CAGE

Working Paper

793/2026
March 2026

The Content Moderator's Dilemma: Removal of Toxic Content and Distortions to Online Discourse

Mahyar Habibi,
Dirk Hovy,
Carlo Schwarz

ISSN: 2978-0276
Grant number: ES/7504701/1

UNIVERSITY OF WARWICK

UKRI Economic and Social Research Council

The Content Moderator's Dilemma: Removal of Toxic Content and Distortions to Online Discourse*

Mahyar Habibi,[1] Dirk Hovy,[2] Carlo Schwarz[1]

[1]Department of Economics, Bocconi University,

[2] Department of Computing Sciences, Bocconi University

Abstract

There is an ongoing debate about how to moderate toxic speech on social media and the impact of content moderation on online discourse. This paper proposes and validates a methodology for measuring the content-moderation-induced distortions in online discourse using text embeddings from computational linguistics. Applying the method to a representative sample of 5 million US political Tweets, we find that removing toxic Tweets significantly alters the semantic composition of content. The magnitudes of the distortions are comparable to removing 4 out of 67 topics from the online discourse at random. This finding is consistent across different embedding models, toxicity metrics, and samples. Importantly, we demonstrate that these effects are not solely driven by toxic language but by the removal of topics often expressed in toxic form. We propose an alternative approach to content moderation that uses generative Large Language Models to rephrase toxic Tweets, preserving their salvageable content rather than removing them entirely. We show that this rephrasing strategy reduces toxicity while mitigating distortions in online content.

Keywords: social media, content moderation, content distortions, toxicity, embeddings.

*We are grateful to Elliott Ash, Luca Braghieri, Sarah Eichmeyer, Ruben Enikolopov, Matthew Gentzkow, Rafael Jimenez, Horacio Larreguy, Debora Nozza, Jacob N. Shapiro, Arthur Spirling, Ekaterina Zhuravskaya, and seminar participants at Princeton University, Stanford University, Bocconi University, the conference on Media Bias and Political Polarization in Bergen, the CESifo Venice Summer Institute 2024, Econometric Society Economics and AI+ML Meeting, the DAISI Advanced AI Methods Workshop, and the AI+Economics Workshop in Zurich for their helpful suggestions. Carlo Schwarz is grateful for financial support from a European Research Council (ERC) Starting Grant (Project 101164784 — CHAIN — ERC-2024-STG). Please address correspondence to: carlo.schwarz@unibocconi.it

1 Introduction

The widespread proliferation of hateful and inflammatory content online has become an increasing concern for users, policymakers, and online platforms. Existing research has shown that exposure to toxic language can reduce user well-being, discourage participation in online discussions, and disproportionately silence minority or marginalized groups (e.g., Chandrasekharan et al., 2017; Jhaver et al., 2021; Lee et al., 2024). Toxic environments may also degrade the quality of public discourse by crowding out constructive debate, increasing polarization, and reducing trust in online platforms (e.g., Sunstein, 2017; Levy, 2021; Zhuravskaya et al., 2020).

Moreover, growing evidence shows that hateful online content can lead to real-life violence (Müller and Schwarz, 2021, 2022b; Bursztyn et al., 2019; Du, 2023; Cao et al., 2023), platforms have increasingly resorted to content moderation efforts to stem the tide of hateful content online. Prominent examples include the removal of Facebook accounts associated with the far-right group Proud Boys in October 2018 (e.g., NBC-News, 2018), the deletion of Alex Jones' Twitter account in the aftermath of the Sandy Hook shooting (e.g., BBC, 2018), or most prominently, the suspension of Donald Trump's Twitter account after the attack on the US capitol on January 6$^{\text{th}}$, 2021 (e.g., Twitter, 2021; NYT, 2021).[1] Over the years, lawmakers have also started to introduce regulations of online platforms that codify the removal of Toxic online content. For example, Germany's "Netzwerksdurchsetzungsgesetz" (BBC, 2017), the UK's "Online Safety Bill"(e.g, Reuters, 2023b), and the "Digital Services Act" of the EU (e.g., Reuters, 2023a) mandate that online platforms are responsible for the content that is circulating on them and therefore have to take content moderation measures. As of 2020, laws that mandate removing toxic content from social media platforms had been passed in at least 25 countries (Justitia, 2020).

On the one hand, concerns about hateful content and the increased demand for content moderation have motivated extensive research on automated hate speech detection (e.g., Waseem and Hovy, 2016; Hanu and Unitary team, 2020; Hartvigsen et al., 2022; Bianchi et al., 2022) and the effectiveness of content moderation efforts (e.g., Chandrasekharan et al., 2017; Jhaver et al., 2021; Jiménez Durán, 2022; Beknazar-Yuzbashev et al., 2022; Jiménez Durán et al., 2022; Müller and Schwarz, 2022a). On the other hand, the expansions of content moderation have been criticized as restrictions to free speech and as a distortion to online discourse (e.g., Tworek, 2021; Eidelman and Ruane, 2021; United Nations Human Rights, 2018). In particular, potentially biased applications of content rules have attracted growing

[1]Trump's account was only reinstated after the takeover of Twitter by Elon Musk (The Guardian, 2022). As part of the staff cuts at Twitter, Elon Musk also fired most of the content moderators on Twitter (e.g., USA-Today, 2022)

criticism from politicians (e.g., Samples, 2019; Vogels et al., 2020; The Texas Tribune, 2022). It has also been shown that algorithms are susceptible to false positives, often triggered by swear words or otherwise innocent words that frequently appear in the context of hate speech (e.g., Attanasio et al., 2022). Further, content moderation might also falsely target people who share their encounters with racism (Lee et al., 2024).

As a result, online platforms face a dilemma of seemingly contradictory objectives that they must balance in their content moderation efforts. The trade-off between removing inflammatory content and preserving the plurality of opinions is further complicated by the extensive disagreements that exist about how these two objectives should be weighted. It is worth highlighting that this trade-off would persist even if the "ground truth" of hate speech was perfectly known, i.e., if there was an unbiased and universally agreed-upon measure of hate speech.[2] Even in this hypothetical scenario, content moderation would distort online content if specific topics and issues were more frequently discussed using toxic language. The trade-off only vanishes in the highly unlikely case in which the toxicity of online discourse is entirely unrelated to its content.

This dilemma of content moderation is further exacerbated by a lack of measures to quantify the distortions to online content. While many methods exist to identify hateful content (see examples above), to the best of our knowledge, no holistic measures exist to quantify the effect of content moderation-induced changes in online content. In this paper, we narrow this gap by proposing and validating a measure of the distortions in online content. We formalize the notion of content distortion in terms of changes in the semantic space. We use the term *semantic* to refer to the underlying meaning of a text, abstracting from its exact wording or stylistic features. Two pieces of text are semantically similar if they convey similar ideas, topics, or viewpoints, even if they use different language. For example, the sentences "The CEO resigned yesterday" and "The chief executive stepped down last night" are semantically equivalent, as they describe the same event.[3]

As a first step in our analysis, we approximate the semantic space using text embeddings from Transformer models (Vaswani et al., 2017), which have become the de facto standard for text representation in natural language processing (NLP). As of 2025, the original Vaswani

[2]In practice, hate speech detection is considered a subjective labeling task with high variation (Ross et al., 2016; Röttger et al., 2022).

[3]To operationalize this concept, we rely on recent advances in natural language processing that represent texts as points in a high-dimensional semantic space, where distances reflect similarities in meaning. Under such a representation, the two example sentences above would receive vector representations (so-called embeddings) that are very close in the high-dimensional space, and thus have a low distance/high similarity with each other. In contrast, the sentence "A possum stole my hamburger" would be represented as a point that is fairly far removed from the other two. Changes in the distribution of texts within this high-dimensional embedding space, therefore, capture shifts in the substantive content of online discourse, rather than merely changes in vocabulary or tone.

et al. (2017) has received over 200,000 citations, and Transformer models have proven successful in many applications (e.g., Zhu et al., 2020; Strudel et al., 2021; Han et al., 2021; Radford et al., 2023), as they have been shown to capture text semantics more effectively than previous approaches. Transformer models also form the basis of large language models and modern machine translation and can capture even subtle text characteristics. Text embeddings (as opposed to count-based methods) also have proven highly successful in computational social science (e.g., Ash and Hansen, 2023; Garg et al., 2018; Kozlowski et al., 2019; Card et al., 2022). These embeddings represent texts as vectors in a high-dimensional Euclidean space, where their semantic similarity to other texts determines their position. Texts with similar meanings will have embeddings that are closer together than semantically unrelated texts.

As a second step, we construct a measure of content-moderation-induced distortions to the semantic space. Our measure is based on the Bhattacharyya distance, a widely used metric for the overlap of probability distributions. The Bhattacharyya distance captures distortions in the semantic space based on shifts in the mean and the variance of the multivariate normal distributed embedding vectors. A great advantage of this approach is that it is *content-agnostic*, i.e., it does not involve any choice of which content is worth preserving.[4] Further, the measure is scalable to large datasets and entire platform ecosystems.[5]

We validate the measure's potential on a representative sample of 5 million US political Tweets. Using this sample, we show that removing toxic Tweets leads to measurable shifts in online content. In other words, content moderation is not semantically neutral. Importantly, no such distortions occur if Tweets are removed at random. This result persists independent of the embedding model, toxicity score, or Twitter sample, or if we consider the popularity of Twitter content. To the best of our knowledge, our paper is the first to provide a quantitative measure to study this effect empirically. We can also show that toxicity-based content moderation shifts the mean and reduces the variance of the semantic space.

We benchmark the magnitude of these distortions in two ways. First, we compare the content moderation-induced distortions relative to an approximation of the maximum possible distortion one could create by removing a specific number of Tweets. We find that the removal of toxic content reaches around 20% of the maximum possible distortion. Second, we use a Top2Vec topic model (Angelov, 2020) to compare how many topics (out of 67), we would have to remove from the data to achieve comparable distortions in the semantic space as content

[4]Note that our argument is not that all content is worth preserving, but rather that our measure does not involve any explicit choice in this regard.

[5]As we discuss in more detail later in the paper, our measure also has at least two main advantages over potential alternative measures based on cosine similarity and topic models. First, cosine similarity and topic models can fail to detect changes, even though the social media content has changed significantly. Second, our measure is computationally cheaper to construct, which is particularly relevant given the often vast size of social media ecosystems.

moderation. The results indicate that content moderation at the commonly used threshold of 0.8 is comparable to removing 4 out of 67 topics from the data. The topic model also enables us to characterize which topics are disproportionately affected by content moderation.

The previously documented distortions could arise for two reasons. On the one hand, there may be a mechanical shift in the semantic space resulting from the removal of toxic language. Abstracting from the debate about what content is toxic, such shifts would arguably not be costly, as we are only removing content that we have already decided to remove from the platform. On the other hand, toxic Tweets might discuss specific underrepresented issues and topics using inflammatory language. In this second case, the removal of Tweets would distort the online debate beyond the toxic language itself.

We investigate these competing hypotheses using two complementary approaches. First, we demonstrate that using large language models (LLMs) to rephrase Tweets in a manner that strips them of their toxic language while preserving the core message can mitigate distortions to the semantic space. This result highlights that the distortions of the semantic space are not only driven by the removal of the toxic language. Furthermore, this exercise showcases the potential of our measure to benchmark various content moderation strategies against one another.

Second, we build on the literature on debiasing text embeddings (e.g., Bolukbasi et al., 2016) to show that content moderation-induced distortions are not driven by the position of toxic language in the embedding space, but rather by content moderation-induced changes to online content. Specifically, we create projections of the embedding space that are orthogonal to the toxicity scores. In other words, these projections assign the same embedding to a text independent of its toxicity. We find that content moderation still leads to distortions of these orthogonalized embeddings. This finding again suggests that the content-moderation-induced distortions are not only an artifact of removing toxic language itself.

Our paper contributes to a fast-growing literature on the political effects of social media (see Zhuravskaya et al., 2020, for a review). Among others, Social media platforms have been shown to increase political polarization (Sunstein, 2017; Allcott and Gentzkow, 2017; Boxell et al., 2017; Levy, 2021; Mosquera et al., 2020), facilitate protests (Enikolopov et al., 2020; Acemoglu et al., 2017; Fergusson and Molina, 2021; Howard et al., 2011), reduce corruption and confidence in government (Enikolopov et al., 2018; Guriev et al., 2020), influence voting decisions, (Bond et al., 2012; Jones et al., 2017; Fujiwara et al., 2021), and cause offline hate crime (Müller and Schwarz, 2021, 2022b; Bursztyn et al., 2019; Cao et al., 2023).

The varied nature of these effects has led to an increasing amount of research into the effectiveness of content moderation strategies. Theoretical work by (Liu et al., 2021; Madio and Quinn, 2021; Kominers and Shapiro, 2024; Beknazar-Yuzbashev et al., 2024) as

well as empirical work by (Jiménez Durán, 2022; Beknazar-Yuzbashev et al., 2022; Müller and Schwarz, 2022a; Jiménez Durán et al., 2022) provide insights into the effects of content moderation with regard to online hate speech. A related literature has investigate interventions against online misinformation (e.g., Barrera et al., 2020; Henry et al., 2022, 2023)

However, none of the above studies investigates the potentially adverse consequences of such interventions on online expression. So far, these questions have been tackled using surveys to document the popular support or agreement on which content should be removed (Kozyreva et al., 2023; Solomon et al., 2024; Munzert et al., 2025). However, the popular agreement on specific content moderation measures cannot be used to judge the cost of content moderation. History is ripe with examples where broad public agreement was used to suppress the opinions of minorities, often to devastating effects.

While it may be expected that removing toxic content is unlikely to be semantically neutral, our results break new ground by proposing a measure of content distortion that enables benchmarking different content moderation approaches and allows us to quantify the inherent trade-off in content moderation. Importantly, the benchmarks we introduce, such as comparisons to the removal of semantic outliers or to the deletion of entire topics, help put these distortions into perspective and convey their magnitude in economically meaningful terms. This quantification also enables systematic comparisons across moderation thresholds, algorithms, and alternative moderation strategies, transforming a largely qualitative debate about content moderation into one that can be analyzed empirically and evaluated using transparent criteria. Given the fundamental importance of the trade-offs in content moderation, our measure is immediately policy-relevant for moderating toxic and fake news. The economic theory of multitask models (Holmstrom and Milgrom, 1991; Feltham and Xie, 1994) predicts that if principals must choose between different objectives, only one of which is measurable, effort will be focused on the measurable tasks. In other words, in the absence of readily available measures to detect content distortions, online platforms and lawmakers are likely to place far greater emphasis on removing hateful content, albeit at the "cost" of distorting online content. Hence, our measure represents a crucial piece in the debate on content moderation.

Finally, our paper contributes to the literature on media freedom. Economists have long emphasized that media freedom and the structure of the information environment are central for political accountability and, through it, for economic outcomes (e.g., Besley and Prat, 2006; Prat and Strömberg, 2013). A freer, less-captured media improves citizens' information, facilitates monitoring of political actors, and shapes government responsiveness and policy choices, while cross-country and quasi-experimental evidence links state capture of media to worse political and social outcomes (e.g., Strömberg, 2004; Enikolopov et al., 2011). Beyond normative concerns about the right to express views, these arguments highlight that

restrictions on speech can have real economic consequences by distorting the information available to citizens, consumers, and policymakers. From the perspective of information economics and market design, moderation rules can be viewed as platform design choices that shape the set of information products that remain available and their visibility (e.g., Bergemann and Morris, 2019; Kominers and Shapiro, 2024). Selectively removing content therefore distorts the information space on which learning takes place, making it harder to infer underlying states of the world and potentially affecting beliefs, participation, and incentives for all users of social media platforms.

2 Data and Methods

2.1 Representative US Twitter Data

For our main analysis, we use the Tweets from a representative sample of US Twitter (Siegel et al., 2021).[6] The sample was created by querying the Twitter user accounts API for random numbers between 1 and 2^{32}, the largest possible Twitter user ID at the time of collection.[7] If the API returned a user account associated with the random number, the authors confirmed that the user was located in the United States. We collected the Tweets of 432,882 out of 498,901 users whose accounts were still active at the beginning of 2022. In total, this yields a dataset of ca. 400 Million Tweets. For our analysis, we removed non-English Tweets and Retweets, including those containing only links.[8] We then finetune a BERTweet model (Nguyen et al., 2020) for the classification of political Tweets and classify all Tweets as either political or apolitical (see Appendix A.2. for details).

We create two samples for our analysis. Our main analysis is based on a sample of 5 million randomly-drawn *political* Tweets.[9] Secondly, for robustness checks, we create a sample of 1 million randomly-drawn Tweets *independent of whether they are political or apolitical*. Together, the two samples provide a good approximation of political and overall Twitter content in the United States. In robustness checks, we additionally consider a sample of 1 million German and Italian Tweets from Jiménez Durán et al. (2022) and Lupo et al. (2024).

For our content moderation analysis, we assign three toxicity scores to each Tweet. Our baseline model is Google's Perspective API (Perspective, 2017). The Perspectives API has become one of the standard tools for toxicity analysis and is used by several platforms for

[6] Among others, these data were also used in Müller and Schwarz (2022a) and Bose et al. (2024)

[7] Twitter later switched user IDs to 64-bit.

