of all the comments under the post. We set a maxi-

mum document frequency of 0.95 and exclude English stopwords. The BERT representation is computed using bert-base-uncased from `transformer` (Wolf et al. 2019) and is similarly the sum of the contextual representation for the post and the average comment representation. We use the second-to-last hidden states as the representation vectors.

User Factors

We also include a set of user factors that are closely connected with contentiousness (Addawood et al. 2017), which are often leveraged to understand the types of people who are participating in these conversations. We use the gender, location, and prolificity of Reddit users as operational features. Our choice of restricting to these factors is partially because this information can be extracted more easily and with high confidence from Reddit. Additionally, gender is often associated with different roles in online conversations and reactions to contentiousness (Herring et al. 2020). Similarly, location is informative due to its contextualized effect in understanding contentiousness in different communities (Chen and Berger 2013). Prolificity can be used to distinguish between newcomers and experienced members.

Gender and location are heuristically determined by textually examining a user's post history for noun phrases. Gender is determined by looking for gendered nouns or pronouns at the end of a noun phrase taking the form "I am...", or any other form of the copula. Genders are counted, with indeterminate and nonbinary genders being excluded from this count. Locations are extracted in a similar fashion, and vectorized as a count vector of geocoded locations or the raw location if the location fails to be geocoded. The distribution of user locations differs from the distribution of user locations generally, indicating a connection with contentiousness. Prolificity is calculated from the number of posts a user has made in conversations on the topic of interest. We define a user as prolific if she/he appears in her/his respective topic more than 25, 50, or 100 times in order to explore differences between different levels of prolificity.

To ensure quality, we validated the gender and locations of users extracted using the above steps. We find a female/male split of 47.84%/52.16%, 25.56%/74.44%, and 33.32%/66.68% for abortion, climate change, and gun control, respectively. The overall gender distribution closely matches surveyed Reddit demographics.[2] We also use location to validate the corpus we collected, where we find that the U.S. is the most common country at 44%, followed by Canada, India, the U.K, Brazil, and Germany. This does not perfectly mirror reported demographics, but comes close.[3]

Subreddit

The subreddit of a conversation is included as a one-hot encoded feature to account for its potentially confounding influence. We found no statistically significant ($p < 0.05$)

---

[1]https://www.perspectiveapi.com/

evidence that these distributions were different using the Kolmogorov-Smirnov 2 sample test (Smirnov 1939). The p for abortion, climate change, and gun control are 0.88, 0.60, and 0.80 respectively.

Prediction of Contentious Conversations

To answer RQ1, we train classifiers to predict whether a conversation is labeled as contentious or otherwise, incorporating the linguistic characteristics and the user factors as features along with subreddit in the classifier. Specifically, we show the performance of the classifier in making two types of predictions: predicting the contentiousness of the conversation using all the comments and predicting the contentiousness at different stages of these conversations. We experimented with two classifiers as baselines: Logistic Regression (LR) using TF-IDF features and Multilayer Perceptron (MLP) using BERT features. We then augmented these baselines to include the additional linguistic and user characteristics described in Conversation Characteristics. All the models were implemented using the off-the-shelf machine learning library `scikit-learn`. The pretrained BERT representation was further finetuned on Reddit comments before being used as input for the MLP. 10-fold stratified cross-validation was used to evaluate all the models and account for potentially confounding factors. The resulting scores are displayed in Table 1.

Conversation Level Contentiousness Prediction

As shown in Table 1, augmenting the baselines with the linguistic and the user features boosts the prediction performance significantly across all three topics. Specifically, in all cases, TF-IDF + All outperforms the baseline TF-IDF. Similarly, BERT + All also displays significant improvements over its baseline though not to the same degree as the TF-IDF baseline. A closer look reveals that the addition of certain features improves the results across all topics. On the other hand, some features only provide topic-specific improvements – for instance, adding the gender features does not improve the performance for the climate change topic. This suggests a distinction between features that are more related to contentiousness in general, and features that are more related to a specific divisive topic. Our modeling of contentiousness through this combination of linguistic and user features highlights these differences which would otherwise go unnoticed if using just the TF-IDF or BERT representations.

