Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Identifying Constructive Conflict in Online Discussions through Controversial yet Toxicity Resilient Posts

Ozgur Can Seckin*†, Bao Tran Truong†, Alessandro Flammini, Filippo Menczer

Observatory on Social Media, Indiana University, Bloomington

Abstract

Bridging content that brings together individuals with opposing viewpoints on social media remains elusive, overshadowed by echo chambers and toxic exchanges. We propose that algorithmic curation could surface such content by considering *constructive conflicts* as a foundational criterion. We operationalize this criterion through *controversiality* to identify challenging dialogues and *toxicity resilience* to capture respectful conversations. We develop high-accuracy models to capture these dimensions. Analyses based on these models demonstrate that assessing resilience to toxic responses is not the same as identifying low-toxicity posts. We also find that political posts are often controversial and tend to attract more toxic responses. However, some posts, even the political ones, are resilient to toxicity despite being highly controversial, potentially sparking civil engagement. Toxicity resilient posts tend to use politeness cues, such as showing gratitude and hedging. These findings suggest the potential for framing the tone of posts to encourage constructive political discussions.

Code — https://github.com/ocseckin/constructive_conflict
Datasets — https://zenodo.org/records/17167317

Introduction

Polarization has intensified globally in recent years (Boxell, Gentzkow, and Shapiro 2024), driven in part by the growing influence of social media (Lorenz-Spreen et al. 2023). In contrast to the original promise of social media as a modern public square that fosters discussions, echo chambers emerge where users preferentially consume content aligned with their ideologies (Garimella and Weber 2017; Flaxman, Goel, and Rao 2016; Barberá et al. 2015; Conover et al. 2011). Algorithms further amplify this selective exposure, segregating users into even more polarized groups, exacerbating affective polarization (Santos, Lelkes, and Levin 2021; Cho et al. 2020; Sasahara et al. 2021).

Theoretically, cross-cutting discussions can promote mutual understanding and greater tolerance for diversity (Kingwell 1994; Gutmann and Thompson 2009; Bohman and Rehg 1997), as well as correcting misperceptions about political out-groups (Voelkel et al. 2023). As such, encouraging individuals to engage with diverse perspectives can be a potential solution to the selective exposure problem and could reduce affective polarization.

However, empirical evidence shows that exposure to opposing views can sometimes reinforce existing beliefs, intensifying polarization (Bail et al. 2018). One of the possible causes is the negative tone accompanying diverse exposure (Efstratiou et al. 2023), which can emphasize ingrained negative emotions toward outgroups (Lerman et al. 2024). This could lead to hostility rather than fostering understanding. The impact of diverse exposure also hinges on the topic. For instance, positive outcomes are less likely in conversations about contentious topics (Santoro and Broockman 2022). Together, the tone and topic of conversations might explain the mixed empirical findings on the effects of diverse exposure on social media (Bail et al. 2018; Levy 2021; Santoro and Broockman 2022).

To effectively bridge people across divides, interventions must consider the nuance in both the tone and topic of conversations. The above literature suggests that interventions introducing diverse exposure hold promise only when disagreements remain civil. However, across platforms, controversies are consistently associated with increased toxicity (Avalle et al. 2024). This raises key questions: Can conversations about critical democratic issues be respectful? Are controversial political topics inherently prone to hostility?

We hypothesize that even highly controversial political topics can be framed in ways that reduce toxic responses — this is supported by the literature on conversational framing (Bao et al. 2021). Qualitative work has shown that toxicity is not intrinsic to political content. For example, topics often associated with high toxicity, such as war and conflicts, are not linked with toxicity when approached from a humanistic perspective. Prior work also demonstrates that toxicity is not tied to specific user groups (Mall et al. 2020), suggesting that politically active users can engage in civil discourse. Understanding the interplay between tone and topic, especially in topics where people tend to disagree, is critical for designing appropriate interventions.

While prior research has examined online controversies (Garimella et al. 2018; Mejova et al. 2014) and the influence of conversational tone on prosociality (Bao et al. 2021)

*Corresponding author. Email: oseckin@iu.edu
†Equal contributions.

or incivility (Zhang et al. 2018; Almerekhi et al. 2019), no work has studied the distinction between the controversiality, toxicity, and resilience to toxic responses of social media posts, or whether linguistic framing can reduce toxicity in conversations of controversial topics. This paper answers the following research questions:

- **Q1:** What is the relationship between the controversiality of a post, its toxicity, and its resilience to toxic responses?
- **Q2:** Which controversial topics are resilient to toxic responses?
- **Q3:** How do constructive and destructive conflicts differ in terms of tone?

In exploring these questions, we make the following contributions:

- First, we develop accurate prediction models for controversiality and toxicity resilience of social media posts. We apply the models to posts in diverse Reddit communities and quantify the relationship between these two attributes. We also quantify the prevalence of toxic posts in these communities.
- Second, we examine the topics of posts with sparks of *constructive conflict* — posts that are resilient to toxicity despite being highly controversial.
- Last, we characterize the difference between linguistic features in constructive and destructive conflicts, characterized by high and low resilience to toxic responses, respectively.

We find that while a significant portion of Reddit posts attract toxic responses, response toxicity is only partially predicted by the toxicity of a post itself. Controversial posts are often political and associate with more toxic responses. However, some political posts maintain resilience to toxicity despite being highly controversial. We demonstrate that such posts use more politeness linguistic strategies. These findings highlight the possibility of fostering constructive conflict on divisive issues.

We begin with a discussion of previous research in computational methods to measure toxicity resilience and controversiality. We then present the methods used to train our models and describe the findings. We conclude with implications for designing algorithmic curation and interfaces that foster civil discourse and reduce affective polarization.

Related Work

Ample work exists on developing and evaluating toxicity detection algorithms in online content (Risch and Krestel 2020; Pavlopoulos et al. 2020; Sheth, Shalin, and Kursuncu 2022), with widely used APIs such as Google's Perspective API (Lees et al. 2022) and OpenAI's omni-moderation (OpenAI 2024). In this paper, we adopt a widely used definition for online and political toxicity: toxic content includes expressions of disrespect that use insulting language, profanity, name-calling, personal attacks, and the use of racist, sexist, or xenophobic terms (Coe, Kenski, and Rains 2014; Kim et al. 2021; Lees et al. 2022). We focus on toxicity triggers, the starting points at which conversations turn toxic. We employ a similar approach to previous work, identifying these triggers as posts likely to provoke toxicity in subsequent threads (Almerekhi et al. 2019).

Many studies quantify the controversiality in various ways. Network-based approaches define controversial topics as highly clustered graphs, reflecting polarized viewpoints among social media communities (Garimella et al. 2018). Other work focuses on interactions such as edit histories on Wikipedia pages to capture the contentiousness of topics (Sumi, Yasseri et al. 2011; Vuong et al. 2008). Controversiality has also been inferred by linked Wikipedia pages (Dori-Hacohen, Jensen, and Allan 2016). Hybrid methods combine user-based, graph-based, and textual features to improve detection accuracy (Koncar, Walk, and Helic 2021; Benslimane et al. 2021). For this study, we adopt a prior definition of controversiality as the perception of whether an issue is likely to evoke disagreement (Sznajder et al. 2019).

Prior research examines the role of linguistic features such as politeness cues in prosocial (Bao et al. 2021) or asocial (Zhang et al. 2018) conversations. Connective language —language that signals openness to differing views— also helps with constructive discussions (Lukito et al. 2024). This line of research has motivated the development of many tools to improve the constructiveness of online discussions. Examples include designing mobile apps to encourage conversations among people differing political views (Doris-Down, Versee, and Gilbert 2013), and to deescalate heated threads in real-time (Chang, Schluger, and Danescu-Niculescu-Mizil 2022).

Methodology

We capture constructive conflicts by operationalizing three properties of posts: *controversiality* (C), *toxicity* (T), and *toxicity attraction* (TA). Note that TA is simply the opposite of toxicity resilience; in the remainder of the paper, we use TA for convenience. To quantify C and TA in social media posts, we develop two distinct models: the *C model* and the *TA model*. Both models leverage DistilBERT, a lightweight, effective, and widely used model for language representation (Sanh 2019). For training and evaluation, we use two data sources. The C model is trained on Wikipedia data, while the TA model is trained on Reddit data. Both models are evaluated using Reddit data. In the following sections, we detail the datasets, model training processes, and evaluation methods for each model.