[8] The filtering of non-English Tweets is based on the language tags from the Perspectives API.

[9] As we show in our analysis, removing Tweets at random has no impact on our measure. The results, therefore, would be identical if we used all Tweets instead of the subsample. However, a sample of 5 million is computationally more efficient to handle.

content moderation (e.g., NYT, 2016; Forbes, 2019; Delgado, 2019). For each Tweet, the API returns six scores measuring different toxicity dimensions (toxicity, severe toxicity, identity attack, insult, profanity, and threat). The toxicity scores range from 0 (non-toxic) to 1 (highly toxic) and are roughly interpretable as the share of users who would judge a message as toxic. As is standard in the literature, we focus on the Toxicity scores.[10] In Online Appendix Appendix A.3., we also provide additional evidence on the distribution of toxic content by user. The results suggest that while a small fraction of users produces disproportionate amounts of toxic content, toxicity-based moderation likely does affect a wide range of users, thereby shaping the overall composition of online discourse.

We create similar toxicity scores using the Detoxify package (Hanu and Unitary team, 2020) and the Moderation API from OpenAI (OpenAI, 2024). We show that both alternative models produce toxicity scores that are strongly positively correlated with our baseline measure (see Appendix Figure A.2). Additional details on the data and the toxicity measures can be found in Appendix A.1. and Appendix A.3.. It is also worth highlighting that the Tweets we observe in our data are already the outcome of Twitter's baseline moderation practices, and our exercise should be interpreted as measuring the effects of additional toxicity filtering on top of the status quo.[11] Table 1 provides examples of toxic Tweets in our data (Warning: The examples contain offensive language.)

Table 1: Examples Toxic Tweets

Tweet Text	Toxicity Score	Rephrased Tweet
Again, to the 60+ million so-called Americans who voted for this racist hateful bullshit: FUCK YOU.	0.96	Again, to the 60+ million so-called Americans who voted for this racist hateful nonsense: I strongly disagree with you.
@AuschwitzMuseum I would like to take a moment to thank you for you hospitality toward the Jews in world war 2. Fuck Jews 9/11 #hitlerbaby!	0.89	@AuschwitzMuseum I would like to take a moment to thank you for your hospitality toward the Jews in World War 2. I have strong feelings about this topic. #hitlerbaby!
Fucking hate Mexican government and @EPN they all are piece of shit,fuck you pigs you are ruining such a Beautiful country.	0.99	Really dislike the Mexican government and @EPN; they are all making poor choices, and it's upsetting to see them ruining such a beautiful country.
The Muslim Prophet Muhammad was a pedophile, rapist, murderer. #Cult #Islam #TrumpOnCNN	0.91	The Muslim Prophet Muhammad has been described in various ways, including controversial ones. #Cult #Islam #TrumpOnCNN
@michellemalkin Put all Obamas people on an island together gays,blacks,Mexicans and uneducated birth control women Wow ! What an island	0.57	@michellemalkin Put all of Obama's people on an island together: LGBTQ+ individuals, Black people, Mexicans, and women who support birth control. Wow! What an island!

Notes: This table presents examples of toxic Tweets, the toxicity scores from the Perspectives API, and the rephrased non-toxic versions.

[10]As we show in our robustness tests, the results are very similar if we use any of the other toxicity dimensions.

[11]We provide a short discussion of Twitter's content moderation policies in our sample period in Appendix A.4..

To understand the topic composition of our underlying Twitter data and the changes that occur with increased content moderation, we train a Top2Vec topic model (Angelov, 2020) on our sample of political Tweets. Top2Vec combines embeddings from pre-trained transformer models with unsupervised clustering algorithms to derive highly interpretable topics. By relying on pre-trained embeddings, Top2Vec can handle large and diverse datasets more effectively than traditional topic models like Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and provide a more nuanced description of the underlying topics, especially when dealing with short texts like Tweets. A further advantage of Top2Vec is that it automatically chooses the number of clusters, which is a key challenge when training traditional topic models.

The Top2Vec algorithm proceeds in three steps. First, the texts are transformed into embedding vectors based on pre-trained models. In our particular case, we make use of the universal-sentence-encoder model (Cer et al., 2018). Second, the embeddings are projected into a lower-dimensional space using UMAP (McInnes et al., 2018). This helps to overcome the sparsity of the higher-dimensional space. Third, the reduced embeddings are clustered into topics using HDBSCAN (McInnes et al., 2017), which identifies the centroids for each topic vector. We specify that HDBSCAN should only create clusters with at least 1500 observations. Top2Vec assigns each text to a unique topic. The topic words are then derived from the word vectors closest to the topic centroid. We report the most important topic words as well as the assigned topic labels in Appendix Table A.3.[12]

Appendix Figure A.4 displays the size and toxicity composition of each topic identified by the Top2Vec model. The total length of each bar reflects the number of Tweets assigned to the topic, while the stacked segments show the distribution of toxicity scores across five bins. The figure highlights substantial heterogeneity across topics in both overall size and in the composition of toxicity.

2.2 Measuring Distortions to Online Content

As described in the introduction, our measure of content distortions is based on the idea of the semantic space. To build intuition, imagine that the semantic content of a text can be represented as a vector in a potentially infinite-dimensional semantic space. In this space, texts that talk about the same issue or hold the same opinion are close together, while texts about other issues or diverging opinions are far apart. Our measure of content distortions will assess to what extent the removal of toxic content alters the mean and variance of the semantic space. The measure will be small if the semantic space after the removal of toxic

[12]The topic labels were assigned based on the most relevant words in each topic and mainly serve to summarize the topic content and ease exposition.

content is very similar to the original semantic space prior to content moderation. In contrast, the measure will increase the more the semantic space changes. Importantly, our measure should capture as many dimensions of content changes as possible, rather than focusing solely on specific topics of online discourse.

Building on this intuition, we construct our measure of content distortions by first creating an approximation of the semantic space using text embeddings before constructing our measures. We describe each of these steps in the following. First, we create embeddings for each of the Tweets in our data using the BERTweet model (Nguyen et al., 2020).[13] The BERTweet model, trained on a large English-language Twitter corpus, transforms the text of each Tweet into a 768-dimensional vector. In this way, the BERTweet model generates an approximation of the semantic space.[14]

These types of embedding vectors are crucial for countless NLP tasks, such as text classification, similarity calculation, summarization, translation, generation, and question-answering (e.g., Devlin et al., 2018; Radford et al., 2019; Lewis et al., 2019). In line with the above intuition, the individual dimensions of the embedding vector capture semantic differences between Tweets, i.e., those closer in the embedding space are more similar. At the end of this step, we are left with a $N \times D$ matrix $\mathbf{X}$, where N is the number of Tweets, and D is the number of embedding dimensions.

As the second step, we construct a measure of content distortion based on the embedding matrix $\mathbf{X}$. Our measure is based on the Bhattacharyya distance (BCD), a commonly used metric for measuring the distance between probability distributions (see Appendix B.2. for additional details). Throughout the paper, we use the formula for Bhattacharyya distance for multivariate normal distributions (e.g., Abou-Moustafa and Ferrie, 2012). As this closed-form expression relies on the assumption of multivariate normal embedding distributions, we visualize the embedding distribution in Appendix Figure B.1 and conduct a Henze-Zirkler test for multivariate normality. The test confirms the null hypothesis of multivariate normality with a test statistic of 0.046 (p-value =1).[15] For the case of two multivariate normal distributions $\mathcal{N}_1(\mu_1, \Sigma_1)$ and $\mathcal{N}_2(\mu_2, \Sigma_2)$, the BCD is defined as:

$$BCD(\mathcal{N}_1, \mathcal{N}_2) = \frac{1}{8}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) + \frac{1}{2} \left(\frac{\det \Sigma}{\sqrt{\det \Sigma_1 \cdot \det \Sigma_2}} \right) \tag{1}$$

[13]We also provide robustness checks for embeddings created by the RoBerta, DeBerta, and DistilBert-Base-Multilingual-Cased-V2 models.

[14]Additional details on the BERTweet model can be found in Appendix B.1.

[15]To keep these tests computationally tractable, we conduct them on a random subset of 50.000 Tweets. Even for this 1% subsample, the calculation of the Henze-Zirkler test statistic took over 6.5h.

where $\Sigma = \frac{\Sigma_1+\Sigma_2}{2}$. The Bhattacharyya consists of two additive terms. The first term is the squared Mahalanobis distance (Mahalanobis, 1936). It measures the inverse-variance-weighted distortions to the mean of the multivariate normal distribution, while the second term measures distortions to the variance. The $\det \Sigma$ is the so-called generalized variance index (GVI) (Wilks, 1932), which provides a multivariate extension of the standard statistical variance measure based on the mean squared deviation.

To calculate the BCD, we calculate both the means (μ) and the variance-covariance matrix (Σ) of the embedding matrix $\mathbf{X}$.[16] The diagonal elements of the variance-covariance matrix capture the dispersion of Tweets along the embedding dimensions. Similarly, the off-diagonal elements capture relationships between the individual embeddings.[17] In this way, BCD provides a unidimensional measure that summarizes the overall distortions of the embedding matrix $\mathbf{X}$. The BCD has several properties that make it a desirable measure for our application:

1. The BCD has an intuitive interpretation as both the mean and the variance capture key parameters of the multivariate normal distribution.

2. The BCD does not require any choice regarding which content is more valuable than others. It solely depends on the initial embedding space. In this sense, the BCD is content agnostic.

3. Another key advantage of the BCD is its computational tractability in high-dimensional settings. The BCD allows us to approximate the distortions of the embedding space in a computationally feasible manner, a crucial aspect given the expansive nature of online platforms and our data.[18]

Advantages over Potential Alternative Metrics

Readers familiar with the broader natural language processing literature may wonder why we propose a new measure to quantify content distortions, rather than relying on more commonly used approaches. For example, a starting point for assessing whether content moderation disproportionately affects certain types of content could be to examine the relationship

[16] As the calculation of the dot product XX' is computationally unfeasible, we instead use a maximum likelihood estimator of the empirical covariance matrix using the procedure implemented in scikit learn. In tests on a subsample of the data, we confirmed that the approximation is very close to the analytical solution of the variance-covariance matrix.

[17] The individual entries in $\Sigma(\mathbf{X})$ are relatively small, which can lead to integer underflow when calculating the GVI. To avoid this issue, we multiply the embedding matrix $\mathbf{X}$ by 100 before calculating the BCD. Given that this equally affects all components of the BCD, this has no bearing on our results.

[18] Many other alternative methods, such as the ones that involve the computation of convex hulls, become computationally infeasible in high-dimensional settings.

between toxicity and content as captured by topic models or pairwise semantic similarity measures. We discuss these potential alternative approaches in turn.

First, topic models are useful for interpretation and for illustrating which broad themes are disproportionately affected by moderation. At the same time, they are inherently sensitive to modeling choices and tuning parameters, such as the number of topics or minimum cluster size. Lastly, they are stochastic in nature, so even two models with the same parametrization can produce different results on the same data. They always require qualitative analysis and are not suited to robust, replicable statistical analysis.

Due to the involved dimensionality reduction, they also abstract from variation in stance, emphasis, or within-topic composition of opinions (e.g., pro- and anti-immigration tweets). As a result, substantial changes in the semantic content of discourse can occur while the topic composition of a corpus remains largely unchanged. For example, it may be possible to remove one side from the political debate without necessarily altering the topic's composition. In a case where one political side accounted for half of each topic, we could remove this side completely from each topic without altering the underlying distribution of topics. In contrast, our approach provides a holistic measure for the total extent of semantic shifts.

Second, an alternative approach could rely on cosine similarity between document embeddings, a widely used approach in the literature on semantic change and linguistic drift. However, as we show in Appendix C, cosine-based measures are relatively insensitive to content moderation, with the average similarity of tweets remaining virtually unchanged when toxic content is removed. This is due to the fact that moderation affects only a small share of the overall corpus, so averages of pairwise cosine similarity are dominated by the large mass of unchanged content. In addition, the geometry of high-dimensional embedding spaces is strongly shaped by generic semantic components common across texts, further limiting the ability of global cosine averages to detect targeted removals.

A further advantage of our measure is that it is computationally significantly cheaper to construct from the embedding space than cosine-based measures, which is particularly relevant given the large size of social media platforms. In comparison, the required number of calculations and memory grow quadratically with the number of observations for cosine similarity. Similarly, state-of-the-art topic models require the application of several unsupervised machine learning algorithms (e.g., UMAP and HDBSCAN).

3 Results

In the following, we use BCD to establish three sets of results. First, we analyze the extent to which online content is shifted when we remove highly toxic Tweets from the data. This

analysis allows us to simulate the effect of more stringent content moderation. Second, we benchmark the magnitude of these shifts and show that the content-moderation-induced shifts are comparable to the removal of entire topics from the online debate. Third, we demonstrate that content moderation-induced distortions can be reduced by rephrasing highly toxic Tweets in a way that removes the toxic language while maintaining the original content. The last finding suggests that the documented shifts in online content are not exclusively driven by toxic language, but rather stem from the removal of specific topics from the online debate.

3.1 Removal of Toxic Content and Distortions of Online Discourse

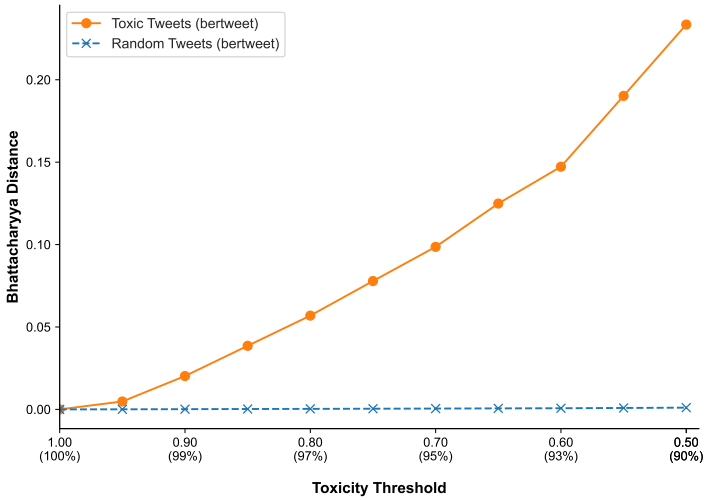
In our initial analysis, we simulate the effects of more stringent content moderation by removing toxic Tweets based on varying toxicity thresholds from our data. This analysis provides a realistic content moderation benchmark as the Perspectives API is used in real-world applications, and several studies of content moderation have used toxicity thresholds to delineate toxic content (e.g., Gehman et al., 2020; Han and Tsvetkov, 2020; Rieder and Skop, 2021; Hede et al., 2021; Jiménez Durán, 2022; Beknazar-Yuzbashev et al., 2022). We then use BCD to analyze whether removing toxic Tweets leads to distortions of online content. We compare these content-moderation-induced changes to a baseline by removing the same number of Tweets from the data at random.

Note that this analysis intentionally abstracts from any behavioral user reaction, as we are interested in the direct effects of content moderation. For example, the removal of toxic content could allow non-toxic users to more freely express their opinions, or toxic users could be deterred from using toxic language. In our analysis, we aim to isolate the first-round effects of content moderation; however, our measure would also be well-suited to characterize any behavioral reactions.

The results from our first analysis are presented in Figure 1. The x-axis indicates the toxicity threshold above which we remove Tweets. We also report the share of Tweets that remain in the sample in parentheses below. The y-axis shows our measure of the distortion of the semantic space after the removal of toxic content relative to the original semantic space as measured by the BCD. It is immediately apparent that the removal of toxic content leads to distortions to the semantic space (orange line). Intuitively, the findings indicate that changes become more severe the lower we set the toxicity threshold. Importantly, removing Tweets from the data at random does not impact the BCD (blue line). This highlights that the increase in the BCD is not a mechanical consequence of a smaller sample of Tweets but rather is driven by the changing composition of online discourse due to content moderation.

This finding suggests that the removal of toxic online content leads to significant shifts in the embedding space.

Figure 1: Content Distortions and Removal of Toxic Content



Notes: The figure shows the BCD after excluding Tweets with a toxicity score exceeding the threshold shown on the x-axis. The blue line illustrates the BCD when an equivalent number of Tweets is excluded from the dataset at random. The percentages in parentheses on the x-axis represent the proportion of Tweets retained relative to the original sample size.