Early Prediction of Contentiousness

To predict contentiousness on partial conversations, we evaluated the previously trained models on increasing percentages of each conversation's comments. From a practical perspective, stakeholders such as moderators could be interested in solutions that predict contentiousness early with high performance *i.e.* without seeing all the comments or seeing only a limited number of comments in addition to the titles and posts. To this end, we test both TF-IDF + All and BERT + All models by gradually increasing the proportion of conversation comments in the order they were posted, beginning at 10% and ending at 100% and increasing in 10%

intervals. The results of this experiment for each topic are shown in Figure 1.

Overall, we found that the inclusion of more comments and features increases the accuracy and F1 score significantly for TF-IDF + All. BERT + All also improves over time, but less substantially. TF-IDF + All approaches the performance of BERT + All at 100% of comments, but BERT + All is much better at early stage prediction than TF-IDF + All. Specifically, models can obtain 88% of peak performance for accuracy after the first 20% of comments have been included, consistent across all three topics. The new features that we explore in this work provide significant performance benefits over the baseline, while remaining explain-able, both in full-prediction and partial prediction contexts.

Explanation of Contentious Case-Studies

This section investigates RQ2 on how different linguistic and conversational characteristics of conversations relate to their contentiousness and how these differ between different contentious topics. To do so, we utilize the regression coefficients from the aforementioned explainable classifier (TF-IDF + All), and further convert them into odds ratios for ease of explainability. We visualize the top 10 predictors with high odds ratios as well as the bottom 10 predictors with low odds ratios for each topic in Figure 2.

Negative Discourse Acts Matter The presence of certain discourse acts, in particular *disagreement*, *negative reaction*, and *question* are indicative of contentiousness across all three topics, with odds ratios greater than 1 for each. *Disagreement* and *negative reaction* are discourse acts with a primarily negative connotation, so their appearance here is to be expected. *Question*, however, is less so. Its appearance here indicates that the communication between the various sides of a divisive topic highly contributes to conversation contentiousness. Since *answer* and *elaboration* are significantly lower in terms of odds ratio, this would imply that communication is either one-sided or incomplete in these contentious conversations.

The discourse act bigrams that include *disagreement*, *negative reaction*, and *question* are also heavily represented in contentious conversations, with the vast majority of them having odds ratios greater than 1. This is even the case when paired with positive discourse acts, such as *agreement elaboration*, and *appreciation*, the cause of which may be that negative discourse acts are causing an effect similar to that of the negativity bias (Pratto and John 1991; Soroka, Fournier, and Nir 2019), a psychological phenomenon where negative events have greater psychological effects than positive ones, in the readers of these comments.

The positive discourse acts *elaboration*, *agreement*, and *appreciation*, however, have significantly lower average odds ratios, 1.002, 1.017, and 0.984, respectively. Bigrams containing these positive discourse acts similarly have lower odds ratios, again seen in Table **??**. Each of these discourse acts indicates a greater level of inter-group cooperation and communication within these divisive topics. The least controversial act *announcement* can be characterized generally