Toxicity Attraction

The Reddit dataset consists of submissions (equivalent to original posts on other platforms and hereafter referred to as *posts*) and comments from 50 subreddits collected using the Pushshift API[1] between January 1 2021–December 31, 2023. These subreddits are chosen from the 150 most active subreddits[2] to capture a broad range of topics with mostly text content, including political (e.g., r/Conservative, r/politics), religious (e.g., r/atheism), and general interest subreddits (e.g., r/Music, r/NoStupidQuestions, r/gaming, r/re-

[1] pushshift.io

[2] reddit.com/best/communities/1

lationships, r/RandomThoughts, r/Showerthoughts). We aim to analyze conversations that have enough interactions. To do this, we extract only submissions with at least five comments. We further standardize conversation length by randomly sampling up to ten direct comments per submission (i.e., excluding replies to comments). Next, we remove posts (submissions and comments) that were deleted by authors or moderators. We find these posts either by matching them with the labels "[removed]" or "[deleted]," or by identifying standardized moderator messages. These messages contain the keywords "removed" or "deleted" and occur more than 20 times within a subreddit. We further preprocess the text in each post by removing URLs, multiple white spaces, and HTML-specific characters, such as ">." At the end of this process, the dataset contains 1,441,907 submissions and 12,866,675 associated comments, where each submission has between five and ten comments. See Appendix for the complete list of subreddits and the number of submissions in each of them.

The toxicity of posts can be measured using various models. We compare multiple models on a manually annotated subset of the Reddit data. The OpenAI omni-moderation-2024-09-26 model (OpenAI 2024) was used throughout this work to determine toxicity scores, as it is highly accurate, achieving a ROC-AUC score of 0.91 on a set of manually annotated Reddit posts (see Appendix for details on model selection).

We begin by exploring the relationship between toxicity and toxicity attraction. For this purpose, a submission or comment is classified as toxic when its toxicity score exceeds 0.5 — this threshold was chosen to balance precision and recall (see Appendix for more details on threshold selection). A submission is considered to be toxicity attracting if it has at least one toxic comment. Around 8% of submissions and 12% of comments exhibit toxicity. A significant portion of submissions, 53%, are toxicity-attracting. Only 6% of all submissions are both toxic and toxicity-attracting. This relationship is illustrated in Fig. 1, highlighting the distinction between a post's toxicity and its potential to attract toxic responses. The distinction between toxic and toxicity attracting posts remains when different thresholds are used to define toxicity attracting, although the absolute overlap varies, as expected (see Appendix). From here on, we abandon any threshold-based labels altogether: TA score is treated as a continuous measure — the average toxicity score of all comments a submission receives—and no further binarization or cut-offs are applied.

Our aim is to develop a TA regression model that takes the text of a submission as input and predicts its TA score, i.e., the average toxicity score of the comments it elicits. The TA model outputs a score between zero (not toxicity-attracting) and one (highly toxicity-attracting). It is trained using text merged from the title and body of Reddit submissions. We split the dataset into 70-15-15% for training, validation and testing, respectively. Fig. 2 shows the distribution of TA scores.

We use a pre-trained DistilBERT model, and adapt it for the TA regression task by adding a single node with linear activation function as the last layer. We apply a grid search for hyperparameter tuning, exploring learning rates (2×10^{-5}, 10^{-4}) and weight decay values (0, 0.05). The model is fine-tuned for a maximum of 5 epochs with a batch size of 32 and early stopping based on root mean squared error (RMSE) with a patience of 2 epochs. Fine-tuning takes about 7.5 hours on a single A100X NVIDIA GPU. The top performing model is optimized with a learning rate of 2×10^{-5} and a weight decay of 0. This model achieves a Spearman correlation of 0.70 on test set (0.74 in validation, both statistics $p < 0.01$) with a RMSE of .10 (.09).

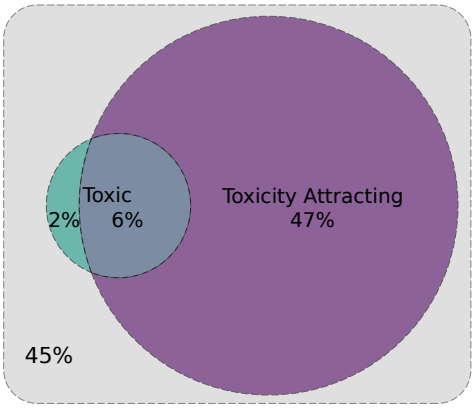

Figure 1: Overlap between toxic and toxicity-attracting submissions. 45% of all the submission are neither toxic nor toxicity attracting.

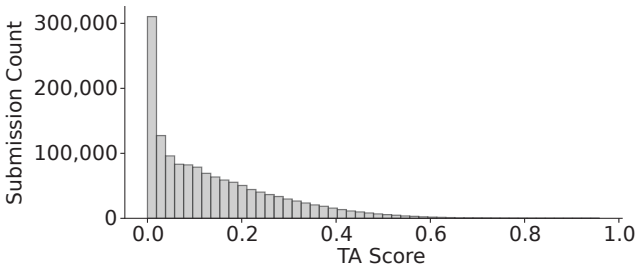

Figure 2: Distribution of toxicity attraction scores.

As a baseline to evaluate our model's accuracy, we also evaluate a linear regression TA model (ordinary least squares) where the independent variable is the submission toxicity score. This baseline achieves a RMSE of 0.13 and a Spearman correlation coefficient of 0.26 ($p < 0.01$), demonstrating that toxicity alone is a weak predictor of toxicity attraction. This relationship is further visualized in Fig. 3.

Controversiality

Our controversiality model is based on both a Wikipedia dataset and a definition of controversy by Sznajder et al. (2019). The dataset consists of 3,561 Wikipedia topics along with their short descriptions.[3] At least ten annotators were asked to rate each topic on whether people would likely

[3]The dataset is released under CC-BY-SA and is available at research.ibm.com/haifa/dept/vst/debating_data.shtml

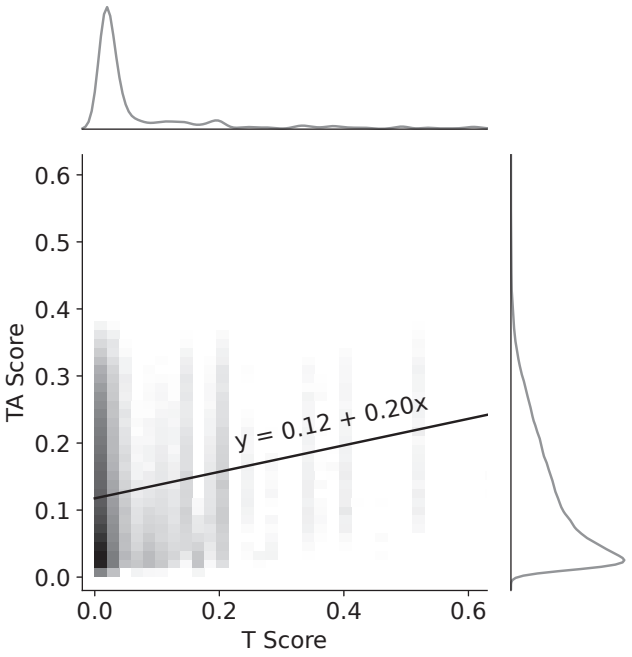

Figure 3: Distribution of submission toxicity and toxicity attraction scores. Color intensity in bins reflects log-transformed, normalized counts. The solid line represents the baseline regression model, with a coefficient standard error of 0.01.

argue about it, achieving a high Cohen's Kappa agreement of 0.53. We remove one entry missing a description and four entries with nonsensical descriptions, resulting in a total of 3,556 topics.

Examples of non-controversial topics are "Snow leopard," "copyright," and "theater"; controversial topics are "Israel, Palestine, and the United Nations," "Feminism," and "Pornography"; and topics with intermediate controversiality scores are "Michael Jackson's health and appearance," "Jehovah," and "1956 Winter Olympics."

The C model is trained using the title and description of Wikipedia entries. We split the dataset into 70-15-15% for training, validation and testing, respectively. We employ a pre-trained DistilBERT model, and adapt it for the C regression task by adding a single node with linear activation function as the last layer. We apply a grid search for hyperparameter tuning, exploring learning rates (2×10^{-5}, 10^{-5}, 10^{-4}) and weight decay values (0, 0.01, 0.05). The model is trained for 10 epochs with a batch size of 16 and early stopping based on RMSE with a patience of 3 epochs. The model is optimized using RMSE loss, but the selection of the best model uses Spearman correlation as we focus on preserving relative order rather than absolute values. The top-performing model is optimized with a learning rate of 2×10^{-5} and a weight decay of 0.05. Training takes around ten minutes on a single A100X NVIDIA GPU. This model achieves a Spearman correlation of 0.77 ($p < 0.01$) and a RMSE of 0.17 on the validation set, and a Spearman correlation of 0.76 ($p < 0.01$) with a RMSE of 0.16 on the test set. This demonstrates high accuracy in identifying controversial Wikipedia topics.