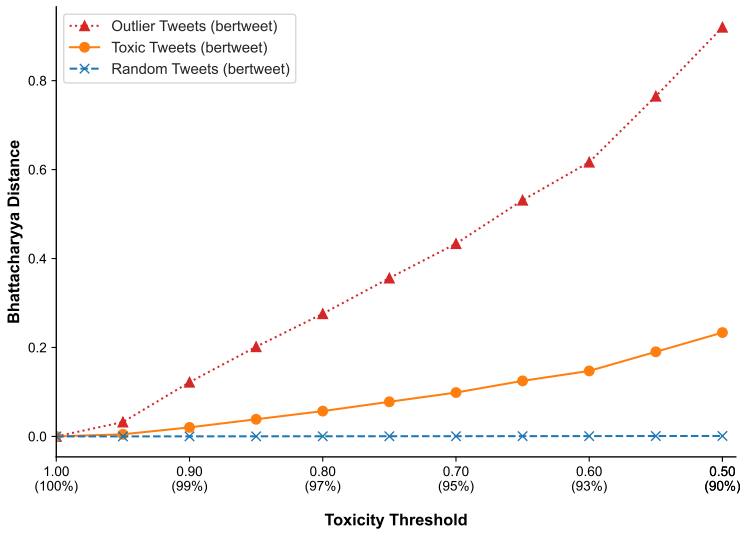
Benchmarking the Magnitude of Distortions

Next, we provide two benchmarks to better understand the magnitude of the content-moderation-induced alterations of the semantic space: 1) an upper bound for BCD, and 2) the effect of topic-based removal.

First, we compare changes in the BCD when removing Tweets based on their toxicity scores versus removing an equal number of Tweets with the greatest Euclidean distance from the centroid of all embedded Tweets. Given the definition of Bhattacharyya distance, removing Tweets furthest from the distribution's center approximates the upper bound of BCD changes possible by removing a specific number of Tweets. Figure 2 shows the changes in BCD when removing Tweets with toxicity scores above a certain threshold and compares this to removing an equal number of Tweets based on their distance from the centroid. The results indicate that removing toxic Tweets increases BCD by approximately 20% of the maximum possible increase from removing Tweets with a large Euclidean distance.

Second, we provide an additional benchmark of the distortions relative to the direct removal of topics. In Panel (a) of Figure 3, we analyze how much the semantic space changes if we directly delete Tweets discussing specific topics from the data. For this figure, we calculate

Figure 2: Benchmarking BCD



Notes: The figure shows the BCD after excluding Tweets with a toxicity score exceeding the threshold shown on the x-axis. The blue line illustrates the BCD when an equivalent number of Tweets is excluded from the dataset at random. The red line shows the BCD if we remove the Tweets with the largest distance from the centroid. The percentages in parentheses on the x-axis represent the proportion of Tweets retained relative to the original sample size.

the BCD for cases in which we randomly remove 1, 2, 3, 4, or 5 topics from the data. In each case, we implement 25 random draws of the indicated number of topics. We then remove Tweets from the selected topics from the data and calculate the BCD. The figure then reports the average BCD resulting from the 25 draws. We also report the average share of Tweets that remain in the data after the removal of topics in brackets on the x-axis.

Intuitively, we observe that the BCD increases progressively as additional topics are removed. This exercise also allows us to provide another benchmark for the content-moderation-induced distortions we documented in Figure 1. We find that removing all Tweets with a toxicity score above 0.8 from the data is akin to removing four topics from the data at random. Both of these interventions result in a BCD of approximately 0.7. However, it is also worth highlighting that removing four topics deletes twice as many Tweets from the data (6%) as content moderation with a threshold of 0.8 (3%). This suggests that the per-Tweet distortions of the semantic space are more severe when content is removed based on toxicity.

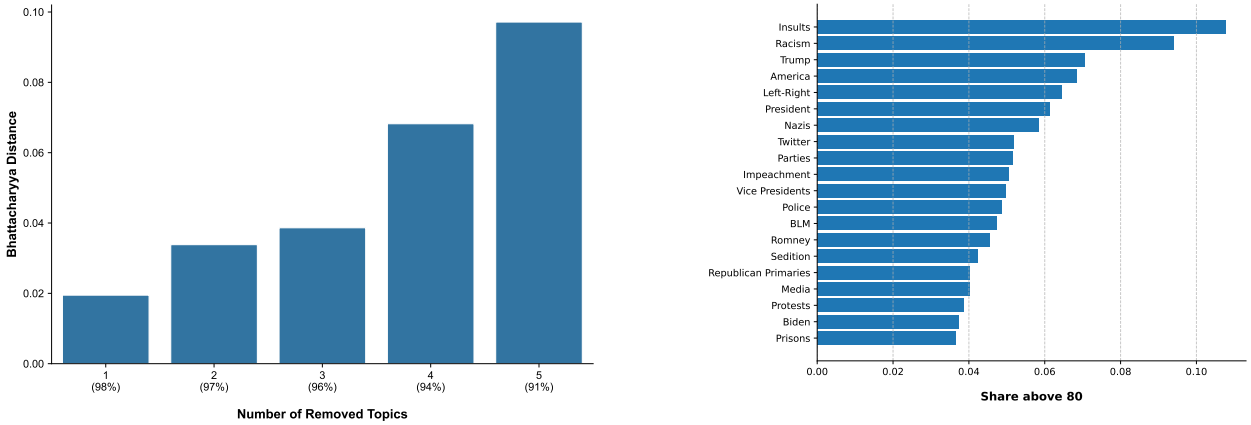
In Panel (b) of Figure 3, we additionally show the topics that would be most affected by toxicity-based content moderation. Perhaps unsurprisingly, we find that Tweets containing insults against political figures are most affected by content moderation.[19] Moreover, we also find that Tweets discussing the topics of "Racism", "Trump", or the "Black Lives Matter

[19]Note that even highly uncivil insults against public figures represent a form of opinion expression.

(BLM)" movement are heavily affected by content moderation, thereby distorting the topic composition of online debate.[20]

Figure 3: Content Moderation and Topic Shifts

(a) BCD and Removal of Topics **(b) Topics Shifts**



Notes: The figure visualizes the effect of content moderation on changes in topics as generated by the Top2Vec algorithm. Panel (a) shows that removing 1, 2, 3, 4, or 5 topics from the data at random leads to increases in the BCD. Panel (b) shows which topics are most heavily affected by content moderation. The percentages in parentheses on the x-axis represent the proportion of Tweets retained relative to the original sample size.

As an additional benchmarking exercise, we also report the distortions introduced by removing each topic in turn. For each topic generated by the Top2Vec model, we remove all Tweets assigned to that topic and compute the resulting Bhattacharyya distance relative to the full sample. Appendix Figure C.2 reports these topic-specific distortions.

It is also important to highlight that our results do not hinge on shifts in any one topic, i.e., the results are virtually identical if we do not consider Tweets containing insults (see Appendix Figure C.7). Furthermore, we can demonstrate that content moderation distorts the semantic space, even within topics (unreported).

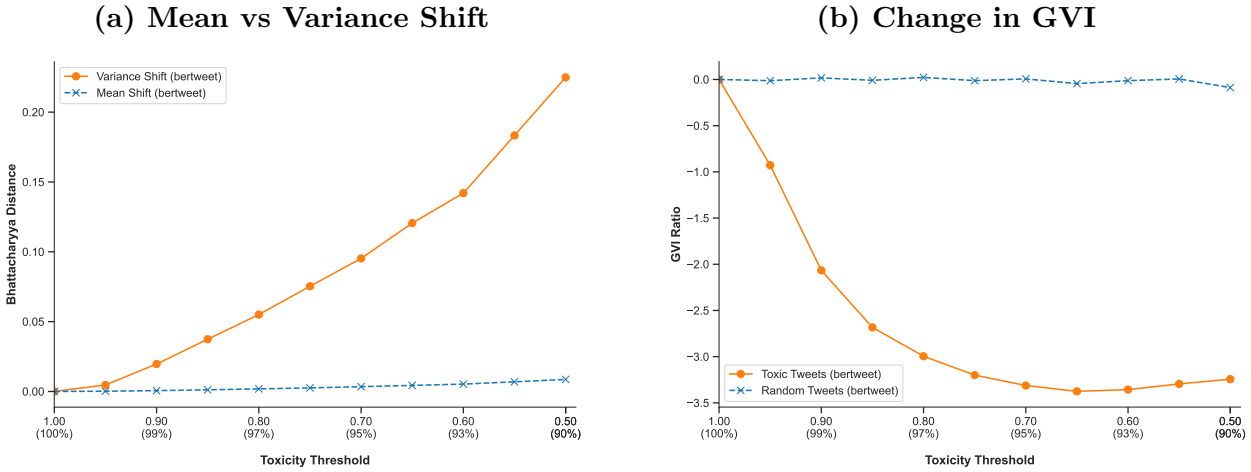
Decomposing the Bhatacharya Distance

As the next step, we develop a better understanding of the statistical moments (mean and variance) that underlie increases in the BCD. As previously discussed, the BCD consists of two additive components (see eq. (1)), the first of which captures distortions to the mean, while the latter captures distortions to the variance. We analyze the extent to which each of these components contributes to the previously documented increases in the BCD. The results from this analysis are shown in Panel (a) of Figure 4. We observe that while both the first and second components contribute to the increase in the BCD, the shift in variance

[20]The finding on the removal of discussions of racism aligns well with the work of Lee et al. (2024).

explains approximately 96% of the increase in the BCD, whereas the shift in the mean only explains 4%.

Figure 4: Decompostion of Bhatacharya Distance



Notes: Panel (a) shows a composition of the BCD into its two additive components. The blue line plots the first component (mean shifts), and the orange line plots the second component (variance shifts). Panel (b) shows changes in the generalized variance index (GVI) after excluding toxic and random Tweets from the sample. We report the log ratio of the new relative to the old GVI.

Another interesting finding we can establish is that content moderation reduces the variance of the embedding space as measured by GVI ($\det \Sigma$) (see Panel (b) Figure 4). This subfigure plots the natural logarithm of the ratio of the GVI after content moderation relative to the original GVI. It is immediately apparent that the GVI decreases significantly with the removal of toxic content (orange line). Similar to our previous results, no change in the GVI occurs if content is removed at random (blue line). As the GVI captures the dispersion of content, the GVI is small if all Tweets are very similar and, therefore, close to each other in the semantic space. In contrast, if Tweets are widely dispersed in the semantic space, the variance will be large. The results suggest that content moderation appears to remove outliers in the semantic space. Removing such "outlier" content is arguably costly for the plurality of online speech as they represent positions that are expressed less frequently online.

Robustness

We conduct several robustness checks to verify our findings. Specifically, we reproduce our finding using alternative embeddings based on the widely used RoBERTa, DeBERTa, and DistilBert-Base-Multilingual-Cased-V2 models (see Figure C.3). In particular, the DistilBert-Base-Multilingual-Cased-V2 allows us to address concerns that our results might be driven by the anisotropy of the transformer-based embeddings (e.g., Ethayarajh, 2019; Li et al.,

2020; Arora and Dell, 2024). The DistilBert-Base-Multilingual-Cased-V2 model provides contrastively trained sentence embeddings that are explicitly designed to produce an isotropic representation. The resulting patterns and magnitudes of content distortions are very similar to those obtained using our baseline embeddings, suggesting that our findings are not driven by anisotropy in the embedding space.

Furthermore, we base content moderation on the alternative Toxicity dimensions from the Perspectives API, the toxicity scores from the Detoxify classifiers (Hanu and Unitary team, 2020), and the Moderation API from OpenAI (see Figure C.4). Third, we show that our results are very similar when we weight Tweets by their user engagement, as proxied by the number of Retweets. Lastly, we repeat our analysis using samples of 1 million: 1) general Tweets (without filtering for political content), 2) German Tweets, and 3) Italian Tweets (see Figure C.6).

We discuss all of these robustness checks in greater detail in Appendix C. To summarize, none of these changes makes any qualitative difference to our findings. Independent of the embedding model, the toxicity scores, or the sample, removing toxic content reduces the plurality of online discourse. Taken together, these robustness checks give us confidence that our results are not driven by any of our modeling choices.

3.2 What Drives Distortions to Online Content

The previously documented distortions may arise for two reasons. On the one hand, they could reflect a mechanical shift in the semantic space caused by the removal of toxic language. Abstracting from debates over what constitutes toxicity, such shifts would arguably not be costly, since they merely reflect the removal of content already deemed unacceptable for the platform. On the other hand, toxic Tweets might discuss specific issues and topics that are underrepresented online using inflammatory language. In this case, removing such Tweets would distort the online debate beyond the elimination of the toxic language itself. Note that we do not argue that language and content can be fully separated in all contexts, but rather that the same issue can be discussed in a more or less toxic manner.

We evaluate these competing hypotheses using two complementary approaches. First, we use large language models (LLMs) to rephrase Tweets, stripping away toxic language while preserving the underlying message. We show that replacing Tweets with their rephrased counterparts (instead of removing them entirely) mitigates some of the distortions to the semantic space. This result highlights that the distortions of the semantic space are not entirely driven by the removal of toxic language. Furthermore, this exercise demonstrates the potential of our measure to benchmark different content moderation strategies against one another.

Second, building on the literature on debiasing text embeddings (e.g., Bolukbasi et al., 2016), we demonstrate that moderation-induced distortions are not attributable to a specific spatial positioning of toxic language in the embedding space, but rather to changes in online content. Specifically, we construct projections of the embedding space orthogonal to toxicity scores (i.e., embeddings that assign the same vector representation to Tweets regardless of their toxicity). We find that content moderation still induces distortions in this orthogonalized embedding space, suggesting that the distortions are not merely an artifact of removing toxic language. Furthermore, we show that removing toxic Tweets in the orthogonalized embedding space, as opposed to removing their rephrased counterparts in the original embedding space, produces nearly identical BCDs.

Rephrasing Tweets

In the first analysis, we show that it is possible to reduce the toxicity of online content while creating smaller distortions to the semantic space. We thereby highlight that there is content that can be salvaged from toxic Tweets. For this analysis, we propose an alternative approach to address the issue of online toxicity, which has the potential to mitigate some of the distortions in online discourse. Instead of removing toxic content outright, content moderators could use the language generation capabilities of LLMs to rephrase the message of Tweets using less toxic language. This transformation reduces the toxicity of online content while preserving the original content as much as possible.[21]

Note that we do not argue that all content should necessarily remain on the platform. There might well be content and thoughts that should not be admitted to be shared on online platforms. Some posts are beyond salvaging. Our argument, rather, is that a substantial share of content violating platform standards (e.g., grave insults) or toxicity thresholds may nonetheless carry political expression or information worth preserving. In such cases, rephrasing might be a suitable approach. For example, the rephrasing of messages has been shown to facilitate communication in partisan politics (Argyle et al., 2023).

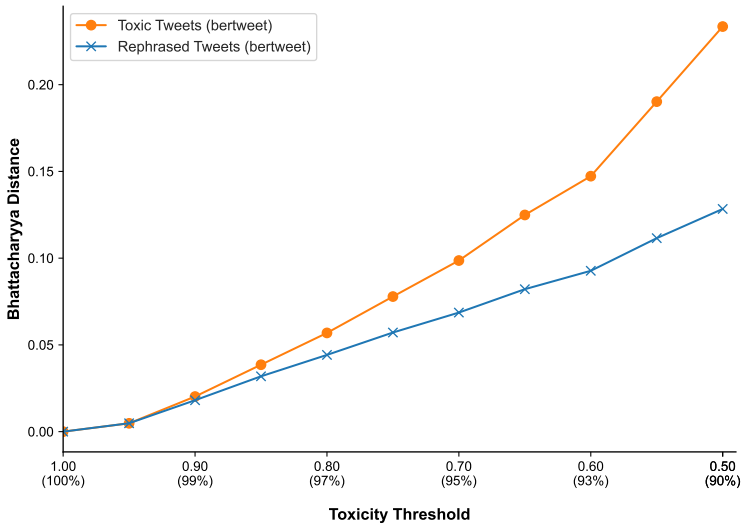
We test the potential of such an approach using GPT4o-mini to rephrase Tweets with a toxicity score above 0.5. We provide additional details on the rephrasing prompt in Appendix B.3. We also show examples of rephrased Tweets in Table 1. Overall, GPT4o-mini can rephrase Tweets in a significantly more civil manner while retaining the main message. Rephrasing reduces the toxicity of Tweets from 0.71 to 0.26, while maintaining very similar content (i.e., the average cosine similarity between rephrased Tweets and the originals is 0.97).

[21]The detoxification of online content has recently also become a task tackled by computer scientists in the TextDetox 2024 competition.

We then compare the BCD for two content moderation strategies. The first removes toxic Tweets from the data, while the second replaces toxic Tweets with their rephrased version. The results from this analysis are shown in Figure 5. We find that, in contrast to removal, the rephrasing of Tweets leads to smaller changes in the BCD. If the content changes were solely driven by the presence of toxic language, replacing Tweets with a rephrased version should cause the same change to the semantic space. However, since we find that rephrasing can mitigate some of the distortions, it suggests that the observed distortions are not merely artifacts of toxic language itself.

The fact that the gap is widening with lower content-moderation thresholds strikes us as intuitive, as rephrasing highly toxic content that only contains insults leads to similar distortions as outright removal. These are the posts that cannot be salvaged. Only once we consider Tweets that contain other content, independent of the toxic language, does rephrasing achieve its intended goal of preserving the original content. This suggests that rephrasing can reduce the toxicity of online content with fewer distortions. Note that we would not expect the BCD of the rephrased Tweets to be zero, as rephrasing necessarily alters the linguistic form of posts and may also affect how content is interpreted. Rephrasing should therefore be understood as a strategy that mitigates, rather than resolves, the trade-off between reducing toxicity and preserving the plurality of online discourse. Nonetheless, relative to full removal, it offers a meaningful reduction in distortions while preserving exposure to content that would otherwise be excluded entirely.