Features	Abortion				Climate change				Gun control			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
TF-IDF	0.586	0.578	0.583	0.546	0.644	0.614	0.581	0.581	0.653	0.640	0.617	0.613
+ Discourse act	0.660*	0.628*	0.613	0.599	**0.696**	0.649	0.617	0.624	**0.721***	**0.694***	0.657	0.668
+ Gender	0.644*	0.620*	0.595	0.585	0.639	0.611	0.569	0.575	0.685	0.667*	0.614	0.630
+ LIWC	0.665*	0.650	0.617	0.608	0.672	**0.671***	0.554	0.590	0.687	0.685	0.660	0.652
+ Location	0.624	0.594	0.581	0.567	0.641	0.617	0.586	0.582	0.693	0.672	0.644	0.647
+ Prolific	0.657*	0.618*	0.606	0.596	0.669*	0.640*	0.605*	0.607*	0.703*	0.684*	0.630	0.648
+ Sentiment	0.596	0.581	0.596	0.556	0.676	0.634*	0.625	0.616	0.665	0.650	0.625	0.625
+ Toxicity	0.609	0.586	0.608	0.565	0.661	0.638	0.614	0.610	0.660	0.646	0.618	0.618
+ Subreddit	0.550	0.551	0.547	0.548	0.649	0.659	0.607	0.631	0.597	0.610	0.541	0.573
+ All	**0.704***	**0.753***	**0.605**	**0.663**	**0.707**	**0.766**	0.593	**0.660**	0.683*	**0.727***	0.573	0.631
BERT	0.664	0.652	0.641	0.617	0.757	0.729	0.678	0.693	0.728	0.701	0.709	0.692
+ Discourse	0.702	0.664	0.638	0.639	0.772	0.754	0.692	0.716	0.746	**0.726**	0.703	0.707
+ Gender	0.660	0.620	0.623	0.605	0.761	0.724	0.701	0.702	0.739	0.706	0.714	0.702
+ LIWC	0.686	0.639	0.668	0.635	0.761	0.736	0.686	0.701	0.734	0.705	0.695	0.692
+ Location	0.652	0.626	0.699	0.630	0.746	0.712	0.689	0.689	0.741	0.723	0.698	0.701
+ Prolific	0.698	0.652	0.688	0.652	**0.780**	0.742	**0.714**	**0.719**	**0.752**	0.724	**0.716**	**0.711**
+ Sentiment	0.657	0.653	0.632	0.617	0.768	0.744	0.708*	0.717*	0.732	0.707	0.703	0.693
+ Toxicity	0.648	0.612	0.683	0.618	0.760	0.735	0.689	0.700	0.728	0.701	0.698	0.688
+ Subreddit	0.558	0.552	0.619	0.582	0.666	0.652	0.708	0.678	0.594	0.580	0.689	0.629
+ All	**0.715**	**0.686**	**0.712**	**0.678**	0.774	**0.755**	0.685	0.712	0.740	0.723	0.707	0.706

Table 1: Comparison of the performance for each feature set. We highlight the model with the best performance in bold. * denotes having statistically significant differences from their corresponding baselines at the $p < .05$ level.

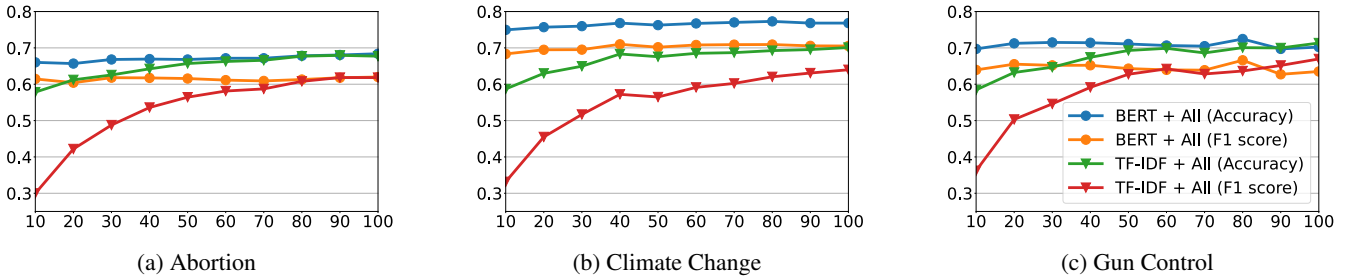(a) Abortion (b) Climate Change (c) Gun Control

Figure 1: Accuracy and F1 score for early prediction of contentiousness using TF-IDF + All and BERT + All models. The percentage of comments used as input is shown on the X axis.

as comments which are repositories of information, usually laid out in list form. These tend to be highly upvoted comments which guide the course of the conversation.

Location Varies Across Topics Location is, as shown by Figure 2, variable in terms of both which specific locations are represented and the importance of them across topics. Two classes of location can be seen in the features for abortion: locations which are generally more accepting of abortion and locations which are generally less accepting of abortion. We see examples of the first category in *in Germany* (odds ratio of 0.993), and *in Europe* (1.005) and examples of the second category in *in Georgia* (1.109) and *in North Texas* (1.076)[4]. The posts of these users, however, have very little to do with their place of residence. This may indicate that a wide spread of people with complex views on abortion or on a diminished importance of particular locations. We also notice a third category of locations, which are

places which are heavily featured in news related to abortion, with *Seattle*, *San Diego*, and *New York* having odds ratios of 1.202, 1.203, and 1.118, respectively.