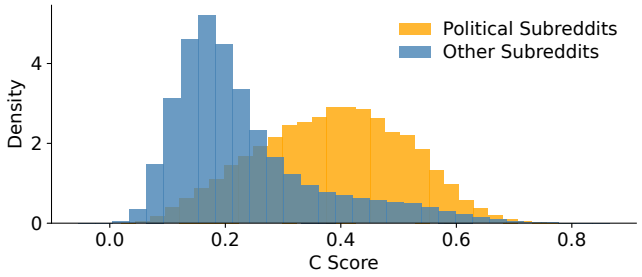

Figure 4: Distribution of C scores among submissions in political and other subreddits.

Once trained, the C model takes the text of a social media post and outputs a controversiality score between zero (non-controversial) and one (highly controversial). We assess the performance of the model by comparing the results to a manually annotated subset of the Reddit dataset. The C scores are skewed, with most posts being non-controversial. To make sure the annotation set encompasses a broad range of C scores, we sample ten posts from each decile. Two authors independently annotate comments in the sample in two rounds using the instruction: "Mark 1 if this comment mentions a topic that people are likely to argue about, 0 otherwise." After the first round of independently labeling a set of 200 posts, the authors collectively review their judgments to resolve disagreements and refine the definition of controversy. In the second round, they independently annotate sets of 100 posts, with 50 posts overlapping among the sets to assess inter-rater agreement. This results in a Cohen's Kappa of 0.82, indicating strong agreement. Finally, the model predictions are evaluated using a labeled set of 250 posts created from both rounds, achieving an ROC-AUC of 0.90. This demonstrates high accuracy in identifying controversial content on social media.

Although a submission's controversiality is often associated with its topic, the C model captures how controversiality manifests beyond topical content alone. Figure 4 presents the distribution of C scores within political and other subreddits, showing notable variation. This indicates that our model does not merely reflect topics (which would result in minimal variation, around 0), but instead identifies nuanced signals of controversy. Appendix provides examples to further illustrate how wording differences can influence whether a post provokes controversy (Salminen et al. 2020).

Linguistic Features

Let us outline a number of linguistic features that may play a role in the toxicity attraction of submissions.

- *Question-asking*. Asking questions in online discussions has been shown to enhance likeability (Huang et al. 2017) and perceived trustworthiness (Saltz, Jalan, and Acosta 2024). We represent the question-asking feature as the ratio of sentences with a question mark to all sentences in a post.

- *Elaboration*. Prior work suggests that details and coherence improve the perceived quality of communication (Crossley and McNamara 2016). Incorporating rhetorical devices and supporting evidence — such as narratives and factual content — can further improve persuasiveness and acceptance (Feng et al. 2023). It is therefore reasonable to assume that authors of longer, more elaborated texts aim to increase acceptance, making such posts less likely to elicit toxicity. We base our definition of *elaboration* on lexical diversity (Johansson 2008), captured by the number of unique words within key lexical groups (i.e., nouns, verbs, adjectives, and adverbs) in a post. The groups are identified using part-of-speech tagging via the NLTK library (Bird, Klein, and Loper 2009). Alternative operationalizations of elaboration in posts produce qualitatively comparable results (see Appendix).
- *Hedge usage*. Hedging words/phrases such as "perhaps," "it seems," and "I believe" can be used in contentious discussions to soften assertions and reduce vulnerability to criticism (Lakoff 1973; Crystal 1988). As a rhetorical strategy, hedging can enhance the persuasiveness and reception of written communication (Rezanejad, Lari, and Mosalli 2015). While excessive hedging may undermine perceptions of authoritativeness, moderate use has been shown to improve appearances of sociability, such as warmth and approachability (Hosman 1989). Prior research shows that conversations starting with hedged remarks maintain civility longer than those beginning with forceful questions or direct language (Zhang et al. 2018). These findings suggest that strategic hedging may help reduce the likelihood of eliciting toxic responses. In our analysis, we quantify *hedge usage* as the ratio of hedge-signaling words to the total word count of a post (Islam, Xiao, and Mercer 2020).
- *Gratitude*. Expressions of gratitude have been shown to strengthen social connections (Algoe 2012), foster reciprocal kindness (McCullough et al. 2001), and signal prosocial intentions (You 2006). *Gratitude* is a variable representing the ratio of gratitude words to all words in a post. To identify such words, we expand and use a lexicon of gratitude words, such as "thank you," and "grateful for" (Bao et al. 2021). See Appendix for the full lexicon.
- *Name-calling*. Name-calling directed at individuals or groups is often used as a tactic to belittle or harass others (Lenhart et al. 2016; Duggan 2017). Such harsh or demeaning language has been linked to online incivility (Coe, Kenski, and Rains 2014). We expect this association to hold in our dataset — specifically, that increased usage of proper nouns is associated with higher elicited toxicity. To capture this, we define the *name-calling* feature based on part-of-speech tagging via the NLTK library (Bird, Klein, and Loper 2009), specifically the ratio of proper nouns ("NNP" and "NNPS") to the total number of words in a post.
- *Polarity*. Polarity measures the negative or positive emotions of a post. The polarity of posts can have contagion effects — negative sentiment often amplifies adversarial reactions in already polarized discussions (Chang, May, and Lerman 2023), while positive sentiment tends to foster more positivity (Ferrara and Yang 2015; Kramer, Guillory, and Hancock 2014; Coviello et al. 2014). We measure the *polarity* of submissions using the scores generated by VADER, a widely adopted lexicon-based sentiment analyzer (Hutto and Gilbert 2014).

Results

Controversiality and Toxicity Resilience

We study the relationship between the controversiality of posts and their resilience to toxic responses using the predicted C and TA scores. Aligning with our intuition, controversial posts are more likely to draw toxic responses. The Spearman correlation between TA and C values is 0.48 ($p < 0.001$). This suggests that controversiality is a stronger predictor for the toxicity attraction of a post than its toxicity.

Topics

Let us examine how controversiality and resilience to toxic responses depend on different topics. To begin, we consider political and non-political content, as illustrated in Fig. 5 (middle panel). Submissions from political subreddits (i.e., r/Conservative, r/Liberal, r/politics) are significantly more controversial than non-political subreddits (e.g., r/RandomThoughts, r/gaming, and r/changemyview). Median C values are significantly different for political and non-political subreddits: 0.39 and 0.20, respectively; Mood's median test statistic 16,937, $p < 0.01$. Furthermore, submissions from political subreddits are more likely to attract toxicity than non-political subreddits. The median TA values for these subreddits are significantly different: 0.22 and 0.09, respectively; Mood's median test statistic 21,399, $p < 0.01$. While the prevalence of political content in non-political subreddits is less than in subreddits dedicated to political topics, the former may occasionally include political discussions.

The distribution of predicted TA and C scores for Reddit submissions and the topics associated with these scores are visualized in Fig. 5. The topics are generated using BERTopic, an effective and widely used topic modeling technique (Grootendorst 2022); implementation details are provided in Appendix. We remove topics supported by less than ten submissions. Fig. 5a-d illustrate combinations of TA and C scores across quadrants, divided by the median TA and C scores. For each quadrant, we randomly draw six topics from political, and six from non-political subreddits.

As expected, most low-controversy submissions are non-political (Fig. 5a,b). Among these, high TA submissions often involve personal insults or body references (Fig. 5a). Low TA submissions typically cover mundane topics such as books, car loans, dentists, and fashion, which are less likely to trigger strong disagreement or toxic language. Note that only one political topic appear in the low-controversy region (Fig. 5b).