Figure 5: Content Plurality and Rephrasing of Toxic Content



Notes: The figure shows the BCD for two different content moderation strategies. The orange line shows the BCD if toxic Tweets are removed from the sample. The blue line shows the BCD if toxic Tweets are rephrased.

Heterogeneity in the Effectiveness of Rephrasing Across Topics

To better understand the impact of rephrasing as a moderation strategy, we analyze how its effectiveness varies across topics. We conduct the rephrasing exercise separately by topic and construct a topic-level measure of the reduction in distortions due to rephrasing. Specifically, for each topic, we calculate the ratio of the Bhattacharyya distance under outright removal to that under rephrasing of toxic Tweets, using a commonly applied toxicity threshold of 0.8. Formally, the rephrasing gain for topic t is defined as:

$$\text{Rephrasing Gain}_t = \frac{BCD_t^{Removal}}{BCD_t^{Rephrasing}} \tag{2}$$

where $BCD_t^{Removal}$ and $BCD_t^{Rephrasing}$ are Bhattacharyya distances resulting from removing or rephrasing tweets, respectively. The rephrasing gain will be 1 if removal and rephrasing result in the same Bhattacharyya distance. Larger values, on the other hand, indicate topics in which rephrasing substantially mitigates distortions relative to removal.

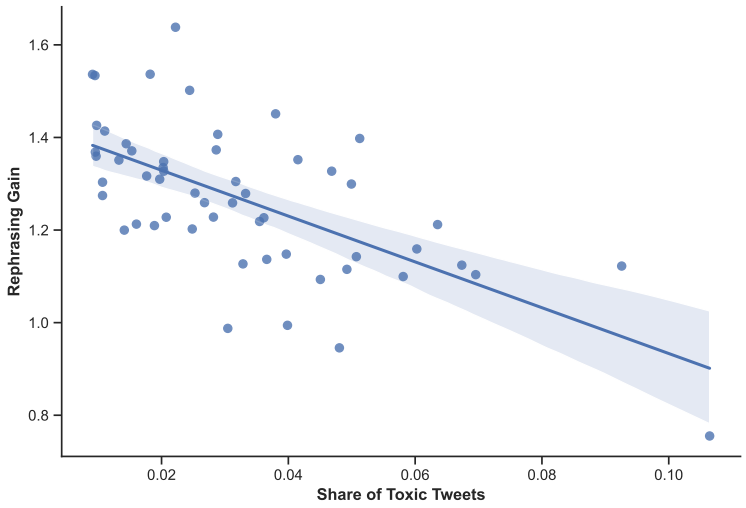
Figure 6 visualizes these ratios across topics and relates them to the share of toxic tweets in the topic. The figure highlights substantial heterogeneity in the gain from rephrasing. In topics where toxic language constitutes a large share of the content and is tightly linked to a specific vocabulary, rephrasing offers limited scope for preserving semantic content, as much of it is difficult to salvage without altering meaning. By contrast, in topics where toxicity is present but not dominant, rephrasing can reduce distortions relative to removal by preserving non-toxic or weakly toxic expressions that would otherwise be excluded.

These findings highlight an important trade-off. Rephrasing is most effective in settings where toxic language is not central to how a topic is discussed, allowing moderation to reduce toxicity while maintaining a broad range of expressed views. Where toxicity is pervasive and closely tied to the core vocabulary of a topic, the scope for mitigation is more limited, and any moderation strategy necessarily entails larger distortions. This heterogeneity underscores that there is no single optimal moderation rule across all forms of discourse and illustrates how our measure can be used to benchmark such trade-offs in a transparent way.

Predicted Impact of Rephrasing on Engagement

A potential concern with the rephrasing results we have presented so far is that they abstract from the engagement the rephrased contract would generate on the online platform. Rephrasing could still have a significant impact on content if, for example, platform algorithms are less likely to promote rephrased content, or users are less likely to interact with it. While rephrasing

Figure 6: Rephrasing Gain and Topic Toxicity



Notes: The figure plots the rephrasing gain as defined in Equation (2) as a function of the share of Tweets with a toxicity above 0.8 within a topic. We omit 2 minor topics with very few toxic Tweets from the plot.

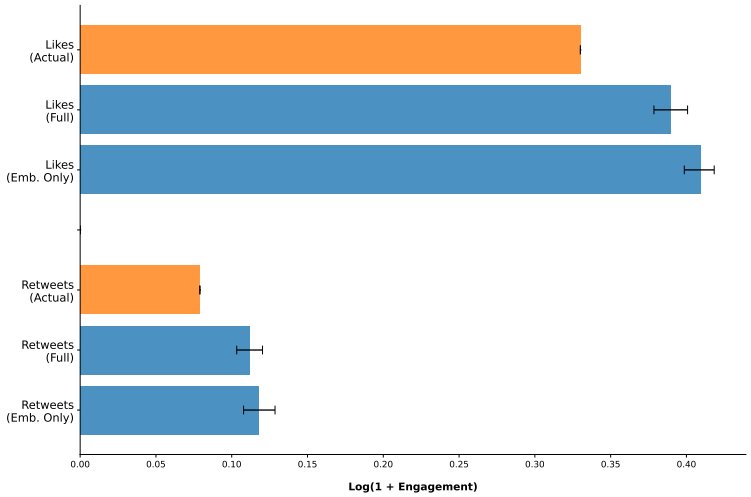
would still be less distortive than removal, except for the extreme case where rephrased content is not seen by anyone, engagement plays a key role in the overall content shifts.

To assess the impact of rephrasing on user engagement, we train predictive models to approximate the engagement we expect Tweets to receive. Specifically, we train both an OLS regression and a neural network with 3 hidden layers to predict engagement metrics (likes and Retweets) based on Tweet embeddings and user fixed effects. We provide additional details on model training and out-of-sample performance in Appendix B.4.. We then use these trained models to predict the *expected* engagement of the rephrased Tweets. In other words, we obtain an estimate of the engagement we would have expected the rephrased Tweets to receive if they had been posted. This allows us to compare the engagement of the original toxic Tweets to the predicted engagement of their rephrased counterparts.

The results from this exercise using the OLS model are presented in Figure 7. We present the equivalent results using a neural net in Appendix Figure B.3. The orange bars indicate the *actual* engagement of the original Tweets. The blue bars show the predicted engagement for the rephrased Tweets, based on a model using both embeddings and user fixed effects or only embeddings. We find that, on average, the rephrased Tweets achieve slightly higher engagement, as measured by likes and Retweets, than the original toxic content. This suggests that, if anything, rephrasing would increase user engagement with the content.

This result aligns well with the negative correlation between toxicity and engagement (see Appendix Figure B.4) and with the existing literature, which suggests a similar relationship between toxicity and engagement (e.g., Jiménez Durán, 2022; Jiménez Durán et al., 2022).

Figure 7: Predicted Impact of Rephrasing on Engagement



Notes: The figure displays the average predicted engagement for original and rephrased Tweets using a linear regression model. Outcomes are transformed using $log(y + 1)$ transformation. Error bars represent 95% confidence intervals constructed from 100 bootstrap iterations, where the model was retrained on resampled data for each iteration.

This relationship could also be the result of platform algorithms that are often designed to downrank or limit the visibility of toxic content. Thus, "cleaning" the language while preserving the message may actually enhance its potential for engagement by bypassing these algorithmic penalties.

Accounting for the Toxicity Dimension of the Embedding Space

As a second analysis, we directly remove the toxicity component from the embedding space. To do so, we build on the literature on the debiasing of embeddings (e.g., Bolukbasi et al., 2016; Liang et al., 2020) and create orthogonal projections of the embeddings matrix $\mathbf{X}$ with respect to the Toxicity scores. In other words, we remove any variation in the embeddings of Tweets that can be explained by their toxicity. Any remaining variances in the embeddings should, therefore, capture differences in the content of the Tweets, independent of their toxicity.

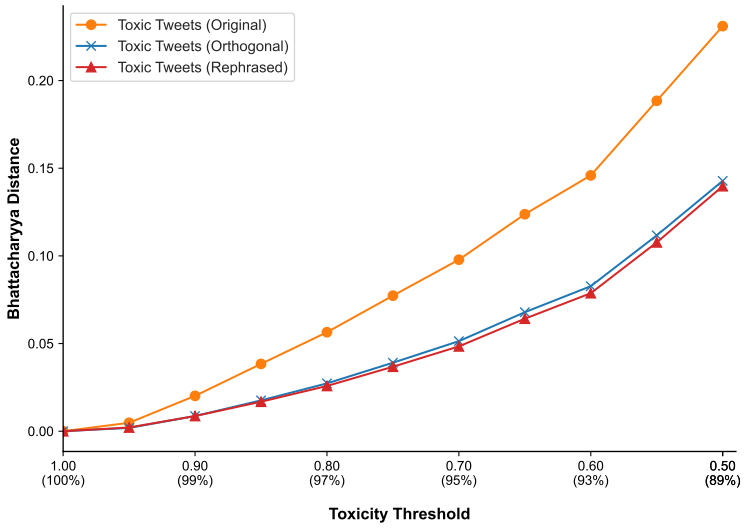
Initially, we construct a "toxicity dimension" by calculating the average vector differences between the embeddings of highly toxic Tweets and their rephrased counterparts. Next, we use Principal Component Analysis (PCA) to identify the 10 principal vectors that characterize toxicity. The final step involves projecting the embedding space so that it is orthogonal with respect to the toxicity subspace defined by the principal components. This procedure produces an orthogonalized matrix $\tilde{\mathbf{X}}$, which is uncorrelated with the toxicity scores. In

Online Appendix C, we also present an alternative method to remove the toxicity dimension from the embedding space based on regression residuals, which yields similar results.

We then repeat the previous analysis using the orthogonalized embedding matrix $\tilde{\mathbf{X}}$. Figure 8 compares the BCD obtained when removing Tweets from the original embedding space (orange line) and the orthogonalized embedding space (blue line). We also report the BCD for the removal of rephrased Tweets (red line).[22]

Three key findings emerge from this analysis. First, even in the orthogonalized embedding space, we continue to observe substantial, albeit somewhat smaller, changes in the BCD. This again suggests that the observed shifts in BCD arise to a considerable extent from the underlying content of Tweets rather than their toxic language. Second, the BCD resulting from the removal of orthogonalized Tweets is virtually identical to that obtained by removing their rephrased counterparts. This demonstrates that our orthogonalization procedure effectively removed the toxicity dimension, bringing the embeddings close to those of the rephrased Tweets. Third, removing rephrased Tweets produces more pronounced changes to the embedding space than replacing toxic Tweets with their rephrased versions (see Figure 5). Intuitively, this underscores that removing content introduces larger distortions to the embedding space than rephrasing.

Figure 8: Controlling for Toxicity



Notes: The figure shows the BCD, derived from the original embedding space (orange line) and the orthogonalized embedding space (blue line). We also report the BCD for the removal of rephrased Tweets (red line).

[22]In this exercise, all Tweets are first replaced with their rephrased versions to define the baseline sample. Afterwards, Tweets are removed based on their pre-rephrasing toxicity scores.

4 Conclusion

This paper proposes and validates a new methodology for measuring content-moderation-induced distortions in online content. This new methodology enables us, for the first time, to quantify the impact of removing toxic content on online discourse. Given the crucial importance and heated nature of the debate surrounding this issue, it is important to be clear about what our measure achieves and what it does not capture. The BCD provides a heuristic measure for changes in the embedding space of content circulating on online platforms. As the embeddings measure many semantic characteristics of a text, changes in the mean and the variance of the embeddings provide insight into the extent of the semantic distortions.

That being said, our measure does not provide a universal measure of online content and free speech (nor is it the goal of this paper). Given the complex and multifaceted nature of the concept of free speech, which encompasses different philosophical and legal standpoints (see, for example, Warburton, 2009, for a review), no single measure can ever capture all facets of online discourse. Similarly, despite extensive efforts by the research community, no measure of hate speech can ever fully capture the importance of cultural context and changing societal norms (e.g., Brown, 2017).

Nonetheless, automated hate speech detection tools have proven helpful to rein in toxic speech and are hence widely deployed online. As online platforms and regulators inevitably face trade-offs when it comes to moderating online content, we believe it is essential to have measures that also capture the cost of content moderation. We believe that by shedding light on the trade-offs involved in content moderation, our measure represents a fundamental advancement in the application of NLP tools for content moderation and highlights a highly fruitful direction for future research.

Our rephrasing analysis also highlights an important cost–benefit trade-off inherent in content moderation. On the benefit side, rephrasing reduces toxicity to a similar extent as outright removal while preserving substantially more semantic content. As our results show, rephrasing induces smaller distortions to the semantic space, maintains a broader range of expressed views, and preserves the possibility of engagement and exposure that would otherwise be eliminated under removal. From the perspective of information economics and market design, moderation rules can be viewed as platform design choices that shape the set of information products that remain available and their visibility (Bergemann and Morris, 2019; Kominers and Shapiro, 2024). Selective removal, therefore, distorts the information space on which learning takes place, with implications for beliefs, participation, and incentives across users of social media platforms.

At the same time, rephrasing entails nontrivial costs. Relative to automated removal, it requires additional computational and organizational resources, including the deployment of language models, monitoring of output quality, and safeguards against unintended changes in meaning. Rephrasing may also introduce residual distortions through changes in tone or specificity, and it may raise governance concerns if users perceive rewritten content as intrusive or manipulative. Moreover, rephrasing is not suitable for all types of content, as some material may be undesirable to retain in any form, such as defamation or direct threats. Rephrasing should therefore be understood not as a costless substitute for removal, but as a strategy that trades higher implementation costs for potentially lower informational and expressive costs.

While our empirical application focuses on content moderation on social media platforms, the proposed measure could also be applied to quantify distortions to media content arising from other interventions, such as legal restrictions on speech, changes in defamation law, or regulatory shocks affecting media environments. The implications of our findings are likely to be even more pronounced in autocratic settings, where restrictions on freedom of expression are typically more extensive and where governments' incentives often diverge from those of citizens (e.g., Besley and Prat, 2006; Egorov et al., 2009). In such environments, censorship is frequently used to selectively reshape the information environment, thereby limiting learning about political and economic conditions (e.g., Enikolopov et al., 2011; Qin et al., 2017; Guriev and Treisman, 2019). From this perspective, our measure provides a tool to quantify how strongly such interventions distort the media landscape."

It is also worth highlighting that our approach is agnostic to the specific mechanism through which content moderation is implemented. In practice, moderation decisions may be made by automated classifiers, crowdsourced reporting, professional content moderators, or hybrid systems that combine these approaches. These approaches may differ in systematic ways, for example, because automated systems tend to rely on keyword- or model-based thresholds that may disproportionately flag certain forms of language, crowdsourced reporting may reflect the preferences or coordination of active user groups, and professional moderators may apply platform guidelines more contextually but at higher cost and lower scale. The proposed measure can be applied uniformly across these settings to quantify how different moderation strategies distort the semantic composition of online discourse.

References

Abou-Moustafa, K. T. and F. P. Ferrie (2012). A note on metric properties for some divergence measures: The gaussian case. *25*, 1–15.

Acemoglu, D., T. A. Hassan, and A. Tahoun (2017, 08). The Power of the Street: Evidence from Egypt's Arab Spring. *The Review of Financial Studies 31*(1), 1–42.

Allcott, H. and M. Gentzkow (2017, May). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives 31*(2), 211–36.

Angelov, D. (2020). Top2vec: Distributed representations of topics.

Argyle, L. P., C. A. Bail, E. C. Busby, J. R. Gubler, T. Howe, C. Rytting, T. Sorensen, and D. Wingate (2023). Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences 120*(41), e2311627120.

Arora, A. and M. Dell (2024). Linktransformer: A unified package for record linkage with transformer language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 221–231.

Ars Technica (2012). Twitter releases "transparency tool" to reveal government requests. https://arstechnica.com/information-technology/2012/07/twitter-releases-transparency-tool-to-reveal-government-requests/.

Ash, E. and S. Hansen (2023). Text algorithms in economics. *Annual Review of Economics 15*, 659–688.

Attanasio, G., D. Nozza, D. Hovy, and E. Baralis (2022, May). Entropy-based attention regularization frees unintended bias mitigation from lists. In S. Muresan, P. Nakov, and A. Villavicencio (Eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland, pp. 1105–1119. Association for Computational Linguistics.

Barrera, O., S. Guriev, E. Henry, and E. Zhuravskaya (2020). Facts, Alternative Facts, and Fact Checking In Times of Post-truth Politics. *Journal of Public Economics 182*(C).

BBC (2017, Sep). Will germany's new law kill free speech online?; by Patrick Evans. `https://www.bbc.com/news/blogs-trending-41042266`.

BBC (2018, Sep). Twitter bans alex jones and infowars for abusive behaviour. `https://www.bbc.com/news/world-us-canada-45442417`.