Climate change may present a more political view of location. California (1.212) and New York (1.184) both had high values for the topic of climate change, and are both heavily featured in national news about climate change as well as politics concerning climate change. Both of these states are primarily represented by the United States' Democratic Party whose constituents who are, on average, significantly more concerned about climate change on average than those of the United States' Republican party.[5] Gun control, however, presents a very specific view of location and contentiousness, with locations generally associated with gun violence tragedies being highly predictive of contentiousness. For instance, Aurora, Colorado (which has an odds ratio of 1.196) where, in 2012, a shooting was carried out

---

[4]https://www.pewforum.org/religious-landscape-study/compare/views-about-abortion/by/state/

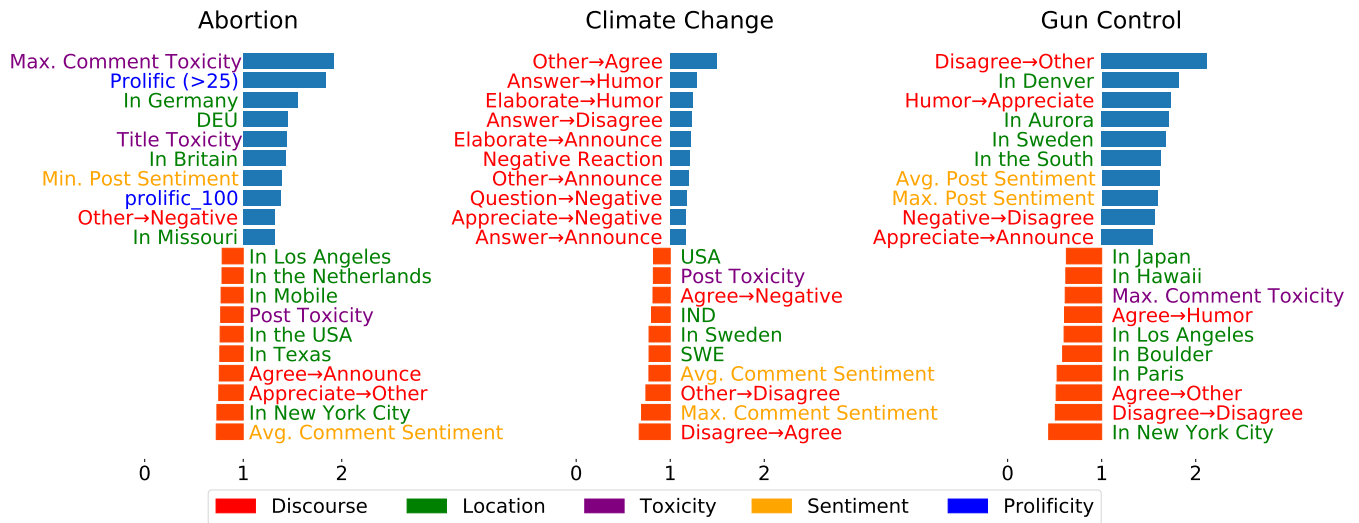[5]https://www.pewresearch.org/science/2015/07/01/chapter-2-climate-change-and-energy-issues/

Figure 2: Odds ratios of the top 20 categories for contentious and non-contentious conversations across "abortion", "climate change", and "gun control". Feature names are colored based on their category. Prolificity, negative discourse acts, and title toxicity are generally associated with contentiousness across all three topics. Location, other discourse acts, and sentiment vary across topics. The feature importance across topics indicates a divide between contextual features and generally predictive ones.

which killed 12 and injured 70.[6] Many users that we have identified as being from the Denver area, very near to Aurora, have participated in contentious conversations specifically about the shooting. Humans tend to react to events spatially close to them more concretely and emotionally(Fujita et al. 2006) which is a likely cause, along with the particularly volatile subject matter, for these conversations to become contentious.