Constructive conflicts — submissions that are controversial (high C), yet resilient to toxicity (low TA) — are demonstrated in Fig. 5d. These submissions mention policy-related

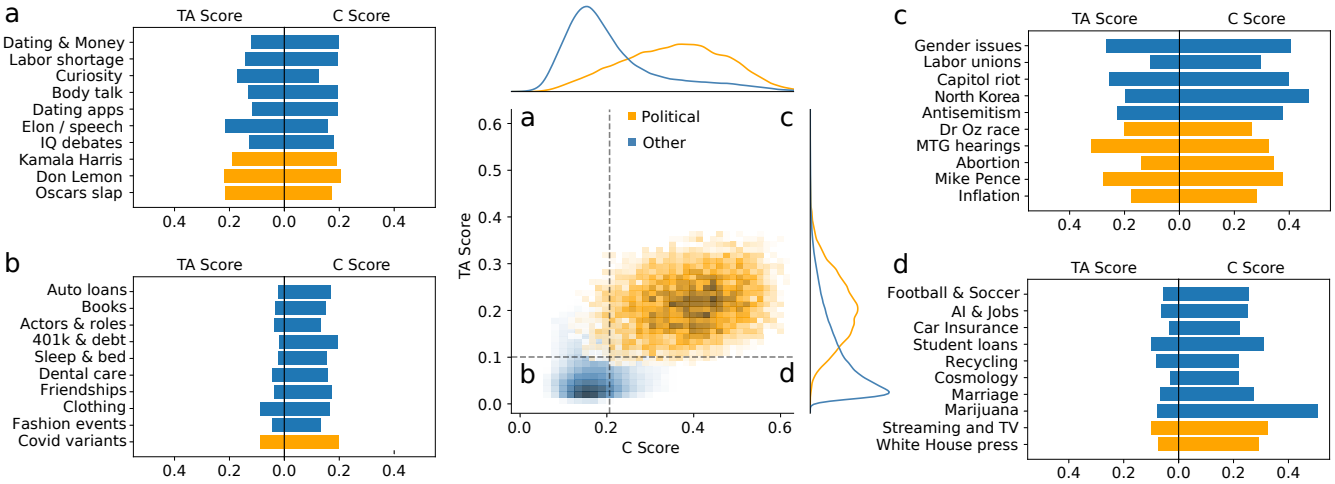

Figure 5: The relationship between toxicity attraction (TA) and controversiality (C) in Reddit. Middle panel: Distribution of TA and C scores of submissions in political (orange) and non-political (blue) subreddits. The dashed lines represent median TA and C values, separating quadrants of low/high C/TA. The bar plots in panels (a)-(d) illustrate topics sampled from submissions with TA and C values in these quadrants, along with their average scores. Topics are obtained using BERTopic and are abbreviated for clarity (see Appendix for the full keyword lists). (a) Topics sampled from submissions with high TA, low C scores. (b) Topics sampled from submissions with low TA, low C scores. (c) Topics sampled from submissions with high TA, high C scores. (d) Topics sampled from submissions with low TA, high C scores.

topics such as student loan forgiveness, artificial intelligence, big bang, and marijuana usage. Other issues in the high C - low TA quadrant, not shown in the figure, are climate change and alternative energy sources, healthcare and insurance policies, and vegetarianism. This underscores the possibility of discussing critical controversial issues in a respectful manner on social media platforms.

In contrast, toxicity-attracting submissions are about religious themes or highly polarizing political issues such as abortion rights, the January 6 riot, racial issues, and LGBTQ+ rights (Fig. 5c). Other topics in this quadrant (not depicted) are the Israel-Gaza conflict, the Ukraine war and school shootings.

Linguistic Features

The topics associated with destructive conflict — submissions that are controversial (high C) and toxicity-attracting (high TA) — deal with important issues that merit dialogue (Fig. 5c). Could conversations about these topics be reframed to become civil and constructive? To explore this question, we compare the linguistic cues in highly controversial submissions with low and high TA scores.

Consistent with our intuition, submissions employing civil features tend to attract less toxicity. In particular, question-asking, elaboration, hedges, gratitude, and positive polarity are negatively correlated with TA scores (Table 1, left column). Conversely, negative linguistic cues, such as negative polarity and name-calling, are associated with higher TA scores. These findings align with prior research, which demonstrates that politeness cues such as question-asking and elaborated text foster more prosocial conversations (Bao et al. 2021), while the absence of hedge and gratitude expressions are linked to conversations that begin civilly but eventually derail into incivility (Zhang et al. 2018).

Most of these patterns hold for non-controversial submissions as well. However, name-calling notably increases toxicity attraction in controversial submissions, while having no significant effect on non-controversial submissions. A possible explanation is that referencing individual names in controversial discussions may signal direct confrontation or the targeting of political figures or parties, which can heighten response toxicity. On the other hand, calling someone by name in non-controversial settings does not necessarily have a negative connotation. We also run a regression model for TA scores using C scores along with each linguistic feature as independent variables, showing that the effects of most features are robust (see Appendix).

Let us further illustrate the role of linguistic features in evoking toxicity with examples from our dataset. We select a set of submissions discussing highly polarized issues according to a recent Gallup Poll (Newport 2024). Table 2 shows pairs of submissions about the same high-C topics, but markedly different TA scores for comparison.

For the topic of abortion, a question-seeking post that includes a hedging symbol ("think") has a significantly lower TA score (39th percentile) than a post with name-calling and an accusatory tone (TA in the 99th percentile). For the topic of healthcare, a factual question about policy outcomes demonstrates how elaboration and asking questions can reduce the likelihood of attracting toxicity compared to inflammatory language targeting. These submissions have TA scores in the 10th and 75th percentiles, respectively. For the topic of gun laws, the same features result in TA scores in the

Linguistic feature	Controversial		Non-controversial	
	Corr.	%	Corr.	%
Question asking	-0.44	47	-0.38	50
Elaboration	-0.18	100	-0.31	100
Hedge usage	-0.16	55	-0.17	56
Gratitude usage	-0.06	3	-0.11	5
Positive polarity	-0.10	33	-0.13	46
Negative polarity	0.11	48	0.08	26
Name calling	0.36	81	0.002	64

Table 1: Correlations between linguistic feature usage and TA scores for controversial and non-controversial Reddit posts, along with the percentage of posts in which these features appear. Point-biserial correlation is used for binary features, i.e., "Positive Polarity" and "Negative Polarity," while Spearman correlation is used for other features. All correlations are significant with ($p < 0.01$) except for name calling under non-controversial, which is non-significant.

30th and 95th percentiles, respectively. The effect of showing gratitude and hedged phrasing is seen across examples on different topics, where it results in lower TA scores for submissions about LGBTQ+ and climate change (TA scores in the 33rd and 6th percentiles, respectively). On the contrary, dismissive rhetoric increases TA (scores in the 99th and 92nd percentiles for submissions about LGBTQ+ and climate change, respectively). These examples show that highly polarizing topics can be discussed civilly using various linguistic features.

Discussion

The rise of affective polarization and its link to social media usage (Lorenz-Spreen et al. 2023), has raised concerns about the potential harm caused by algorithms employed by social media platforms. These concerns have spurred research into alternative, "bridging" algorithms that promote cross-ideological understanding and respect (Ovadya and Thorburn 2023; Piccardi et al. 2024a). One way to operationalize bridging heuristics is by encouraging diverse exposure to cross-cutting content (Gutmann and Thompson 2009; Bohman and Rehg 1997; Levy 2021). We propose that constructive conflicts — posts that are controversial, yet resilient to toxicity — can provide a foundational condition to promote such exposure.

We develop accurate machine learning models to quantify the controversiality and toxicity resilience of posts, and analyze the results across diverse Reddit communities. Previous research shows that across platforms like Facebook, Gab, and Twitter (but not Reddit), controversy is positively correlated with increased toxicity (Avalle et al. 2024). Our work bridges the literature on online incivility and conversational framing, showing that political controversy and toxicity are not inherently linked, and that language framing plays a critical mediating role. This suggests the feasibility of achieving a constructive online environment.

The proposed attributes can be flexibly integrated into recommendation algorithms to align with specific goals. For instance, users in open ecosystems like Bluesky could use these metrics to create custom feed generators (e.g., FeedGenerator[4]) to prioritize posts based on toxicity resilience, controversiality, or both. Toxicity resilience predicts the likelihood of toxic subsequent comments to a post, accounting for posts at risk of high toxicity that might be overlooked in content-level analyses. More importantly, optimizing only for low-toxicity content may disproportionately favor neutral content and overlook controversial content that attracts diverse perspectives. By simultaneously considering controversiality together with toxicity resilience, we can create a ranking heuristic to demote negativity while retaining a diversity of viewpoints.

Since posts that attract toxicity are often linked to high-arousal emotions that drive online engagement (Ferrara and Yang 2015; Robertson et al. 2023), demoting these posts could reduce overall platform engagement. The most valid usability test to examine the effect of ranking heuristics on engagement would involve deploying the ranking in a field experiment (Piccardi et al. 2024a). However, such a design is time-consuming and costly, and therefore remains outside the scope of the current work.