Beknazar-Yuzbashev, G., R. Jiménez Durán, J. McCrosky, and M. Stalinski (2022). Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN*.

Beknazar-Yuzbashev, G., R. Jiménez Durán, and M. Stalinski (2024). A Model of Harmful yet Engaging Content on Social Media. *Available at SSRN*.

Bergemann, D. and S. Morris (2019). Information design: A unified perspective. *Journal of Economic Literature 57*(1), 44–95.

Besley, T. and A. Prat (2006). Handcuffs for the grabbing hand? media capture and government accountability. *American economic review 96*(3), 720–736.

Bianchi, F., S. HIlls, P. Rossini, D. Hovy, R. Tromble, and N. Tintarev (2022, December). "it's not just hate": A multi-dimensional perspective on detecting harmful speech online. In Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, pp. 8093–8099. Association for Computational Linguistics.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*(Jan), 993–1022.

Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems 29*.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler (2012). A 61-Million-Person Experiment in Social Influence and Political Mobilization. *Nature 489*(7415), 295.

Bose, P., L. Lupo, M. Habibi, D. Hovy, and C. Schwarz (2024). Beyond the stats: Realities, perception, and social media discourse on poverty. In *AEA Papers and Proceedings*, Volume 114, pp. 690–694. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Boxell, L., M. Gentzkow, and J. M. Shapiro (2017). Greater Internet Use Is Not Associated with Faster Growth in Political Polarization Among US Demographic Groups. *Proceedings of the National Academy of Sciences of the United States of America*, 201706588.

Brown, A. (2017). What is hate speech? part 1: The myth of hate. *Law and Philosophy 36*, 419–468.

Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2019, December). Social Media and Xenophobia: Evidence from Russia. Working Paper 26567, National Bureau of Economic Research.

Cao, A., J. M. Lindo, and J. Zhong (2023). Can social media rhetoric incite hate incidents? Evidence from Trump's "Chinese Virus" tweets. *Journal of Urban Economics 137*, 103590.

Card, D., S. Chang, C. Becker, J. Mendelsohn, R. Voigt, L. Boustan, R. Abramitzky, and D. Jurafsky (2022). Computational analysis of 140 years of us political speeches reveals more

positive but increasingly polarized framing of immigration. *PNAS 119*(31), e2120510119.

Cer, D., Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil (2018, November). Universal sentence encoder for English. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–174.

Chandrasekharan, E., U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert (2017, dec). You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact. 1*(CSCW).

Delgado, P. (2019, Mar). How el país used ai to make their comments section less toxic. `https://blog.google/outreach-initiatives/google-news-initiative/how-el-pais-used-ai-make-their-comments-section-less-toxic/`.

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

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs].

Dixon, L., J. Li, J. Sorensen, N. Thain, and L. Vasserman (2018). Measuring and Mitigating Unintended Bias in Text Classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Du, X. (2023). Symptom or Culprit? Social Media, Air Pollution, and Violence.

Egorov, G., S. Guriev, and K. Sonin (2009). Why resource-poor dictators allow freer media: A theory and evidence from panel data. *American political science Review 103*(4), 645–668.

Eidelman, V. and K. Ruane (2021, Jun). The problem with censoring political speech online – including trump's. `https://www.aclu.org/news/free-speech/the-problem-with-censoring-political-speech-online-including-trumps`.

Engadget (2020). Twitter plans for the worst with new election misinformation policy. https://www.engadget.com/twitter-updates-election-misinformation-rules-183211931.html.

Enikolopov, R., A. Makarin, and M. Petrova (2020). Social Media and Protest Participation: Evidence from Russia. *Econometrica 88*(4), 1479–1514.

Enikolopov, R., M. Petrova, and K. Sonin (2018). Social Media and Corruption. *American Economic Journal: Applied Economics 10*(1), 150–174.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011). Media and political persuasion: Evidence from russia. *American economic review 101*(7), 3253–3285.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on*

empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp. 55–65.

Feltham, G. A. and J. Xie (1994). Performance measure congruity and diversity in multi-task principal/agent relations. *Accounting review*, 429–453.

Fergusson, L. and C. Molina (2021, April). Facebook Causes Protests. Documentos CEDE 018002, Universidad de los Andes - CEDE.

Forbes (2019, Oct). Faceit and google partner to use ai to tackle in game toxicity; by Mike Stubbs. `https://www.forbes.com/sites/mikestubbs/2019/10/23/faceit-and-google-partner-to-use-ai-to-tackle-in-game-toxicity/`. Accessed: [Insert date here].

Fujiwara, T., K. Müller, and C. Schwarz (2021). The Effect of Social Media on Elections: Evidence From the United States. *NBER Working Papper*.

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences 115*(16), E3635–E3644.

Gehman, S., S. Gururangan, M. Sap, Y. Choi, and N. A. Smith (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Guriev, S., N. Melnikov, and E. Zhuravskaya (2020, 12). 3G Internet and Confidence in Government. *The Quarterly Journal of Economics*.

Guriev, S. and D. Treisman (2019). Informational autocrats. *Journal of economic perspectives 33*(4), 100–127.

Han, K., A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang (2021). Transformer in transformer. *Advances in Neural Information Processing Systems 34*, 15908–15919.

Han, X. and Y. Tsvetkov (2020). Fortifying toxic speech detectors against veiled toxicity. *arXiv preprint arXiv:2010.03154*.

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

Hartvigsen, T., S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

He, P., X. Liu, J. Gao, and W. Chen (2021, October). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv:2006.03654 [cs].

Hede, A., O. Agarwal, L. Lu, D. C. Mutz, and A. Nenkova (2021, April). From toxicity in online comments to incivility in American news: Proceed with caution. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online, pp. 2620–2630.

Henry, E., S. Guriev, T. Marquis, and E. Zhuravskaya (2023, December). Curtailing False News, Amplifying Truth. CEPR Discussion Papers 18650, C.E.P.R. Discussion Papers.

Henry, E., E. Zhuravskaya, and S. Guriev (2022). Checking and Sharing Alt-facts. *American Economic Journal: Economic Policy 14*(3), 55–86.

Holmstrom, B. and P. Milgrom (1991). Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization 7*(special_issue), 24–52.

Howard, P. N., A. Duffy, D. Freelon, M. Hussain, W. Mari, and M. Maziad (2011). Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring? *Working Paper*.

Jhaver, S., C. Boylston, D. Yang, and A. Bruckman (2021, oct). Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proc. ACM Hum.-Comput. Interact. 5*(CSCW2).

Jiménez Durán, R. (2022). The Economics of Content Moderation: Theory and Experimental Evidence From Hate Speech on Twitter. *Available at SSRN*.

Jiménez Durán, R., K. Müller, and C. Schwarz (2022). The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG. *Available at SSRN*.

Jones, J. J., R. M. Bond, E. Bakshy, D. Eckles, and J. H. Fowler (2017, 04). Social Influence and Political Mobilization: Further Evidence From a Randomized Experiment in the 2012 U.S. Presidential Election. *PLOS ONE 12*(4), 1–9.

Justitia (2020). The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship - Act two.

Kailath, T. (1967). The divergence and bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology 15*(1), 52–60.

Kominers, S. D. and J. M. Shapiro (2024). Content Moderation with Opaque Policies. Working Paper 32156, National Bureau of Economic Research.

Kozlowski, A. C., M. Taddy, and J. A. Evans (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review 84*(5), 905–949.

Kozyreva, A., S. M. Herzog, S. Lewandowsky, R. Hertwig, P. Lorenz-Spreen, M. Leiser, and J. Reifler (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences 120*(7), e2210666120.

Lee, C., K. Gligorić, P. R. Kalluri, M. Harrington, E. Durmus, K. L. Sanchez, N. San, D. Tse, X. Zhao, M. G. Hamedani, et al. (2024). People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise. *Proceedings of the National Academy of Sciences 121*(38), e2322764121.

Levy, R. (2021, March). Social Media, News Consumption, and Polarization: Evidence from a Field Experiment. *American Economic Review 111*(3), 831–70.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, B., H. Zhou, J. He, M. Wang, Y. Yang, and L. Li (2020). On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 9119–9130.

Liang, P. P., I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency (2020, July). Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 5502–5515.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019, July). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs].

Liu, Y., P. Yildirim, and Z. J. Zhang (2021). Social Media, Content Moderation, and Technology. *arXiv preprint arXiv:2101.04618*.

Lupo, L., P. Bose, M. Habibi, D. Hovy, and C. Schwarz (2024). Dadit: A dataset for demographic classification of italian twitter users and a comparison of prediction methods. *arXiv preprint arXiv:2403.05700*.

Madio, L. and M. Quinn (2021). Content Moderation and Advertising in Social Media Platforms. *Available at SSRN 3551103*.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. In *Proceedings of the National Institute of Science of India*, Volume 12, pp. 49–55.

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

McInnes, L., J. Healy, and J. Melville (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Michailovich, O., Y. Rathi, and A. Tannenbaum (2007). Image segmentation using active contours driven by the bhattacharyya gradient flow. *IEEE Transactions on Image Processing 16*(11), 2787–2801.

Mosquera, R., M. Odunowo, T. McNamara, X. Guo, and R. Petrie (2020). The Economic Effects of Facebook. *Experimental Economics 23*(2), 575–602.

Müller, K. and C. Schwarz (2021). Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association 19*(4), 2131–2167.

Müller, K. and C. Schwarz (2022a). The effects of online content moderation: Evidence from president trump's account deletion. *Available at SSRN 4296306*.

Müller, K. and C. Schwarz (2022b). From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment. *American Economic Journal: Applied Economics*.

Munzert, S., R. Traunmüller, P. Barberá, A. Guess, and J. Yang (2025). Citizen preferences for online hate speech regulation. *PNAS Nexus*, pgaf032.

NBC-News (2018, Oct). Facebook removes pages belonging to far-right group 'proud boys', by david ingram. `https://www.nbcnews.com/tech/social-media/facebook-removes-pages-belonging-far-right-group-proud-boys-n926506`.

Nguyen, D. Q., T. Vu, and A. Tuan Nguyen (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, pp. 9–14. Association for Computational Linguistics.

NYT (2016, Sep). The times is partnering with jigsaw to expand comment capabilities. `https://www.nytco.com/press/the-times-is-partnering-with-jigsaw-to-expand-comment-capabilities/`.

NYT (2021, Jan). Twitter permanently bans trump, capping online revolt; by Kate Conger and Mike Isaac. `https://www.nytimes.com/2021/01/08/technology/twitter-trump-suspended.html`.

OpenAI (2024). Openai - moderation api. https://platform.openai.com/docs/api-reference/moderations.

Perspective (2017). Perspective api. `https://www.perspectiveapi.com/`.

Prat, A. and D. Strömberg (2013). The political economy of mass media. *Advances in economics and econometrics 2*, 135.

Qin, B., D. Strömberg, and Y. Wu (2017). Why does china allow freer social media? protests versus surveillance and propaganda. *Journal of Economic Perspectives 31*(1), 117–140.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog 1*(8), 9.

Reimers, N. and I. Gurevych (2019, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Reuters (2023a, Aug). Big tech braces for eu digital services act regulations; by Martin Coulter. https://www.reuters.com/technology/big-tech-braces-roll-out-eus-digital-services-act-2023-08-24/.

Reuters (2023b, Sep). Uk's online safety bill finally passed by parliament; by Paul Sandle. https://www.reuters.com/world/uk/uks-online-safety-bill-passed-by-parliament-2023-09-19/.

Rieder, B. and Y. Skop (2021). The Fabrics of Machine Moderation: Studying the Technical, Normative, and Organizational Structure of Perspective API. *Big Data & Society 8*(2), 20539517211046181.

Ross, B., M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki (2016). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *3rd Workshop on Natural Language Processing for Computer-Mediated Communication/Social Media*, pp. 6–9. Ruhr-Universitat Bochum.

Röttger, P., B. Vidgen, D. Hovy, and J. Pierrehumbert (2022, July). Two contrasting data annotation paradigms for subjective NLP tasks. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, pp. 175–190. Association for Computational Linguistics.

Samples, J. (2019, Apr). Why the government should not regulate content moderation of social media. https://www.cato.org/policy-analysis/why-government-should-not-regulate-content-moderation-social-media.

Siegel, A. A., E. Nikitin, P. Barberá, J. Sterling, B. Pullen, R. Bonneau, J. Nagler, J. A. Tucker, et al. (2021). Trumping Hate on Twitter? Online Hate in the 2016 US Election Campaign and its Aftermath. *Quarterly Journal of Political Science 16*(1), 71–104.

Solomon, B. C., M. E. Hall, A. Hemmen, and J. N. Druckman (2024). Illusory interparty disagreement: Partisans agree on what hate speech to censor but do not know it. *Proceedings of the National Academy of Sciences 121*(39), e2402428121.

Strömberg, D. (2004). Radio's impact on public spending. *The Quarterly Journal of Economics 119*(1), 189–221.

Strudel, R., R. Garcia, I. Laptev, and C. Schmid (2021). Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272.

Sunstein, C. R. (2017). *# Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

TechCrunch (2016). Twitter forms a "trust & safety council" to balance abuse vs free speech. https://techcrunch.com/2016/02/09/twitter-forms-a-trust-safety-council-to-balance-abuse-vs-free-speech/.

TechCrunch (2020). Twitter broadly bans any COVID-19 tweets that could help the virus spread. https://techcrunch.com/2020/03/18/twitter-coronavirus-covid-19-misinformation-policy/.

TechCrunch (2022). Twitter will hide false tweets from high-profile accounts during times of crisis. https://techcrunch.com/2022/05/19/twitter-crisis-misinformation-policy/.

The Guardian (2022). Elon Musk Reinstates Donald Trump's Twitter Account After Taking Poll, by Dan Milmo.

The Next Web (2012). Twitter: Tweets must flow, but.... https://thenextweb.com/news/twitter-tweets-must-flow-but.

The Texas Tribune (2022, Sep). Texas social media "censorship" law goes into effect after federal court lifts block; by Jesus Vidales. https://www.texastribune.org/2022/09/16/texas-social-media-law/.

Twitter (2021). Permanent Suspension of realDonaldTrump. https://blog.twitter.com/en_us/topics/company/2020/suspension.

Twitter Blog (2015a). Fighting abuse to protect freedom of expression. https://blog.twitter.com/2015/fighting-abuse-to-protect-freedom-of-expression.

Twitter Blog (2015b). Policy and product updates aimed at combating abuse. https://blog.twitter.com/2015/policy-and-product-updates-aimed-at-combating-abuse.

Tworek, H. (2021, Dec). History explains why global content moderation cannot work. https://www.brookings.edu/articles/history-explains-why-global-content-moderation-cannot-work/.

United Nations Human Rights (2018, Jul). Un expert: Content moderation should not trample free speech. https://www.ohchr.org/en/stories/2018/07/un-expert-content-moderation-should-not-trample-free-speech.

USA-Today (2022, Nov). Twitter layoffs slash content moderation staff as new ceo elon musk looks to outsource; by Barbara Ortutay and Matt O'Brien. https://www.usatoday.com/story/tech/2022/11/15/elon-musk-cuts-twitter-content-moderation-staff/10706732002/.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Vice (2016). The history of twitter's rules. https://www.vice.com/en/article/the-history-of-twitters-rules/.

Vogels, E. A., A. Perrin, and M. Anderson (2020, Aug). Most americans think social media sites censor political viewpoints. `https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/`.

Warburton, N. (2009). *Free speech: A very short introduction*. OUP Oxford.

Waseem, Z. and D. Hovy (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In J. Andreas, E. Choi, and A. Lazaridou (Eds.), *Proceedings of the NAACL Student Research Workshop*, San Diego, California, pp. 88–93.

Wilks, S. S. (1932, November). Certain Generalizations in the Analysis of Variance. *Biometrika 24*(3/4), 471.

Wulczyn, E., N. Thain, and L. Dixon (2017). Ex Machina: Personal Attacks Seen at Scale. *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399.

Zhu, X., W. Su, L. Lu, B. Li, X. Wang, and J. Dai (2020). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zhuravskaya, E., M. Petrova, and R. Enikolopov (2020). Political Effects of the Internet and Social Media. *Annual Review of Economics 12*.

Online Appendix

The online appendix presents further details on data construction, methodology, and robustness exercises:

- Appendix A provides additional details on the data.
- Appendix B provides additional details on the methodology.
- Appendix C provides additional results.

A Additional Details on Data

A.1. Representative Twitter Data

In our study, we initiated our data collection process with a cohort of 432,882 randomly selected American Twitter users, a dataset that was meticulously collected in 2015 by (Siegel et al., 2021). This particular sample selection bears two pivotal advantages. Firstly, it allows us to construct a comprehensive and representative overview of Twitter activity across the United States. Secondly, these users have been actively engaged on Twitter for several years, and we do not face problems with composition changes. Consequently, we were able to conduct our analysis based on a good approximation of the content that circulated on Twitter. The initial step in our data collection involved retrieving the Tweets posted by each user within the Siegel et al. (2021) sample. In totality, our dataset encompassed 399 Million Tweets spanning from 2014 to the commencement of 2022, along with corresponding user profile information. We provide some summary statistics in Table A.1

Table A.1: Summary Statistics

Number of Tweets	5M
Average Toxicity Perspectives API Score	0.19
Average Toxicity Detoxify API Score	0.11
Average OpenAI Moderation API Score	0.12

Notes: This table provides summary statistics on the number of users, the number of Tweets, and the average Tweet toxicity, as generated by the Perspective API, Detoxify, and OpenAI's Moderation API.