The examination of locations in contentious conversations shows that location is highly variable across topics as well as within them. For abortion, location may only play a broader role in contentiousness prediction. In climate change, locations appear to be related to political affiliation. In gun control, locations may be extremely specific and related to similarly specific events. As such, detecting contentiousness in conversations concerning issues involving very specific, spatially close locations can be viewed very differently from broader issues involving more spatial separation.

Prolificity and Account Age Are Inversely Correlated with Contentiousness Prolificity and account age are two related user factors which relate to contentious conversations in very different ways. Prolificity tends to correlate with contentious conversations across all topics. This would indicate that users that are, to a certain extent, prolific within a given community are more likely to participate in potentially difficult conversations due to feeling more comfortable engaging with the community on these issues. For instance, conversations featuring high numbers of prolific users in the topic of abortion overwhelmingly featured religion, a topic which many are uncomfortable talking about in unfamiliar spaces. Non-contentious conversations still involved religion, but not nearly to the same extent. Account age, on the other hand, is correlated with a lack of contentiousness.

Gender Predicts Abortion Gender has an extremely variable impact on contentiousness depending heavily on whether or not gender is an inherent issue for the contentious topic being discussed. This is immediately evident in our analysis: the presence of specific genders is only appreciably different when looking at abortion, with male having an odds ratio of 1.128 compared to the female value of 1.061. The fact that gender can be so influential in the topic of abortion as to raise the odds ratio of male so high is a testament to the gender differences that can appear in contentious conversations (Herring et al. 2020) and that abortion is a topic that particularly concerns women more than men. Those studying conversations using our method in a particular community which may deal with female-centric issues more frequently than usual should pay more attention to gender when determining contentiousness. Notably, these results suggest that the differences in which female and male users engage with contentious discussion found in Herring et al. are highly variable with topic.

Toxicity Equals Contentiousness? High levels of toxicity in post titles are strong predictors of contentiousness, as demonstrated by odds ratios that are greater than 1.1 in all topics. It might be that some users choose contentious or provocative titles to attract more users to participate in conversations; similar trends have been found on Twitter (Reis et al. 2015), where the sentiment of the news headline is strongly related to the popularity of the news. As demonstrated in Figure 2, however, the toxicity of other parts of conversations are weighted differently between topics. For instance, *max. comment toxicity* contributes to contentiousness across all three topics while *avg. comment toxicity* only contributes to contentiousness in abortion. This might be because different topics have their own characteristics, with people involved in discussions about climate change being more likely to use milder language in comments, with an av-

erage toxicity level of 0.19. On the other hand, participants in abortion and gun control conversations are more likely to use more toxic language in comments, with an average toxicity level of 0.26 and 0.25, respectively. *Selftext toxicity* is relatively less predictive, with odds ratios of less than 1 among all topics. This might be because posts usually serve a pragmatic purpose, either as expansions or explanations of their title in order to provide more information about the poster's intent and context for discussions. It may also be that the content of the title colours commenters' opinions more strongly than the selftext, following the human preference for first impressions.

LIWC Offers Contextual View of Language Choice on Contentiousness In general, LIWC features offer a comprehensive view of users' linguistic styles and word choices and their influences on controversy. For simplicity, here we discuss two consistently informative LIWC features: *you* and *friends*. Here the linguistic category *you* contains second-person pronouns such as "you", and "yours". *you* is positively associated with contentiousness across all topics which might communicate to us that a repetitive utilization of *you* words increases social distance between conversation participants, and relatively related to relationship quality (Simmons, Chambless, and Gordon 2008). Such situations can be observed in our cases. The impact of the category *friends* is more variable, with climate change viewing it the most indicative of contentiousness (1.13) and abortion viewing it the least (0.91). This might indicate that higher mentions of *friend* words in abortion make them harder to serve as a distinction signal in contentiousness prediction, as abortion involves large amount of descriptions about relationships, for which people need to refer to related words related to friends. In both contentious and non-contentious conversations, we can easily find such discussions. Conversely, we found that in climate change *friends* tended to be associated with sarcasm, or negative experiences with friends. In general, these results are highly contextual with regards to target and community.