Observational analysis can offer insights into the potential effects of our framework on engagement, but it cannot establish causality, and may not even yield accurate correlations. Specifically, data available through the platform reflects the engagement and exposure shaped by existing recommendation algorithms, while posts surfaced through any alternative rankings would likely generate different engagement dynamics. Consequently, simply correlating proposed ranking scores with engagement metrics from Reddit's current system likely overestimates or misrepresents the true impact of our framework. Further research is needed to fully understand and address this trade-off before integrating the heuristic into platform ranking algorithms.

The proposed methods enable the early identification of discussions at risk of escalating toxicity. Platforms and moderators can leverage these predictions to implement safeguards, such as adding warnings or context labels to posts that are likely to elicit toxic responses.

Our analysis reveals that even conversations on deeply divisive issues like abortion, gun laws, and LGBTQ+ rights can resist toxic responses through strategic language choices. Specific linguistic strategies, such as questioning, hedging, and expressing gratitude, increase a post's resistance to toxic responses. These findings inform the design of platform affordances that encourage healthier online discussions. For instance, prompts could guide users toward interactions that reduce the risk of toxicity. Another possible application of our analysis is in platform interfaces that highlight politeness cues when displaying comments.

While powerful large language models could be useful aids to algorithmic ranking (Piccardi et al. 2024b) and rephrasing issues (Costello, Pennycook, and Rand 2024), their high computational cost can cause delay in real-time user experience, decreasing user engagement. The proposed

[4]github.com/bluesky-social/feed-generator

Topic	Submission	TA %	C %
Abortion	VP Kamala Harris continues to tell Christians they should support the killing of pre-born humans.	99	91
	Should abortion be limited? I'm asking if abortion should be limited around 12-15 weeks? By then the fetus has a heartbeat, nervous system, functional organs, and yes eyes and mouth. Also women have bumps by then too. 99% of women know they are pregnant before 12 weeks and can make decision. I don't think abortion should be totally banned but 15 week is a reasonable compromise.	39	96
Healthcare	If the US was actually serious about supporting small businesses, they would enact universal healthcare. Working for a small business is effectively a death sentence, either financially or physically or both. Corporations hold the golden phallus of health care to dangle in front of your face. No health care only strengthens corporate asphyxiation on workers.	75	98
	Would making healthcare free in the U.S. be more expensive due to higher taxes, or cheaper due to no insurance payments/co-pays? I've been wondering this since the topic of "If we make healthcare free our taxes will go up". Although this is obviously true, I find that people forget they also wont be directly shilling out money every year for their insurance. So my question is in the title, I know taxes in other countries are a lot higher than in the U.S. because of that reason, but when incorporating the money you technically save from insurance and copayments would it still be that much of a difference?	10	88
Gun Laws	Enough is enough. Anyone who opposes gun control in this country is psychopathic unamerican scum.	95	98
	What are actual ways where gun laws can be re-established where both parties can be satisfied (whether for or against gun-control)?	30	99
LGBTQ+	Hating someone for being gay is a lifestyle choice.	99	84
	what's it called when you feel as if you were born the wrong gender but don't wanna go through with a transition of any sort because you know you wouldn't really be that gender? i in no way mean to invalidate anyone at all but i've been having some thoughts bout myself but i know it wouldn't be the same as being born as the opposite gender, i wanna look more into this so if someone could point me in the right direction of what to be looking up that would be great, and thank you in advance :)	33	51
Climate Change	This idiot still pushing climate change as a hoax. Same as he did with COVID.	92	72
	If global warming becomes a problem couldn't everyone just move to a cooler place Like say the temperatures get really hot Couldn't everyone just move to a place/country that is colder? Or for flooding couldn't everyone just move inland and be all good?	6	81

Table 2: Examples of posts discussing highly polarized issues (according to a recent Gallup Poll (Newport 2024)) in our dataset. Random pairs of Reddit posts about similar topics with markedly different TA scores. Percentiles for TA and C scores are shown, where higher percentiles indicate higher scores. Topics are obtained using BERTopic. Topic names are abbreviated for clarity: Abortion (abortion_roe_abortions_wade), Healthcare (insurance_healthcare_medical_gp), Gun Laws (gun_guns_shootings_weapons), LGBTQ+ (gender_trans_gay_nonbinary), and Climate Change (climate_warming_global_change).

models utilize DistilBERT, a lightweight transformer model with short inference time that enables efficient real-time integration into recommendation systems. Research indicates that BERT-based architectures outperform GPT models in detecting certain linguistic features (Lukito et al. 2024), supporting our architectural choice.

This study has several limitations. First, we rely on the OpenAI omni-moderation model to identify toxic submissions and comments, which means our findings inherit any shortcomings or biases of the model. Second, the proposed toxicity attraction model is trained using Reddit data and our specific definition of toxicity, limiting their generalizability to other platforms. We operationalize the toxicity attraction of a post as the average toxicity of its comments. Our results are robust under an alternative definition, where the TA score is the ratio of a submission's toxic comments. This alternative method strongly correlates with our proposed TA metric (Spearman correlation: $r = 0.92, p < 0.01$). In addition, despite their high accuracy, our proposed models are not immune to prediction errors, such as assigning a high C score to low-controversial posts or a high TA score to posts that are not truly toxicity-attracting. Applications utilizing these models should exercise caution by fine-tuning them with platform- and use case-specific data and regularly retraining them with updated datasets and annotations to maintain reliability. Lastly, implementing the controversiality model in real-world applications would require a frequently-updated method for generating predictions. However, our model shows strong potential in capturing recent topics effectively. For example, despite being trained on 2019 Wikipedia data, it successfully identifies events from subsequent years, such as the January 6 riot, as highly controversial topics, indicating reasonable robustness.

Ethical Statement

We have taken several steps for ethical considerations. First, we do not use any personally identifiable information (PII). We utilized two datasets for our research. The first is the Wikipedia Dataset by (Sznajder et al. 2019), which contains topics and their short descriptions. This dataset does not contain any PII. The second is the Reddit dataset, consisting of public submissions and comments. While the posts may include PII, such as usernames or self-disclosures, the dataset is limited to publicly available posts. We collected this dataset using the Pushshift API, adhering to the platform's terms of service to ensure compliance. In addition, the dataset is stored on a server with restricted access. Lastly, we release the data in Zenodo[5] and the source code in a GitHub repository [6]. To uphold ethical standards, the curated datasets exclude raw text and any PII, providing only post IDs and extracted features, such as polarity, toxicity scores, or indicators like the presence of a question mark. These features cannot be reverse-engineered to reveal PII.

Amid increasing concerns of polarization, our study enhances the understanding of bridging conversations by providing insights on how conflicts are discussed online. We hope to inform the design of social media algorithms and interfaces that would promote dialogues productive to democratic deliberation.

[5]zenodo.org/records/17167317

[6]github.com/ocseckin/constructive_conflict

Acknowledgements

We are grateful to Giovanni Luca Ciampaglia, Do Won Kim, and Saumya Bhadani for helpful discussions. This work was supported in part by the Swiss National Science Foundation (Sinergia grant CRSII5_209250) and by the Knight Foundation. This work used the IU JetStream 2 computational infrastructure through allocation CIS240118 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296 (Hancock et al. 2021; Boerner et al. 2023).

References

Algoe, S. B. 2012. Find, remind, and bind: The functions of gratitude in everyday relationships. *Social and personality psychology compass*, 6(6): 455–469.

Almerekhi, H.; Kwak, H.; Jansen, B. J.; and Salminen, J. 2019. Detecting toxicity triggers in online discussions. In *Proceedings of the 30th ACM conference on hypertext and social media*, 291–292.

Avalle, M.; Di Marco, N.; Etta, G.; Sangiorgio, E.; Alipour, S.; Bonetti, A.; Alvisi, L.; Scala, A.; Baronchelli, A.; Cinelli, M.; et al. 2024. Persistent interaction patterns across social media platforms and over time. *Nature*, 628(8008): 582–589.

Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.

Bao, J.; Wu, J.; Zhang, Y.; Chandrasekharan, E.; and Jurgens, D. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021*, 1134–1145.

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10): 1531–1542.

Benslimane, S.; Azé, J.; Bringay, S.; Servajean, M.; and Mollevi, C. 2021. Controversy detection: a text and graph neural network based approach. In *Proceedings of the 22nd International Conference on Web Information Systems Engineering*, 339–354. Springer.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Boerner, T. J.; Deems, S.; Furlani, T. R.; Knuth, S. L.; and Towns, J. 2023. ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support. In *Practice and Experience in Advanced Research Computing 2023: Computing for the Com-*

mon Good, PEARC '23, 173–176. New York, NY, USA: Association for Computing Machinery. ISBN 9781450399852.