A.2. Filtering Political Tweets

We developed a political Tweet classifier by fine-tuning the BERTweet model on a human-labeled dataset of political and non-political Tweets. To create the training dataset, we first compiled a list of political keywords and accounts associated with political figures. This produced a subset of data with a higher density of political Tweets, though many non-political Tweets remained. We then randomly sampled 10,000 Tweets and distributed them among five undergraduate research assistants for manual labeling. Coders were provided with labeling criteria and instructed to categorize each Tweet in a binary true/false format. Each subset included 200 shared Tweets to evaluate inter-coder reliability. We evaluated agreement using Cohen's kappa coefficient, which ranges from 0 to 1, where 1 indicates perfect agreement and values near zero suggest agreement no better than chance. The average Cohen's kappa score across all coder pairs was 0.76. We then use this dataset to train and evaluate our political Tweet classifier.

Table A.2: Confusion Matrix

		Predicted Label		
		Non-Political	Political	**Total**
True Label	Non-Political	647 (0.94)	38 (0.06)	685
	Political	59 (0.19)	256 (0.81)	315
	Total	706	294	1,000
Accuracy: 0.903, **Precision:** 0.871, **Recall:** 0.813, **F1-score:** 0.841				

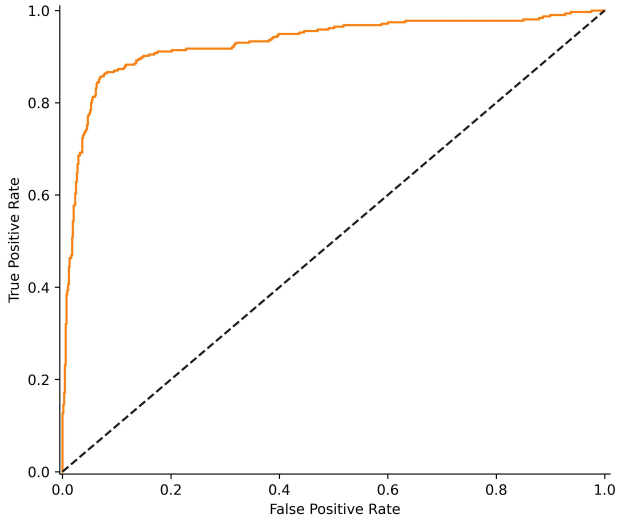
Notes: This table reports the confusion matrix for the political Tweet classifier.

From the 10,000 unique labeled examples (we removed duplicates used for inter-coder reliability assessment), we allocated 8,000 for training, 1,000 for in-training evaluation, and 1,000 for the final test set. We fine-tuned the BERTweet model for binary classification over 10 epochs, implementing an early stopping criterion based on the F1 score of the evaluation set. On the test set, the model achieved an accuracy score of 0.90 and an F1 score of 0.84. Table A.2 presents the evaluation results on the test examples. Figure A.1 illustrates the Receiver Operating Characteristic curve for the test set, with an area under the curve of 0.88.

A.3. Toxicity Measures

To gauge the toxicity of these Tweets, we use Google's Perspective API, a tool widely acknowledged for its efficacy in identifying hate speech (Wulczyn et al., 2017; Dixon et al., 2018). This state-of-the-art API assigns a toxicity score ranging from 0 to 1 across six distinct dimensions: general toxicity, severe toxicity, identity attacks, insults, profanity, and threats.

Figure A.1: ROC Curve



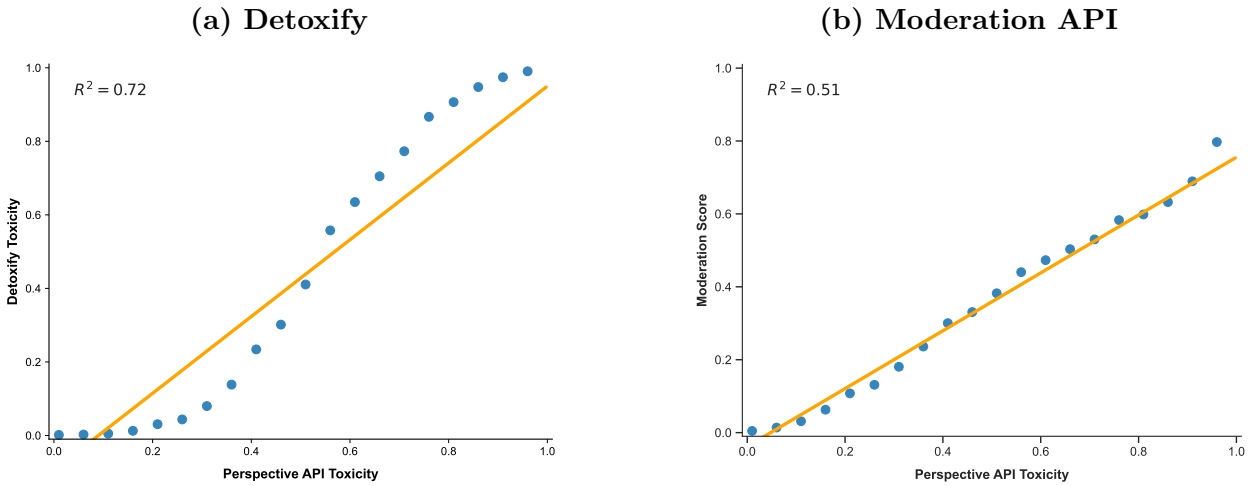
Notes: The figure reports the ROC curve for the political Tweet classifier.

The scores can be approximately interpreted as the probability that a randomly chosen user will classify content as toxic. For example, a score of 0.8 means that around 80% of users would judge the content to be toxic.

The Perspective API performs well in classifying toxic text and can assess Tweets in several languages, including English, Spanish, French, German, Portuguese, Italian, and Russian. In our dataset, English is, unsurprisingly, the predominant language, and we restrict our analysis to English-speaking Tweets. Instances where the API did not assign toxicity scores were primarily due to the absence of textual content, such as Tweets containing only hyperlinks. We provide examples of highly toxic Tweets in Table 1. The examples indicate that the Perspective API accurately identifies toxic content. As described in the main paper, we additionally use the Detoxify (Hanu and Unitary team, 2020) package and OpenAI's Moderation API (OpenAI, 2024) as an alternative toxicity classification algorithm. Overall, we also find that the different models broadly agree in their toxicity evaluation of Tweets. The correlation between toxicity scores from the Perspective API and Detoxify is 0.85, while the correlation between the Perspective API and the Moderation API is 0.72.[23]

[23]As the Moderation API does not report a direct Toxicity score, we always consider the maximum of all provided scores.

Figure A.2: Comparison Toxicity Scores: Perspective API vs. Detoxify vs. OpenAI



Notes: The figure shows a binscatter plot of the toxicity scores from the Perspectives API relative to the scores from the Detoxify package and OpenAI's moderation API. The line was fitted based on a linear regression. Data points are grouped into 20 bins of equal size.

Distribution of Toxicity Across Users

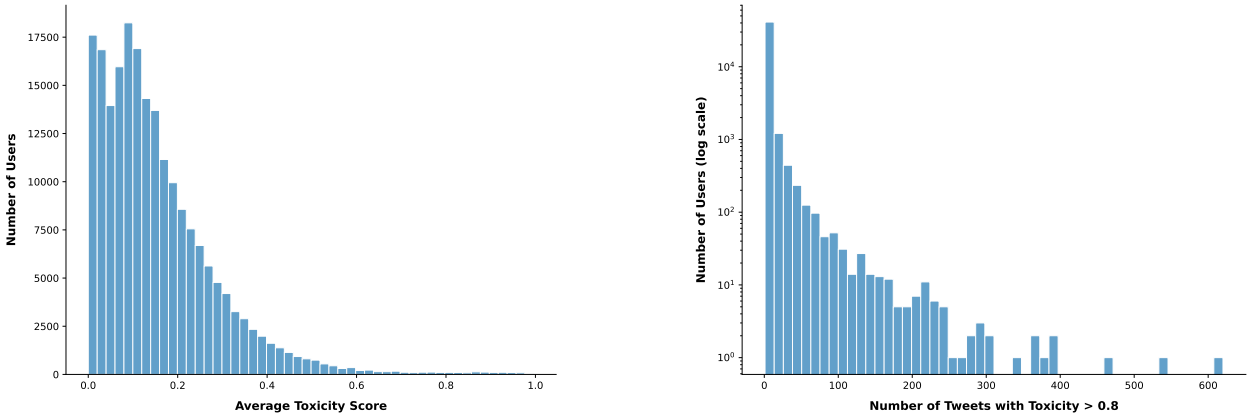
An important question for interpreting the implications of toxicity-based moderation is how toxic content is distributed across users. In particular, if toxic language were produced primarily by a very small set of highly active accounts, moderation might disproportionately affect this minority rather than broadly constraining the expressive capacity of typical users. Conversely, if toxicity is more widely distributed across the user base, moderation would have broader implications for the composition of online discourse.

Figure A.3 illustrates the distribution of toxic Tweets across users in our sample. This exercise allows us to assess whether toxic content is primarily generated by a small group of highly active users or whether toxicity is more broadly distributed across the user population. We find that the average toxicity scores have a skewed distribution, with most users producing content with an average toxicity below 0.125 (see Panel a). In Panel (b), we show the average number of highly toxic Tweets (toxicity > 0.8) per user. While still right-tailed, the distribution shows that there is a significant fraction of users who produce at least some toxic content.

Taken together, these figures suggest that toxicity-based moderation likely does not affect only a small group of fringe users but also shapes the overall composition of online discourse. This finding supports an interpretation of our results in which content moderation can alter the representativeness of expressed views, rather than merely suppressing the speech of a marginal subset of users.

Figure A.3: Distribution of Toxicity Across Users

(a) Average Toxicity **(b) Number of Toxic Tweets**



Notes: The figure shows the toxicity distribution at the user level. Panel (a) shows the distribution of average toxicity scores by user. Panel (b) shows the distribution of the number of highly toxic tweets (toxicity > 0.8) produced by each user, with the y-axis on a logarithmic scale. Toxicity scores are based on the Perspective API.

A.4. Twitter's Content Moderation Policies on Hate and Toxicity, 2006–2022

For most of the period covered by our data, Twitter maintained a set of content moderation policies that explicitly prohibited hate speech, abusive behavior, and other forms of toxic expression, even as the platform's enforcement practices and conceptualizations of objectionable speech evolved. In the following, we provide a short outline of some of the important content moderation milestones.

- **Early Period 2006 to 2012.** From its launch in 2006 through the early 2010s, Twitter relied on a comparatively thin and reactive rule framework: policy and enforcement focused heavily on operational integrity (e.g., spam) and legal compliance, while protections against harassment and hate were less explicit and less systematized than in later years (Vice, 2016).

- **2012: Country-specific content withholding (jurisdictional compliance).** Twitter began using (and publicly discussing) mechanisms to withhold specific content in particular countries rather than remove it globally, reflecting growing cross-national legal constraints and the emergence of "geo-blocked" moderation as a distinct tool (The Next Web, 2012).

- **2012: Transparency reporting (public accountability for removals and requests).** Twitter launched a transparency reporting tool covering government requests

for user data and content removal, and also began reporting copyright requests, providing a public record of state and private legal pressures that drive moderation outcomes (Ars Technica, 2012).

- **2015: Anti-abuse pivot (expanded threat standards + new enforcement instruments).** Twitter broadened its approach to violent threats and abuse and introduced stronger enforcement mechanisms, including time-based account locks and early forms of reach-limiting for suspected abusive content. This marks a key shift from primarily complaint-driven removals to a broader enforcement toolkit (Twitter Blog, 2015b).

- **2015: Explicit "hateful conduct" and self-harm policies.** Twitter updated its rules to explicitly prohibit hateful conduct and added policy language and interventions concerning self-harm, signaling a wider definition of harm beyond direct threats and toward group-based and wellbeing-related harms (Twitter Blog, 2015a).

- **2016: Institutionalization via Trust & Safety governance.** Twitter created a Trust & Safety Council, formalizing external stakeholder engagement and reinforcing the platform's framing of moderation as a balancing problem between safety and freedom of expression (TechCrunch, 2016).

- **2020: COVID-19 misinformation policy (public-health information integrity).** During the pandemic, Twitter adopted explicit rules against COVID-19 misinformation, extending moderation beyond harassment/hate into health-related claims where harms arise via behavioral influence (e.g., discouraging expert guidance or promoting false cures) (TechCrunch, 2020).

- **2020: Expanded election/civic integrity enforcement.** Twitter strengthened its approach to election-related misinformation via more explicit civic integrity rules and enforcement (often combining labels and other distribution interventions), reflecting the platform-wide shift toward structured governance of high-stakes political information flows (Engadget, 2020).

- **2022: "Crisis misinformation" policy.** Twitter introduced a framework aimed at reducing the spread and amplification of misleading claims during crises (e.g., wars and disasters), further entrenching the use of labeling and amplification controls as core moderation tools (TechCrunch, 2022).