Discussion and Conclusion

This work investigated contentious discussions via case studies on three highly divisive topics: abortion, climate change, and gun control. Specifically, we operationalized a set of conversation characteristics — discourse acts, toxicity, sentiment, LIWC, text representation, and user factors — and demonstrated that these features contributed to both reasonable predictive performances and explainability in understanding contentious in conversations. In doing so we discovered that different linguistic and social features have different relationships to contentiousness and the topics themselves. Certain features — toxicity, prolificity, and discourse acts — remained stable predictors of contentiousness regardless of the topic. Other features — location, gender, sentiment, and LIWC — demonstrated highly contextual information and differences in the ways in which users engage with these contentious topics on Reddit. The findings suggest that contentiousness in conversations is subtle and that explainable models are necessary to identify these nuances

for both moderators, users, and researchers.

Implications

Our work has both theoretical and empirical guidelines to assist online media in defining, measuring, understanding, and even intervening with contentious content on social media platforms. First, our research provides conversational characteristics of contentious conversations which enable deep, contextual insights into the nature of these conversations. Second, our proposed generic conversation characteristics can be easily applied to build predictive models for understanding other similarly divisive topics beyond the three studied here, since they are not tailored to any specific data.

Practically, this work could motivate the design of tracking/monitoring systems that can visualize possibly contentious conversations in any specific groups to assist community moderators with content moderation. In addition to warning moderators early on in a post's lifespan whether or not it is of concern, this work has the ability to explain to moderators (and users) why a post is considered contentious and has attracted increased attention. Since the explanation of moderator action is a large contributor to user satisfaction (Jhaver et al. 2019), the explanation portion of our research is key for ensuring quality moderation strategies can be derived from our work. It also sheds light on designing better features to inform new community members on what aspects of the conversations might cause problematic experiences, allowing for increased retention of newer users who are not exposed to contentious content which they may not wish to engage with early on in their time on the subreddit. Since there is some amount of subject-matter familiarity required to properly explain the models we have proposed here, moderators which are very familiar with the content that is posted to their communities naturally fit this role. Our work will also help design novel tutorials to educate users on how to engage in discussions, especially in divisive topics, in a more civil and constructive manner. Particularly, these tutorials can be tailored to individual topics with the use of this work and subject-matter experts.

Generalization

Although our work focused only on three topics: abortion, climate change, and gun control, many of our features and characteristics of these divisive topics can generalize well to other topics. LIWC, toxicity, and sentiment have been shown to be powerful predictors of contentiousness from previous work(Hessel and Lee 2019; Mejova et al. 2014; Dori-Hacohen and Allan 2013). We expect the patterns that we have observed regarding user factors, excluding location, to generalize well. For gender, the strong correlation of gender with contentiousness in the topic of abortion suggests that topics which relate to gendered issues will hold a similar relationship to gender, while non-gendered issues will continue to display little to no relationship with gender. We do not expect that the same locations should emerge in a generalized version of this work, although that locations which are relevant to the issue being discussed will continue to appear. In a generalized version of this work, less specific measures of location diversity (e.g., entropy) may perform better. We

also hypothesize that discourse acts will hold generally for contentious conversations, and leave confirmation of that hypothesis to future work.

Limitations

Our work also has a few limitations. First, the distant-supervision method that we use to label the contentiousness in posts relies on upvote ratio, which may not be a perfectly accurate analogue for contentiousness, though this has been adopted and validated by prior work (Hessel and Lee 2019). The Rediquette (Reddit's rules of conduct) specifically states that users are not to "downvote an otherwise acceptable post because you do not personally like it." Although only a subset of users follow these recommendations, this indicates that the downvote button is mainly reserved for posts that break rules or do not belong on a particular subreddit. We partially circumvent this limitation by controlling for topic. Future work can conduct further annotations to validate the correlations between human judgement and our contentiousness measures. Another important limitation of this work is its correlational nature. We mainly characterize which features of conversations are correlated with contentiousness but not what causes that relationship. However, without true random-assignment experimentation we cannot assume that the relationships described reflect any degrees of causation.

Ethical Considerations

This research study has been approved by the Institutional Review Board (IRB) at the researchers' institution. In this work, we leverage no information that was not publicly available at the time of data collection, leveraging the public post histories of the users who participated in the conversations which we studied. As such, user private information is not disclosed, and none of the posts which were used to compute the gender and location features have been saved or need to be saved during the course of this computation.