Bohman, J.; and Rehg, W. 1997. *Deliberative democracy: Essays on reason and politics*. MIT press.

Boxell, L.; Gentzkow, M.; and Shapiro, J. M. 2024. Cross-country trends in affective polarization. *Review of Economics and Statistics*, 106(2): 557–565.

Chang, J. P.; Schluger, C.; and Danescu-Niculescu-Mizil, C. 2022. Thread with caution: Proactively helping users assess and deescalate tension in their online discussions. In *Proceedings of the ACM on Human-Computer Interaction*, volume 6, 1–37.

Chang, R.-C.; May, J.; and Lerman, K. 2023. Feedback Loops and Complex Dynamics of Harmful Speech in Online Discussions. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 85–94. Springer.

Cho, J.; Ahmed, S.; Hilbert, M.; Liu, B.; and Luu, J. 2020. Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. *Journal of Broadcasting & Electronic Media*, 64(2): 150–172.

Coe, K.; Kenski, K.; and Rains, S. A. 2014. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of communication*, 64(4): 658–679.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 89–96.

Costello, T. H.; Pennycook, G.; and Rand, D. G. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714): eadq1814.

Coviello, L.; Sohn, Y.; Kramer, A. D.; Marlow, C.; Franceschetti, M.; Christakis, N. A.; and Fowler, J. H. 2014. Detecting emotional contagion in massive social networks. *PloS one*, 9(3): e90315.

Crossley, S. A.; and McNamara, D. S. 2016. Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research*, 7(3): 351–370.

Crystal, D. 1988. On keeping one's hedges in order. *English Today*, 4(3): 46–47.

Dori-Hacohen, S.; Jensen, D.; and Allan, J. 2016. Controversy detection in wikipedia using collective classification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 797–800.

Doris-Down, A.; Versee, H.; and Gilbert, E. 2013. Political blend: an application designed to bring people together based on political differences. In *Proceedings of the 6th International Conference on Communities and Technologies*, 120–130.

Duggan, M. 2017. Online Harassment 2017. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/.

Efstratiou, A.; Blackburn, J.; Caulfield, T.; Stringhini, G.; Zannettou, S.; and De Cristofaro, E. 2023. Non-polar opposites: analyzing the relationship between echo chambers and hostile intergroup interactions on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 197–208.

Feng, G. C.; Luo, Y.; Yu, Z.; and Wen, J. 2023. Effects of rhetorical devices on audience responses with online videos: An augmented elaboration likelihood model. *PLoS One*, 18(3): e0282663.

Ferrara, E.; and Yang, Z. 2015. Measuring emotional contagion in social media. *PloS one*, 10(11): e0142390.

Flaxman, S.; Goel, S.; and Rao, J. M. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1): 298–320.

Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1): 1–27.

Garimella, V. R. K.; and Weber, I. 2017. A long-term analysis of polarization on Twitter. In *Proceedings of the International AAAI Conference on Web and social media*, volume 11, 528–531.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Gutmann, A.; and Thompson, D. F. 2009. *Democracy and disagreement*. Harvard University Press.

Hancock, D. Y.; Fischer, J.; Lowe, J. M.; Snapp-Childs, W.; Pierce, M.; Marru, S.; Coulter, J. E.; Vaughn, M.; Beck, B.; Merchant, N.; Skidmore, E.; and Jacobs, G. 2021. Jetstream2: Accelerating cloud computing via Jetstream. In *Practice and Experience in Advanced Research Computing 2021: Evolution Across All Dimensions*, PEARC '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450382922.

Hanu, L.; and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Hosman, L. A. 1989. The evaluative consequences of hedges, hesitations, and intensifies: Powerful and powerless speech styles. *Human communication research*, 15(3): 383–406.

Huang, K.; Yeomans, M.; Brooks, A. W.; Minson, J.; and Gino, F. 2017. It doesn't hurt to ask: Question-asking increases liking. *Journal of personality and social psychology*, 113(3): 430.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 216–225.

Islam, J.; Xiao, L.; and Mercer, R. E. 2020. A lexicon-based approach for detecting hedges in informal text. In *Proceedings of the Language Resources and Evaluation Conference*, 3109–3113.

Johansson, V. 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, 53: 61–79.

Kim, J. W.; Guess, A.; Nyhan, B.; and Reifler, J. 2021. The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, 71(6): 922–946.

Kingwell, M. 1994. *A civil tongue: Justice, dialogue, and the politics of pluralism*. Penn State Press.

Koncar, P.; Walk, S.; and Helic, D. 2021. Analysis and prediction of multilingual controversy on reddit. In *Proceedings of the 13th ACM Web Science Conference 2021*, 215–224.

Kramer, A. D.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24): 8788–8790.

Lakoff, G. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4): 458–508.

Lees, A.; Tran, V. Q.; Tay, Y.; Sorensen, J.; Gupta, J.; Metzler, D.; and Vasserman, L. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 3197–3207.

Lenhart, A.; Ybarra, M.; Zickuhr, K.; and Price-Feeney, M. 2016. Online Harassment, Digital Abues, and Cyberstalking in America.

Lerman, K.; Feldman, D.; He, Z.; and Rao, A. 2024. Affective polarization and dynamics of information spread in online networks. *npj Complexity*, 1(1): 8.

Levy, R. 2021. Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, 111(3): 831–870.

Lorenz-Spreen, P.; Oswald, L.; Lewandowsky, S.; and Hertwig, R. 2023. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature human behaviour*, 7(1): 74–101.

Lukito, J.; Chen, B.; Masullo, G.; and Stroud, N. 2024. Comparing a BERT Classifier and a GPT classifier for Detecting Connective Language Across Multiple Social Media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 19140–19153.

Mall, R.; Nagpal, M.; Salminen, J.; Almerekhi, H.; Jung, S.-G.; and Jansen, B. J. 2020. Four types of toxic people: Characterizing online users' toxicity over time. In *Proceedings of the 11th nordic conference on human-computer interaction: Shaping experiences, shaping society*, 1–11.

McCarthy, P. M.; and Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2): 381–392.

McCullough, M. E.; Kilpatrick, S. D.; Emmons, R. A.; and Larson, D. B. 2001. Is gratitude a moral affect? *Psychological bulletin*, 127(2): 249.

McInnes, L.; Healy, J.; Astels, S.; et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11): 205.

Mejova, Y.; Zhang, A. X.; Diakopoulos, N.; and Castillo, C. 2014. Controversy and Sentiment in Online News. ArXiv:1409.8152 [cs].

Newport, F. 2024. Update: Partisan gaps expand most on government power, climate.

OpenAI. 2024. Moderation Endpoint - OpenAI API.

Ovadya, A.; and Thorburn, L. 2023. Bridging systems: open problems for countering destructive divisiveness across ranking, recommenders, and governance. *Knight First Amend. Inst.*, 23-11.

Pavlopoulos, J.; Sorensen, J.; Dixon, L.; Thain, N.; and Androutsopoulos, I. 2020. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*.

Piccardi, T.; Saveski, M.; Jia, C.; Hancock, J.; Tsai, J. L.; and Bernstein, M. S. 2024a. Reranking Social Media Feeds: A Practical Guide for Field Experiments. *arXiv preprint arXiv:2406.19571*.

Piccardi, T.; Saveski, M.; Jia, C.; Hancock, J. T.; Tsai, J. L.; and Bernstein, M. 2024b. Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity. *arXiv preprint arXiv:2411.14652*.

Rezanejad, A.; Lari, Z.; and Mosalli, Z. 2015. A cross-cultural analysis of the use of hedging devices in scientific research articles. *Journal of Language Teaching and Research*, 6(6): 1384–1392.

Risch, J.; and Krestel, R. 2020. Toxic comment detection in online discussions. *Deep learning-based approaches for sentiment analysis*, 85–109.

Robertson, C. E.; Pröllochs, N.; Schwarzenegger, K.; Pärnamets, P.; Van Bavel, J. J.; and Feuerriegel, S. 2023. Negativity drives online news consumption. *Nature Human Behaviour*, 7(5): 812–822.

Salminen, J.; Sengün, S.; Corporan, J.; Jung, S.-g.; and Jansen, B. J. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PloS one*, 15(2): e0228723.

Saltz, E.; Jalan, Z.; and Acosta, T. 2024. Re-ranking news comments by constructiveness and curiosity significantly increases perceived respect, trustworthiness, and interest. *arXiv preprint arXiv:2404.05429*.