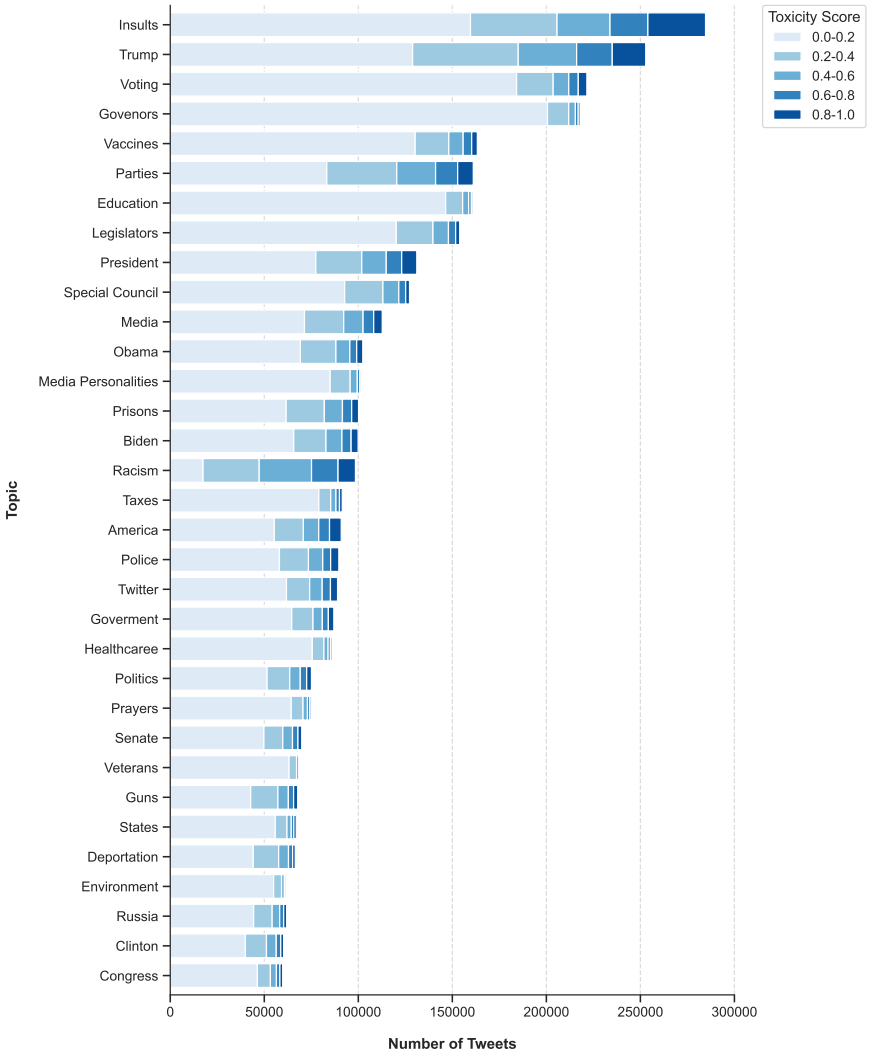
A.5. Additional Details on Topic Model

Table A.3: Most Relevant Words by Topic

Number	Topic Label	Topic Words
0	Insults	idiot hates hahahahaha stfu moron fuk scum yawn dumbass yikes pathetic
1	Trump	trump trumps djt trumpers drumpf trumpster trumpy trumpsters potus donald nevertrump
2	Voting	voter ballot voted voting ballots voters vote votes disenfranchise elections election
3	Govenors	rauner mccrory mayor gubernatorial blasio mayors governor lepage rahm redistricting mayoral
4	Vaccines	vaccinate vaccinating unvaccinated immunization vaccinated vaccine vaccines vaccinations vaccination cdc vaxxers
5	Parties	gops gop repub repubs dems republican republicans democrats rinos democrat bipartisanship
6	Education	devos defund rauner educators privatize underfunded taxpayers nea education educate defunded
7	Legislators	wyden durbin legislator cispa grandstanding legislate lawmakers toomey rauner lawmaker
8	President	potus presidential trump prez obummer president trumps presidency presidents obamas whitehouse
9	Special Council	manafort indictments mueller comey gowdy kushner grassley dershowitz colluded trumpers nunes
10	Media	msnbc foxnews cnn maddow hannity olbermann drudge megyn cspan tyt reuters
11	Obama	obamas obama obummer nobama barack potus prez newsmax presidential teleprompter romney
12	Media Personalities	fareedzakaria govmikedewine katrinapierson krystalball plf johnbrennan flapol narendramodi laurenboebert bhive mikepence
13	Prisons	jailing jailed exonerated prosecutors sentencing sentenced jails indict incarceration imprison correctional
14	Biden	biden bidens kaine vp palin obummer nobama cheney reince joe veep
15	Racism	racists racist racism racial sharpton naacp racially divisiveness klan supremacist blacks
16	Taxes	tax taxation taxing fiscal taxes taxed irs taxable deductions redistribution cbo
17	America	america merica murica unamerican amerikkka americas americans american patriotic usa unpatriotic
18	Police	policemen police cops policeman policing acab nypd sheriffs lapd dorner cop
19	Goverment	govt government gov goverment governmental governments feds bureaucrats privatize privatized govs
20	Healthcaree	obamacare cbo singlepayer medicare aca healthcare repeal medicaid ahca trumpcare uninsured
21	Politics	politics political politic politicians apolitical politically politician politicized pols partisanship divisiveness
22	Prayers	prayers praying prayer pray prayed condolences amen blessings thankful salute praise
23	Twitter	tweeting tweets tweet tweeted retweets twitter retweeted tweeter retweeting retweet covfefe
24	Senate	mcconnell grassley gops rinos senate manchin bipartisanship toomey senators filibuster obstructionist
25	Veterans	veterans servicemen veteran salute honoring commemorate soldiers commemorating vets thanking heros
26	Guns	nra guns giffords firearms shootings gun firearm gunman armed massacres feinstein
27	States	florida fla floridians fl ohioans michiganders broward tallahassee kentucky hoosier ohio
28	Deportation	deportations illegals deporting deportation immigration deport deported amnesty immigrants migrant immigrant
29	Environment	polluters algore renewables epa fracking greenpeace polluting globalwarming politicized deniers exxon
30	Russia	putin kremlin russia crimea kgb russians ukrainian russian ukraine ukrainians oligarchs
31	Clinton	hillary clintons hilary clinton hrc killary huma trump lewinsky gillibrand trumps
32	Congress	congressmen congressional congress senate lawmakers constitutional congressman grassley incumbents wyden senatorial
33	Left-Right	liberal liberals conservatives conservative conservatism liberalism leftists leftist libs rightwing libtard
34	Federal Reserve	bernanke yellen fomc krugman bullish zerohedge schiff jpmorgan cnbc economist recession
35	Protests	protestors protesters protesting protester protests protestor demonstrators protest rioting rioters protested
36	Democracy	democracy democracies undemocratic democratically oligarchy democratic dictatorship dictatorships tyranny unelected demagogue
37	Supreme Court	scotus scalia sotomayor alito ginsburg kagan gorsuch justices recuse grassley rbg
38	Impeachment	impeachment impeach impeached impeachable impeaching indictments grassley dershowitz pardons sedition pardoning
39	Civil Rights	mlk confederate confederacy gettysburg juneteenth tubman proclamation secession naacp commemorating divisiveness
40	Canada	trudeau scheer ndp harper kenney wynne rudd cpc tories conservatives libdems
41	Terrorism	terrorism terrorist terrorists jihadist jihadists radicalized islamist bombings islamists islamophobia qaida
42	Budget	debt deficits cbo krugman debts austerity bernanke trillion deficit bailout defund
43	Syria	airstrikes assad syria daesh isil bashar mosul jihadists isis jihadist syrian
44	UK	libdems miliband farage clegg brexit tories corbyn ukip gove snp labour
45	Israel	israelis netanyahu zionist zionists palestinians aipac antisemitism antisemitic israel zionism idf
46	LGBTY	doma lgbtq gays homophobic gaymarriage homophobia bigots legislating dadt homosexuality gsa
47	Sanders	bernie sanders feelthebern bern hillary dnc hrc electable tyt canvassing disavow
48	Debate	debates debate debating debated discussions cspan rebuttal controversy argument discourse fiorina
49	India	bjp kejriwal modi swamy ndtv gujarat nawaz bihar aap narendra mlas
50	Sedition	sedition inciting facist censorship freedoms fascist incitement fascism ammendment totalitarianism authoritarians
51	Socialism	socialism socialist socialists capitalist marxism marxist communist communism capitalism marxists capitalists
52	Romney	romney romneys mitt romneyryan obummer santorum nobama mittens obama gingrich gops
53	Petitions	petition petitions signed signatures impeach signing amnesty protest oppose cispa sopa
54	China	china jinping chinese taiwan beijing mao gao ccp hk reuters hong
55	Africa	nigeria nigerians nigerian mugabe davidcorndc niger zuma anc lagos boko somali
56	Marijuana	legalization decriminalize legalize legalizing legalized marijuana mmj cannabis thc hemp weed
57	Abortion	prolife prochoice abortions abortion lifers aborted unborn fetus jimmykimmel fetal personhood
58	Southern States	bama alabama auburn lsu mizzou mississippi arkansas clemson birmingham tennessee kentucky
59	Iran	iran iranians ahmadinejad iranian khamenei mullahs tehran rouhani netanyahu shah kissinger
60	Republican Primaries	cruz rubio kasich ted nevertrump fiorina rino reince grassley rinos huckabee
61	Nazis	nazi nazis goebbels nazism hitler holocaust gestapo adolf reich auschwitz fascist
62	Afghanistan	taliban afghanistan afghan afghans kabul qaeda bergdahl insurgents qaida osama jihadist
63	BLM	blm alllivesmatter sharpton naacp protestors antifa disavow supremacists protestor protesters seiu
64	North Korea	dprk nk kim seoul jong rodman korea missiles wwiii korean nukes
65	Twitter	tweeted tweets retweeted tweet confirms tweeting withheld twitpic cancelled retweets retweet
66	Vice Presidents	pence kaine vp impeaching impeachment impeach djt trumpers mattis impeached nevertrump

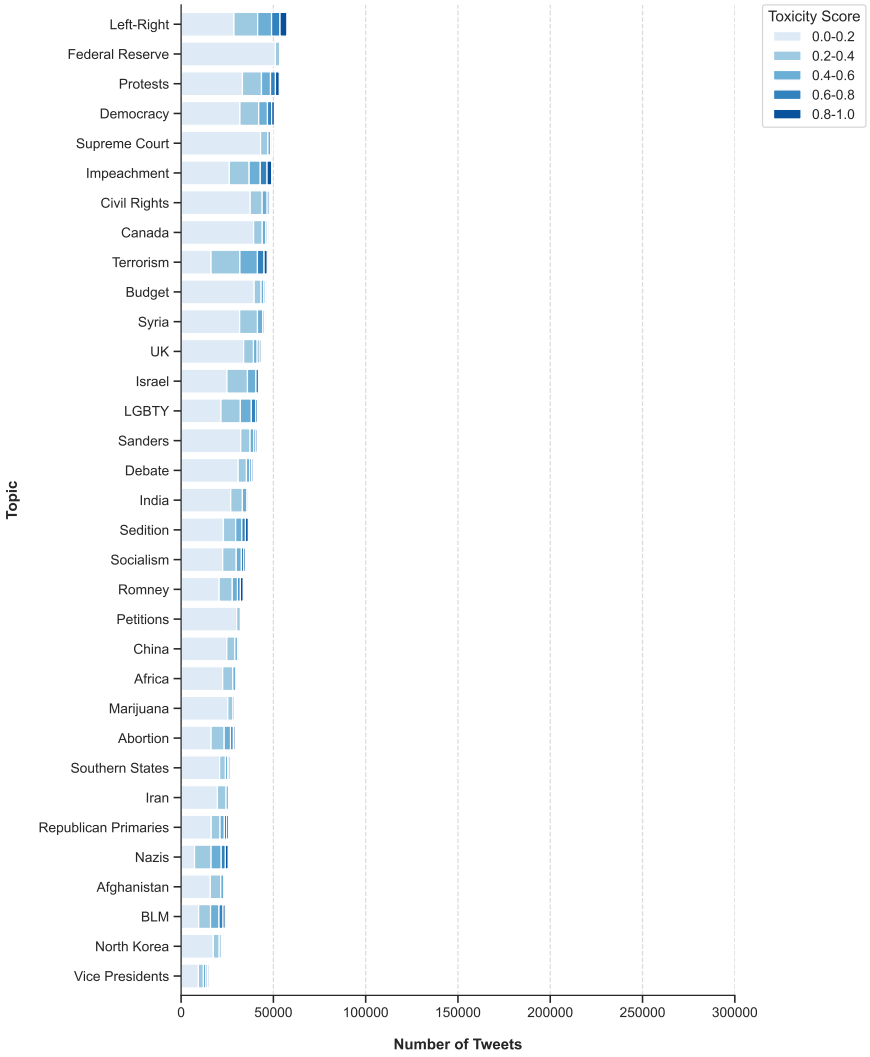
Notes: This table reports the 10 most important topic words for each of the 67 topics generated by the Top2Vec topic model. The labels were assigned by the authors based on the topic words.

Figure A.4: Toxicity Composition of Topics (1/2)



Notes: The figure shows the size and the toxicity composition of each of the topics created by the Top2Vec topic model.

Figure A.4: Toxicity Composition of Topics (2/2)



Notes: The figure shows the size and the toxicity composition of each of the topics created by the Top2Vec topic model.

B Additional Details on Methodology

B.1. Additional Details on Embeddings

To distill information from the raw text of Tweets into quantifiable data, we use three pre-trained models: BERTweet (Nguyen et al., 2020), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and DistilBert-Base-Multilingual-Cased-V2 (Reimers and Gurevych, 2019). These models all use on the BERT-style Transformer architecture, which has become a staple in text classification and natural language understanding tasks. The strength of these models comes from their utilization of an "attention" mechanism, allowing them to evaluate words in the context of their surrounding text, thereby capturing the subtleties of language that are often lost in traditional analysis. We transform the raw Tweets into analyzable embeddings by tokenizing the text into its constituent word tokens. These tokens then serve as input to our models, which produce numerical representations—or embeddings—of each Tweet. These embeddings convey the semantic and syntactic nuances of the language used by Twitter users and form the backbone of our computational analysis. Further details on the specific attributes of these models within our study are provided in the following.

BERTweet

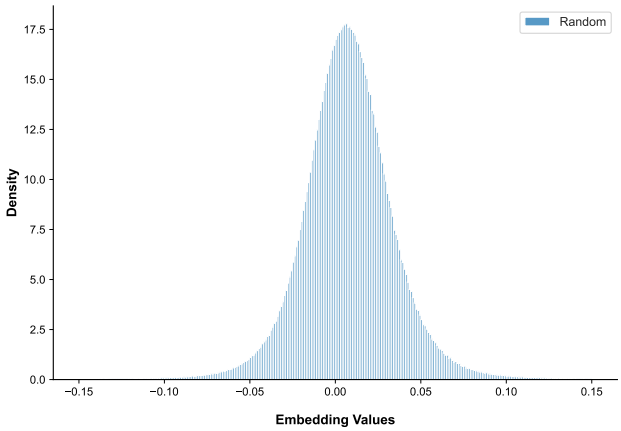
In the rapidly expanding field of Natural Language Processing (NLP), the inception of BERT (Bidirectional Encoder Representations from Transformers) and its subsequent iterations have marked a significant milestone. Introduced by (Devlin et al., 2019), BERT leverages an architecture known as Transformers (Vaswani et al., 2017) to process words in relation to all the other words in a sentence, which contrasts with prior models that viewed words in sequence. The BERT class of language models, with BERTweet as one example, are proficient in tasks such as part-of-speech tagging, named entity recognition, and text classification. The original BERT model was trained on an extensive corpus comprising sources like Wikipedia and books, known for their structured and formal English. However, the nature of Twitter's text, characterized by brevity and idiosyncratic language usage, presented a unique challenge for these models. To address this, BERTweet was specifically trained on an 80GB corpus containing 850 million English Tweets (Nguyen et al., 2020). This vast training corpus allows BERTweet to learn about the distinctive language patterns on Twitter.

To provide a brief overview of the data processing pipeline. Initially, the raw text of Tweets undergoes a tokenization process wherein the text is segmented into "tokens," which are the basic units for the model to understand. Imagine tokenization as the breaking down of a sentence into individual words and symbols, which are then analyzed by the language model.

Once tokenized, these Tweets are fed into the pre-trained BERTweet model. This model excels in interpreting each token in context, producing a vector of size 768 that captures not just the semantics of the individual token but also its relationship to others in the Tweet, all while considering the token's position.

Subsequently, we create an embedding (vector) that captures the content of the Tweet as a whole by taking a weighted average of all token embeddings based on their attention weights. Attention weights are a major component of Transformers that ensures that more influential tokens have a bigger impact on the final Tweet embedding. Put differently, the model distinguishes which words carry more weight in conveying the Tweet's overall message and adjusts the embedding accordingly. Ultimately, each Tweet is distilled into a unit-length vector within a 768-dimensional space, enabling nuanced interpretations and analyses. The 768-embedding dimensions are approximately normally distributed (see Figure B.1).

Figure B.1: Histogram of Embeddings



Notes: The figure shows a histogram of the embedding dimensions. For this figure, we aggregate data across all dimensions of the embedding for a sample of 10,000 Tweets.

RoBERTa

RoBERTa (Liu et al., 2019) is another widely used model that we can use to generate Tweet embeddings. Building upon the BERT foundation, RoBERTa implements several modifications to improve performance. In particular, RoBERTa extended the training duration, increased batch sizes, and exposed the model to a broader spectrum of data, including the large CC-NEWS dataset. RoBERTa also modifies BERT's training process by discarding the next sentence prediction objective and by training on longer sequences. Moreover, RoBERTa introduces variability in the masking pattern of the input data during training, which prevents the model from merely memorizing fixed patterns and encourages a deeper comprehension of

language nuances. Together, these modifications are beneficial for understanding the context more effectively and led RoBERTa to achieve state-of-the-art results on many NLP tasks and benchmarks.

DeBERTa

The third model we are using in our analysis is DeBERTa (He et al., 2021). DeBERTa (Decoding-enhanced BERT with disentangled attention) introduced innovative mechanisms that refine the workings of BERT and RoBERTa models. Its distinctive feature lies in the disentangled attention mechanism, which considers the content and the position of words separately, offering a more nuanced understanding of the text. Each word is represented by dual vectors that capture what the word is and where it stands in a sentence. This allows for a better interpretation of language nuances. Furthermore, DeBERTa's enhanced mask decoder predicts masked tokens by using their absolute positions, an improvement that aids in the pre-training process. With these advancements, DeBERTa improved the model's pre-training efficiency and improved the performance across a variety of downstream tasks, like the generation of new text that mimics human speech.

DistilBert-Base-Multilingual-Cased-V2

DistilBERT-Base-Multilingual-Cased-V2 (Reimers and Gurevych, 2019) is a lighter and faster variant of BERT, specifically trained to handle multiple languages in a single model. DistilBERT applies knowledge distillation during training. This process reduces the model's size and increases inference speed. The multilingual-cased version was trained on multiple languages using cased text, ensuring that it preserves distinctions in capitalization, which can be semantically relevant in many languages. As a result, this model is particularly well-suited for cross-lingual applications.

B.2. Additional Details the Bhattacharyya Distance

This subsection provides additional details on the Bhattacharyya distance (BCD) as a measure of distortions in semantic space, and provides guidance on its interpretation and applicability. The BCD was developed to quantify the distance between two probability distributions. Let P and Q denote two probability distributions over a random vector $X \in \mathbb{R}^d$ with densities $p(x)$ and $q(x)$. The Bhattacharyya coefficient is defined as:

$$BC(P, Q) = \int \sqrt{p(x)\, q(x)}\, dx. \tag{B.1}$$

The Bhattacharyya distance is then defined as:

$$BCD(P, Q) = -\ln\left(BC(P, Q)\right). \tag{B.2}$$

The coefficient $BC(P, Q)$ measures the overlap between two probability distributions and takes values in $[0, 1]$, where 1 indicates identical distributions. The BCD converts this overlap into a divergence measure, with larger values indicating greater separation.