References

[Addawood and Bashir 2016] Addawood, A., and Bashir, M. 2016. "what is your evidence?" a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, 1–11. Berlin, Germany: Association for Computational Linguistics.

[Addawood et al. 2017] Addawood, A.; Rezapour, R.; Abdar, O.; and Diesner, J. 2017. Telling apart tweets associated with controversial versus non-controversial topics. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, 32–41. Vancouver, Canada: Association for Computational Linguistics.

[Anderson and Huntington 2017] Anderson, A. A., and Huntington, H. E. 2017. Social media, science, and attack discourse: How twitter discussions of climate change use sarcasm and incivility. *Science Communication* 39(5):598–620.

[Awadallah, Ramanath, and Weikum 2012] Awadallah, R.; Ramanath, M.; and Weikum, G. 2012. Harmony and dissonance: organizing the people's voices on political controversies. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 523–532.

[Bird and Loper 2004] Bird, S., and Loper, E. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 214–217. Barcelona, Spain: Association for Computational Linguistics.

[Chen and Berger 2013] Chen, Z., and Berger, J. 2013. When, why, and how controversy causes conversation. *Journal of Consumer Research* 40(3):580–593.

[Choi, Jung, and Myaeng 2010] Choi, Y.; Jung, Y.; and Myaeng, S.-H. 2010. Identifying controversial issues and their sub-topics in news articles. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, 140–153. Springer.

[Chor et al. 2019] Chor, J.; Tusken, M.; Young, D.; Lyman, P.; and Gilliam, M. 2019. Factors shaping women's pre-abortion communication with members of their social network. *Journal of community health* 44(2):265–271.

[Conover et al. 2011] Conover, M. D.; Ratkiewicz, J.; Francisco, M. R.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political polarization on twitter. *Icwsm* 133(26):89–96.

[Demszky et al. 2019] Demszky, D.; Garg, N.; Voigt, R.; Zou, J.; Shapiro, J.; Gentzkow, M.; and Jurafsky, D. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2970–3005.

[Devlin et al. 2018] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Dori-Hacohen and Allan 2013] Dori-Hacohen, S., and Allan, J. 2013. Detecting controversy on the web. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 1845–1848.

[Fujita et al. 2006] Fujita, K.; Henderson, M. D.; Eng, J.; Trope, Y.; and Liberman, N. 2006. Spatial distance and mental construal of social events. *Psychological Science* 17(4):278–282.

[Garimella et al. 2017a] Garimella, K.; De Francisci Morales, G.; Gionis, A.; and Mathioudakis, M. 2017a. The effect of collective attention on controversial debates on social media. In *Proceedings of the 2017 ACM on Web Science Conference*, 43–52.

[Garimella et al. 2017b] Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2017b. The ebb and flow of controversial debates on social media. *arXiv preprint arXiv:1703.05994*.

[Guimaraes et al. 2019] Guimaraes, A.; Balalau, O.; Terolli, E.; and Weikum, G. 2019. Analyzing the traits and anomalies of political discussions on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 205–213.

[Herring et al. 2020] Herring, M.; Sweet-Cushman, J.; Prough, E.; and Vultee, F. 2020. Who dominates the conversation? the effect of gender, discussion medium, and controversy on political discussion. *Feminist Media Studies* 1–19.

[Hessel and Lee 2019] Hessel, J., and Lee, L. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1648–1659.

[Hutto and Gilbert 2014] Hutto, C., and Gilbert, E. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *International AAAI Conference on Web and Social Media*.

[Jang and Allan 2018] Jang, M., and Allan, J. 2018. Explaining controversy on social media via stance summarization. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1221–1224.

[Jhaver et al. 2019] Jhaver, S.; Appling, D. S.; Gilbert, E.; and Bruckman, A. 2019. "did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. *Proc. ACM Hum.-Comput. Interact.* 3(CSCW).

[Kim 2014] Kim, Y. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. Doha, Qatar: Association for Computational Linguistics.