Sanh, V. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Santoro, E.; and Broockman, D. E. 2022. The promise and pitfalls of cross-partisan conversations for reducing affective polarization: Evidence from randomized experiments. *Science advances*, 8(25): eabn5515.

Santos, F. P.; Lelkes, Y.; and Levin, S. A. 2021. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, 118(50): e2102141118.

Sasahara, K.; Chen, W.; Peng, H.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2021. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 4(1): 381–402.

Sheth, A.; Shalin, V. L.; and Kursuncu, U. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490: 312–318.

Sumi, R.; Yasseri, T.; et al. 2011. Edit wars in Wikipedia. In *2011 IEEE 3rd International Conference on Social Computing*, 724–727. IEEE.

Sznajder, B.; Gera, A.; Bilu, Y.; Sheinwald, D.; Rabinovich, E.; Aharonov, R.; Konopnicki, D.; and Slonim, N. 2019. Controversy in context. *arXiv preprint arXiv:1908.07491*.

Voelkel, J. G.; Chu, J.; Stagnaro, M. N.; Mernyk, J. S.; Redekopp, C.; Pink, S. L.; Druckman, J. N.; Rand, D. G.; and Willer, R. 2023. Interventions reducing affective polarization do not necessarily improve anti-democratic attitudes. *Nature human behaviour*, 7(1): 55–64.

Vuong, B.-Q.; Lim, E.-P.; Sun, A.; Le, M.-T.; Lauw, H. W.; and Chang, K. 2008. On ranking controversies in wikipedia: models and evaluation. In *Proceedings of the 2008 international conference on Web search and data mining*, 171–182.

You, H. W. I. C. 2006. Gratitude and Prosocial Behavior. *Psychological Science*, 17(4): 319–325.

Zhang, J.; Chang, J. P.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Thain, N.; and Taraborelli, D. 2018. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes**
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes**
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes, please refer to section *Discussion*.**
   (e) Did you describe the limitations of your work? **Yes, please refer to section *Discussion*.**
   (f) Did you discuss any potential negative societal impacts of your work? **Yes, please refer to section *Ethical Statement***
   (g) Did you discuss any potential misuse of your work? **No. We do not anticipate any direct misuse of our work since the model and the results are only applicable in a specified context. We try to minimize the possibility of our results being misinterpreted by carefully presenting the results and acknowledging limitations**
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes, please refer to section *Ethical Statement***
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**
2. Additionally, if your study involves hypotheses testing...
   (a) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **Yes, please refer to sections *Discussion* and *Ethical Statement***
3. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **Yes**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes. We describe the process of hyperparameter tuning, the final parameters used and the data split ratio in section *Methodology*.**
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **No, because it is not relevant to our settings.**
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **Yes. We describe the computational resources in section *Methodology*.**
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes. We describe the model validation in sections *Methodology* and *Results*.**
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? **Yes. Since our models are regression models, we discuss the cost of prediction errors in section *Discussion*.**
4. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? **Yes. We utilize an annotated Wikipedia dataset to train the C model, referenced in section *Dataset* and multiple lexicons to identify linguistic features, referenced in section *Linguistic Features***
   (b) Did you mention the license of the assets? **Yes. We mention the license of the asset, when applicable — in Methods section**
   (c) Did you include any new assets in the supplemental material or as a URL? **Yes**
   (d) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes. Please refer to section Ethics and Broader Impact**

(e) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? **Yes. We ensure that the dataset is findable, accessible, inter-operable and re-usable**

(f) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? **Yes**

Appendix

Dataset

Subreddits and number of submissions: NoStupidQuestions (280,938), antiwork (99,514), politics (95,296), worldnews (94,272), WhitePeopleTwitter (69,887), gaming (62,756), AskUK (58,110), Conservative (57,217), Tinder (45,358), relationship_advice (43,314), movies (42,698), mildlyinfuriating (41,124), dating (33,207), mildlyinteresting (30,440), news (26,512), aww (24,391), RandomThoughts (23,174), atheism (20,440), personalfinance (19,448), popculturechat (19,182), AskAnAmerican (18,297), Bitcoin (17,113), Showerthoughts (16,302), Music (14,674), technology (14,063), relationships (13,818), nottheonion (12,852), CasualConversation (11,576), Frugal (11,206), socialskills (10,781), selfimprovement (10,351), changemyview (10,242), Jokes (10,159), 4chan (9,965), WTF (9,667), RoastMe (8,784), tifu (7,876), science (7,231), introvert (7,041), 3amjokes (5,723), getdisciplined (4,546), MaliciousCompliance (4,011), HistoryPorn (3,463), sports (3,348), UpliftingNews (3,203), nosleep (2,877), GetMotivated (2,497), Liberal (1,493), AskEconomics (773), OutOfTheLoop (697)

Choosing the Toxicity Detection Model

We choose the best toxicity detection model by comparing the model predictions with manual annotations. Three state-of-the-art models are considered: (1) toxic-BERT (Hanu and Unitary team 2020), (2) Perspective API (Lees et al. 2022), and (3) OpenAI omni-moderation-2024-09-26 (OpenAI 2024), the latest harmful content detection model provided by OpenAI as of October 2024.

These models assign scores to multiple categories. The toxicity score inferred by each model is the maximum among the categories. We compare the inferred toxicity scores against our ground truth annotations to assess model performance. To align with our definition of toxicity, we exclude the "self-harm/intent" and "violence" categories from the omni-moderation-2024-09-26 model categories.[7] Similarly, "obscene" and "profanity" are not considered among the categories of toxic-BERT[8] and Perspective API.[9]

Annotators are asked to label a post as toxic or non-toxic. Since one of our models is Perspective API, we provide Google's Perspective API[10] definition of toxicity to annotators, namely, "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion." Two annotators participate in the annotation. In the first round, they independently annotated a set of 100 posts. After this, they compare judgments, followed by a discussion to resolve discrepancies and refine the definition of toxicity. The second round annotation set contains 250 comments, where 150 comments are annotated independently, and 50 comments in common. We calculate the inter-rater agreement using the 50 overlapping comments, achieving a Cohen's Kappa of 0.81, indicating strong agreement. Finally, all the annotated labels, totaling 350 comments, are used to evaluate the performance of the toxicity detection models. The best performance is by omni-moderation-2024-09-26 model, with a ROC-AUC score of 0.91. Perspective API and toxicBERT achieves ROC-AUC scores of 0.87 and 0.84, respectively.

[7] platform.openai.com/docs/guides/moderation

[8] huggingface.co/unitary/toxic-bert

[9] developers.perspectiveapi.com/s/about-the-api-attributes-and-languages

[10] perspectiveapi.com/how-it-works/

Choosing the Threshold for Toxicity Model

To report the overlap between toxic and toxicity attracting submissions, we determine the optimal threshold for toxicity scores using our manually annotated dataset. Among thresholds ranging from 0 to 1, we find that 0.5 achieves the highest performance, yielding an F1-score of 0.68 (see Fig.6).

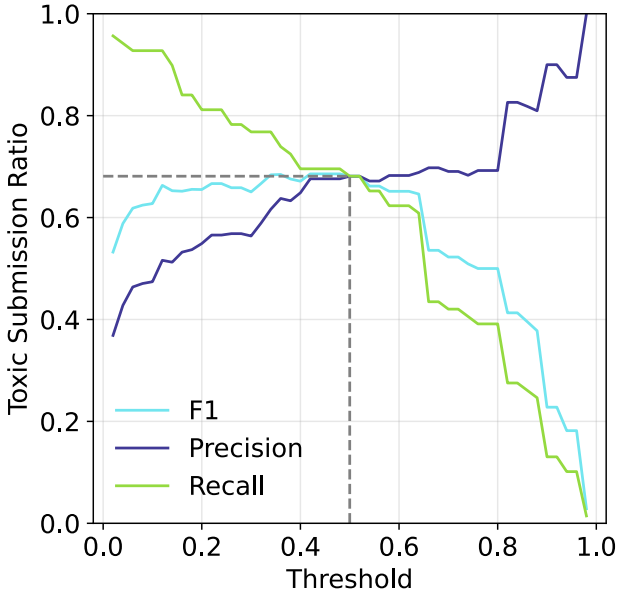

Figure 6: F1, Precision and Recall scores based on different thresholds for toxicity score.