[Matamoros-Fernández 2017] Matamoros-Fernández, A. 2017. Platformed racism: The mediation and circulation of an australian race-based controversy on twitter, facebook and youtube. *Information, Communication & Society* 20(6):930–946.

[Mejova et al. 2014] Mejova, Y.; Zhang, A. X.; Diakopoulos, N.; and Castillo, C. 2014. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152*.

[Mishne, Glance, and others 2006] Mishne, G.; Glance, N.; et al. 2006. Leave a reply: An analysis of weblog comments. In *Third annual workshop on the Weblogging ecosystem*. Edinburgh, Scotland.

[Olteanu et al. 2015] Olteanu, A.; Castillo, C.; Diakopoulos, N.; and Aberer, K. 2015. Comparing events coverage in online news and social media: The case of climate change. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, number CONF.

[Pennacchiotti and Popescu 2010] Pennacchiotti, M., and Popescu, A.-M. 2010. Detecting controversies in twitter: a first study. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics in a world of social media*, 31–32.

[Pennebaker et al. 2007] Pennebaker, J. W.; Chung, C. K.; Ireland, M.; Gonzales, A.; and Booth, R. J. 2007. The development and psychometric properties of liwc2007: Liwc. net.

[Pennebaker, Francis, and Booth 2001] Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

[Pratto and John 1991] Pratto, F., and John, O. P. 1991. Automatic vigilance: the attention-grabbing power of negative social information. *Journal of personality and social psychology* 61(3):380.

[Raman et al. 2020] Raman, N.; Cao, M.; Tsvetkov, Y.; Kästner, C.; and Vasilescu, B. 2020. Stress and burnout in open source: Toward finding, understanding, and mitigating unhealthy interactions. In *International Conference on Software Engineering, New Ideas and Emerging Results (ICSE-NIER)*.

[Reis et al. 2015] Reis, J.; Benevenuto, F.; de Melo, P. O.; Prates, R.; Kwak, H.; and An, J. 2015. Breaking the news: First impressions matter on online news. *arXiv preprint arXiv:1503.07921*.

[Rethmeier, Hübner, and Hennig 2018] Rethmeier, N.; Hübner, M.; and Hennig, L. 2018. Learning comment controversy prediction in web discussions using incidentally supervised multi-task CNNs. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 316–321. Brussels, Belgium: Association for Computational Linguistics.

[Schweizer, Davis, and Thompson 2013] Schweizer, S.; Davis, S.; and Thompson, J. L. 2013. Changing the conversation about climate change: A theoretical framework for place-based climate change engagement. *Environmental Communication: A Journal of Nature and Culture* 7(1):42–62.

[Simmons, Chambless, and Gordon 2008] Simmons, R. A.; Chambless, D. L.; and Gordon, P. C. 2008. How do hostile and emotionally overinvolved relatives view relationships?: What relatives' pronoun use tells us. *Family Process* 47(3):405–419.

[Smirnov 1939] Smirnov, N. V. 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou* 2(2):3–14.

[Soroka, Fournier, and Nir 2019] Soroka, S.; Fournier, P.; and Nir, L. 2019. Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences* 116(38):18888–18892.

[Tausczik and Pennebaker 2010] Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29(1):24–54.

[Tremayne and Minooie 2013] Tremayne, M., and Minooie, M. 2013. Opinion leadership on gun control in social networks: Preferential attachment versus reciprocal linking. *American Communication Journal* 15(4).

[Trott et al. 2020] Trott, S.; Torrent, T. T.; Chang, N.; and Schneider, N. 2020. (re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association*

for Computational Linguistics, 5170–5184. Online: Association for Computational Linguistics.

[Wasike 2017] Wasike, B. 2017. Persuasion in 140 characters: Testing issue framing, persuasion and credibility via twitter and online news articles in the gun control debate. *Computers in human behavior* 66:179–190.

[Wolf et al. 2019] Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

[Zhang et al. 2018] Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1350–1361. Melbourne, Australia: Association for Computational Linguistics.

[Zhang, Culbertson, and Paritosh 2017] Zhang, A. X.; Culbertson, B.; and Paritosh, P. 2017. Characterizing online discussion using coarse discourse sequences. In *Eleventh International AAAI Conference on Web and Social Media*.