Overlap of Toxic and Toxicity Attracting Posts

Fig.7 illustrates how varying the toxicity-attraction (TA) threshold affects the overlap between toxic (T) and toxicity-attracting (TA) submissions. The percentage of toxicity-attracting but not toxic posts for thresholds 0.2, 0.3, 0.4, and 0.5 are 27%, 8%, 7%, and 3%, respectively. The percentage of toxicity-attracting and toxic posts for the same thresholds are 5%, 2%, 2%, and 1%, respectively. The percentage of toxic but not toxicity-attracting posts are 3%, 6%, 6%, and 7%, respectively.

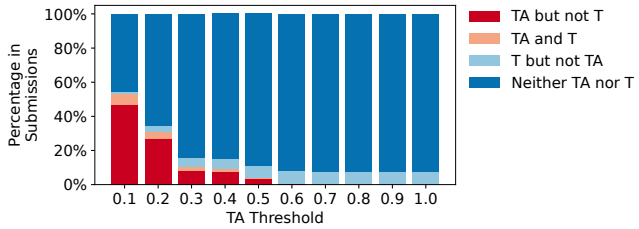

Figure 7: The proportion and overlap of toxic (T) and toxicity-attracting (TA) submissions across varying toxicity-attraction thresholds.

Within-topic Variation of C Scores

Table 3 provides examples of posts discussing the same issue but receiving different C scores. For instance, two submissions about immigration reform in Germany are scored 0.67 and 0.25, indicating that how a topic is framed or expressed plays a role in shaping its C score.

Extracting Topics with BERTopic

We use BERTopic to infer topics reflected in submissions. To do so, we first create a document-term matrix where we ignore terms that occur strictly lower than 10 times in the dataset. We then pass this matrix to BERTopic with number of topics set to "automatic" (nr_topics = "auto"), which automatically reduces the number of topics using HDBSCAN (McInnes et al. 2017). This method clusters the submissions into coherent topical groups and produces, for each cluster, a list of four representative keywords. In our analyses, we do not report any topic clusters supported by fewer than ten submissions. We shortened the keyword lists so the labels would fit cleanly in figures and tables.

Keyword Lists for Fig. 5. Dating & money ("she, gf, her, money"); Labor shortage ("labor, shortage, workers, nobody"); Curiosity ("wonder, interesting, why, wow"); Body talk ("boobs, breasts, nipples, breast"); Dating apps ("apps, dating, app, bumble"); Elon / speech ("elon, centrist, speech, conspiracies"); IQ debates ("iq, intelligence, smart, smarter"); Kamala Harris ("kamala, harris, biden, vp"); Don Lemon ("don, lemon, cnn, cnns"); Oscars slap ("smith, oscars, chris, slap"); Auto loans ("car, loan, vehicle, miles"); Books ("books, book, read, library"); Actors & roles ("actor, actors, role, roles"); 401k & debt ("401k, mortgage, roth, debt"); Sleep ("sleep, wake, asleep, bed"); Dental care ("teeth, dentist, tooth, dental"); Friendships ("friends, friendships, group, social"); Clothing ("wear, dress, wearing, pants"); Fashion events ("fashion, highlight, gala, archives"); COVID variant ("variant, cdc, infection, mild"); Gender issues ("gender, trans, gay, nonbinary"); Labor unions ("union, unions, strike, amazon"); Capitol riot ("capitol, police, riot, officers"); North Korea ("korea, north, kim, missile"); Antisemitism ("hitler, antisemitism, nazi, holocaust"); Dr Oz race ("oz, dr, pennsylvania, john"); MTG hearings ("greene, taylor, rep, committee"); Abortion ("abortion, roe, abortions, wade"); Mike Pence ("pence, mike, trump, capitol"); Inflation ("inflation, biden, joe, bidens"); Football & soccer ("football, sports, soccer, sport"); AI & jobs ("ai, artificial, automation, jobs"); Car insurance ("insurance, car, coverage, damage"); Student loans ("student, loan, loans, forgiveness"); Recycling ("trash, bin, plastic, recycling"); Cosmology ("universe, infinite, bang, expanding"); Marriage ("married, marriage, divorce, marry"); Marijuana ("marijuana, cannabis, weed, recreational"); Netflix ("netflix, streaming, subscription, tv"); White House press ("press, secretary, psaki, jen").

Linguistic Features

Gratitude Lexicon. thanks, contented, blessed, thank you, thankful for, grateful for, greatful for, my gratitude, i appreciate, made me smile, make me smile, i super appreciate, i deeply appreciate, i really appreciate, bless your soul, made my day, tysm, thx, shout out to.

Elaboration. We evaluated three operationalizations of elaboration: (1) total token count, (2) lexical-word count – number of content words (nouns, verbs, adjectives, adverbs), (3) Measure of Textual Lexical Diversity (MTLD) – a lexical-diversity metric in which higher scores reflect a wider range of vocabulary rather than repetitive wording, signaling the expression of a broader set of ideas (McCarthy and Jarvis 2010). All three metrics show highly similar patterns: each correlates significantly with TA score (all $p < 0.01$) for both controversial and non-controversial posts. Text length has the strongest association for non-controversial submissions ($\rho = -0.32$) and a comparable effect for controversial ones ($\rho = -0.24$). Lexical-word count follows the same trend ($\rho = -0.31$ non-controversial, $\rho = -0.18$ controversial). MTLD: $\rho = -0.30$ non-controversial, $\rho = -0.24$ controversial. Lexical-item count is strongly correlated with both alternative elaboration metrics: its Spearman correlation with MTLD is 0.80, and with total token count is an even tighter 0.98, both $p < 0.01$.

Regression Analysis. We also run a ordinary least squares regression analysis with the described linguistic features as independent variables, and TA score as the dependent variable. Before running the regression analysis, we assess multicollinearity by examining the variance inflation factor (VIF) for each variable. Since none of the features exhibits a VIF greater than 3, we include all variables in the analysis. The results are presented in Table 4. All coefficient signs are consistent with the correlations reported in the main text, except for the text length.

Topic	Submission	C Score	C %
Immigration in Germany	Germany is planning to reform its immigration policies. The U.S. should, too.	0.67	99
	Germany to "recruit workers from abroad" to ease airport chaos	0.25	63
Cannabis Usage	Adolescent cannabis use and later development of schizophrenia: An updated systematic review of six longitudinal studies finds "Both high- and low-frequency marijuana usage were associated with a significantly increased risk of schizophrenia."	0.83	99
	Can you get high of weed passively? The dentis just said, I cant smoke or drink for a week or so. Well, I cant smoke, but what if I just burn it and whiff it?	0.15	23
Taxation	Why does sales tax exist?	0.46	90
	Why is the tax not on the price tag? Ok so for reference and if this a stupid question, I'm from a country where the price tag is the final price. Tax included. But I've been to the US (recently I knew Canada has it too) the final price is added at the checkout. I find it kinda annoying because how do you know the exact price before checking out? Is there a way?	0.23	58
Racing Games	What Happened to Racing Games Lately? Anyone knows why there are very less racing games nowadays and most of them are not even that enjoyable. Just a decade ago there was huge catalog of them with the likes of Need For Speed, Midnight Club, Rumble Racing, Blur, Juiced, Driver, Grid, Burnout Series. Heck even crazy taxi not even a racing game was so good for its time. The most wonderful thing was all of these had local couch co-op and were really fun to play. Always wondered if these games had a open world mode. But now years later we have only Forza. Dont get me wrong, Forza also is a very good game but misses the joy of Street racing. Need for speed are released with years of gap and have no split screen and are very short and full of dlc and not that much exciting anymore. Whats up with the racing devs lately. How the heck was most old racing games were so good and not now.	0.18	38
	Looking for a new car racing game. Something inbetween a racing sim (project cars) and an arcade game (nfs). I couldn't really get into the new forza horizon, I have wreckfest but want a tarmacy game lol I'm on PC. Thanks.	0.12	14

Table 3: Examples of posts discussing same issues with differing C scores. C scores and their percentiles are shown, where higher percentiles indicate higher scores(controversiality). Topics are obtained using BERTopic. Topic names are abbreviated for clarity: Immigration in Germany (germany_immigration_immigrants_bolster), Cannabis Usage (marijuana_cannabis_weed_recreational), Racing Games (racing_driving_games_game), and Taxation (tax_price_sales_prices)

Variable	Coefficient	Std Err
intercept***	0.054	0.001
c_score***	0.344	0.001
question_ratio***	-0.067	0.001
gratitude_ratio	-0.017	0.022
proper_noun_ratio***	0.046	0.001
lexical_item_count	1e-06	1e-06
hedge_ratio***	0.017	0.003
polarity***	-0.018	1e-06

Table 4: Ordinary least squares regression results. ***($p < .01$).