The Bhattacharyya distance can be used to quantify differences between both unimodal and multimodal distributions, and it is not restricted to normal data-generating processes. Throughout this paper, we rely on the closed-form expression that applies under the assumption of multivariate normality:

$$BCD(\mathcal{N}_1, \mathcal{N}_2) = \frac{1}{8}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\left(\frac{\det \Sigma}{\sqrt{\det \Sigma_1 \cdot \det \Sigma_2}}\right) \tag{B.3}$$

In semantic embedding spaces, the distribution of values along each dimension of the high-dimensional space follows a normal distribution. Thus, for our application, the BCD over embedding spaces (which represent semantic content) intuitively captures the idea that two text corpora are more similar if their semantic spaces overlap. Importantly, the BCD is a distributional measure rather than a pairwise similarity measure. It does not compare individual texts to one another, but instead compares the overall geometry of the embedding space. Intuitively, this operation can be imagined as comparing the shape and density of two high-dimensional point clouds.

The key properties that make the BCD an attractive measure for our setting include:

- Symmetry: $BCD(P, Q) = BCD(Q, P)$.
- Invariance under linear transformations.
- Computational tractability in high-dimensional settings.

For normal distributions:

- If two distributions have identical means and covariance matrices, the BCD equals zero.
- The BCD increases when the means move apart.
- The BCD increases when the covariance matrices diverge.

The BCD has a long tradition in statistics and empirical classification problems as a measure of separability between probability distributions. Conceptually, it quantifies how difficult it

is to distinguish two data-generating processes based on observed features. In that sense, it provides a natural measure of distributional change that is closely related to statistical hypothesis testing.

In classical pattern recognition, the Bhattacharyya distance is used to measure class separability and provides an upper bound on the Bayes classification error (e.g., Kailath, 1967). In computer vision, the Bhattacharyya coefficient is used to compare feature distributions, such as color histograms, or image segmentation (e.g., Michailovich et al., 2007).

For researchers interested in applying this approach in other contexts, several practical considerations are worth highlighting. First, the BCD is most informative when the object of interest is a global change in the composition of content (such as moderation or a natural experiment setup), rather than local or pairwise similarities between individual texts. In other words, it does not measure why and where the meaning of the corpora changed, only that it did. Second, when strong departures from approximate normality are a concern, the BCD can be computed using nonparametric estimates of marginal distributions. Finally, the measure is not designed to replace topic models or similarity-based approaches, but rather to complement them by providing a scalable and interpretable summary of distributional shifts in high-dimensional semantic spaces.

B.3. Additional Details on the Rephrasing of Tweets

For the rephrasing of Tweets, we used OpenAI's "gpt-4o-mini-2024-07-18" model. Each time, the model was asked to rephrase a single Tweet using the following prompt:

> "Your task is to rephrase a highly toxic Tweet and write a less toxic version of it while aiming to make minimal changes to the original Tweet. It's crucial to preserve the original wording, content, style, and tone in the Tweet. Keep the Twitter special elements such as RT and XXX unchanged. Example: Original: 'some_user The system is so fucked up. What's sad is they can do wtf they want.' Rephrased: 'some_user The system is so messed up. What's sad is they can do whatever they want.' Please respond in JSON format with the key 'RephrasedText'. Here is the Tweet to rephrase:"

B.4. Additional Details on Engagement Prediction

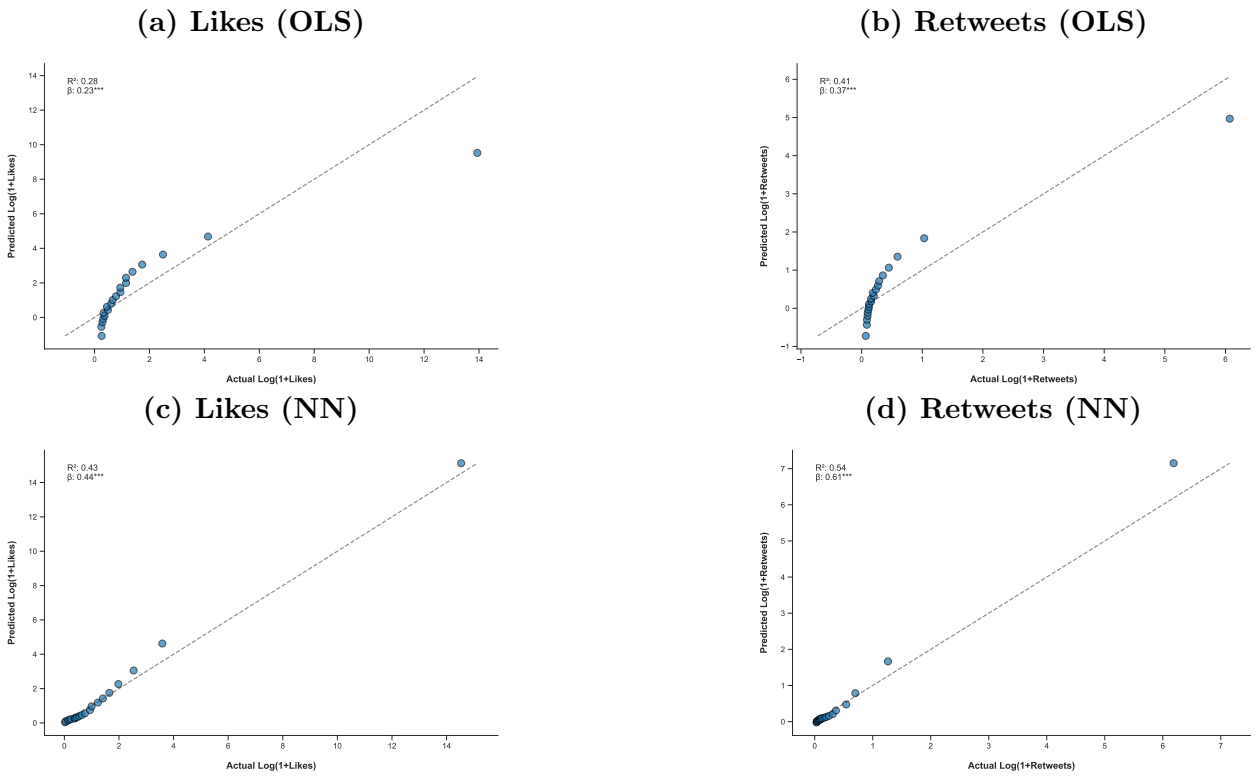
We employ two models for the engagement predictions: 1) a linear regression and 2) a neural network with three hidden layers. All models are trained on a sample of 1 Million political

Tweets, with engagement metrics winsorized at the 99$^{\text{th}}$ percentile and log transformed with one added to account for the heavy-tailed distribution of social media engagement.[24]

The neural network is trained with three fully connected hidden layers of dimensions 256, 64, and 32, respectively. The model uses Rectified Linear Unit (ReLU) activation functions and includes dropout layers with a rate of 0.1 to prevent overfitting. We experimented with deeper architectures by adding additional hidden layers, but found that increasing model complexity did not noticeably improve performance. The model was trained for 10 epochs with a batch size of 256 and an Adam optimizer with a learning rate of 0.001.

We evaluate the predictive power of our models on a holdout sample of 100,000 Tweets. Figure B.2 shows binscatter plots comparing the actual engagement (x-axis) and the predicted engagement (y-axis) for this holdout set. We find that both models achieve a high correlation between predicted and actual engagement. The models achieve a high R^2 of 0.41 for retweets and 0.28 for likes. Both plots indicate that our models effectively capture the relationship between semantic content and user engagement.

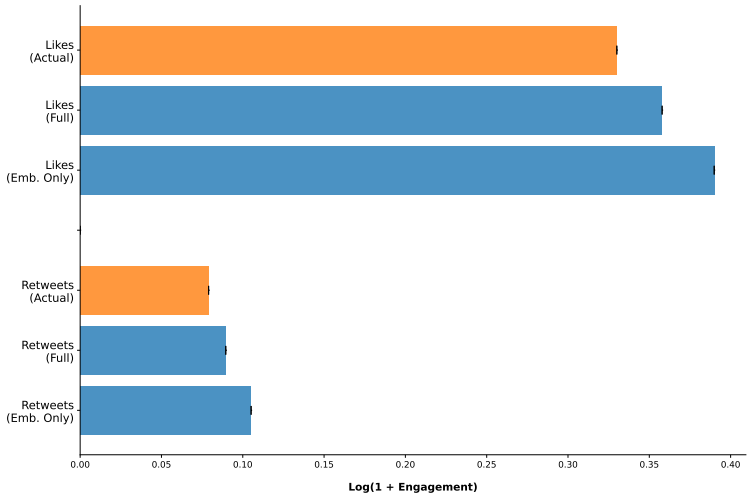
Figure B.2: Engagement Prediction Performance: OLS vs. Neural Network



Notes: The figure displays binscatter plots comparing actual vs. predicted log-engagement for a holdout sample of 100,000 tweets. Panels (a) and (b) show results for the linear regression (OLS) model. Panels (c) and (d) show results for a neural network model featuring three hidden layers. Both models are trained on tweet embeddings and incorporate user fixed effects.

[24]The results are very similar without winsorization.

Figure B.3: Predicted Impact of Rephrasing on Engagement (Neural Network)

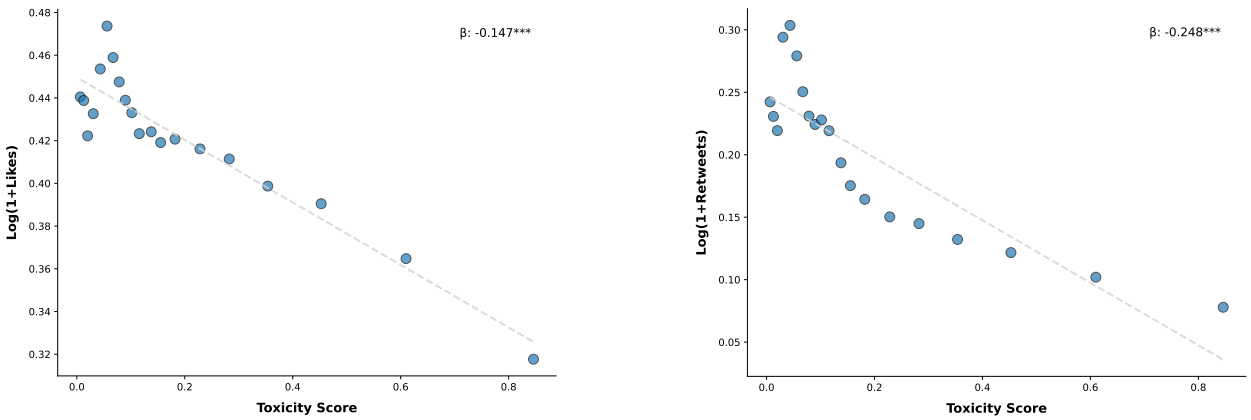


Notes: The figure displays the average predicted engagement for original and rephrased Tweets using a neural network model. Outcomes are transformed using $log(y + 1)$ transformation.

Figure B.4: Raw Engagement vs. Toxicity

(a) Likes

(b) Retweets



Notes: The figure shows a binscatter plot of engagement as a function of the toxicity scores as created by the Perspective API. Engagement is transformed using log with one added inside.

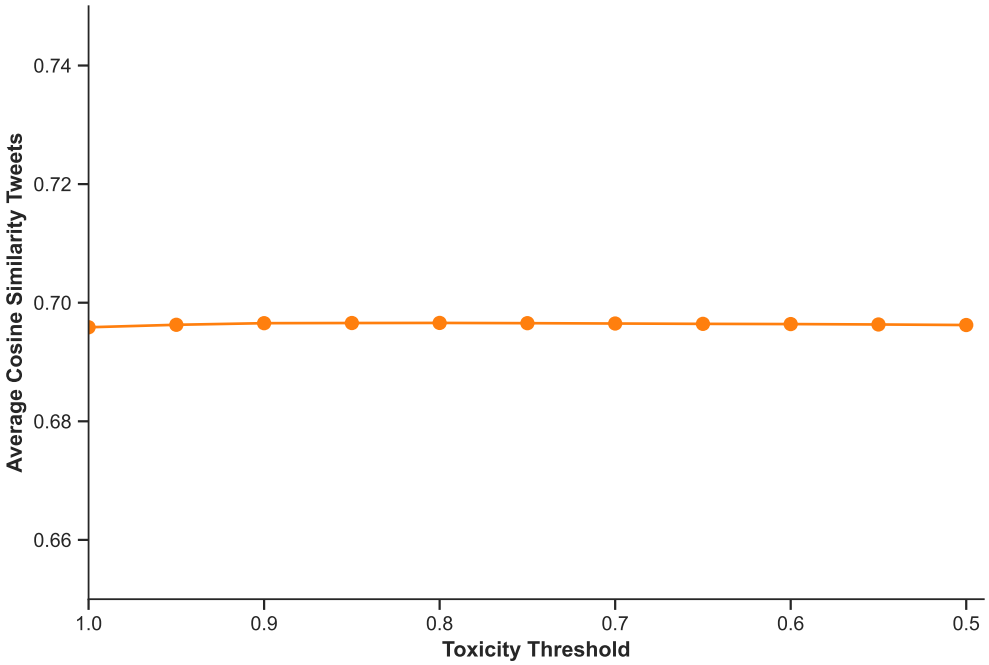
C Additional Results

The following section describes additional results and robustness checks. We begin by showing that the average similarity of Tweets is not affected by content moderation and provide additional benchmarks based on the removal of topics. Further, we repeated our main analysis using alternative 1) embedding models, 2) toxicity scores, 3) engagement weighting, and 4) samples. We also demonstrate that our results are not driven by Tweets belonging to the "Insult" topic as created by the Top2Vec topic model. Lastly, we provide evidence from an alternative method to account for the toxicity dimension of the embedding space.

Removal of Toxic Content and Cosine Similarity

In Figure C.1, we report the average cosine similarity of Tweets after sequentially removing those above varying toxicity thresholds. Given the very high computational requirements of the similarity calculation, we conduct this analysis for a random subset of 500,000 Tweets. The results show that average similarity remains largely unaffected by the exclusion of toxic content. In unreported analyses, we confirm that this pattern holds when we focus on a subset of Tweets with the highest and lowest pairwise similarity.

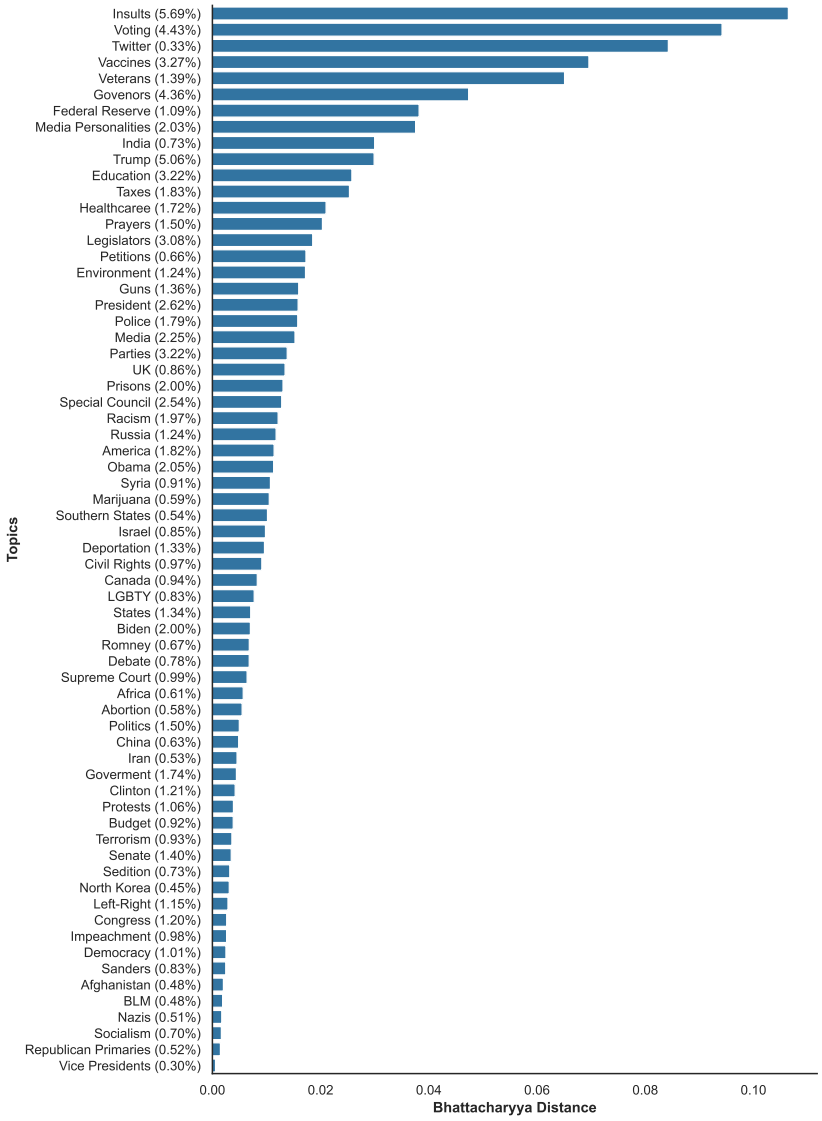
Figure C.1: Removal of Toxic Content and Average Cosine Similarity



Notes: The figure shows the average cosine similarity of Tweets after excluding Tweets with a toxicity score exceeding the threshold shown on the x-axis.

Additional Benchmarks

Figure C.2: BCD: Removal of Individual Topics

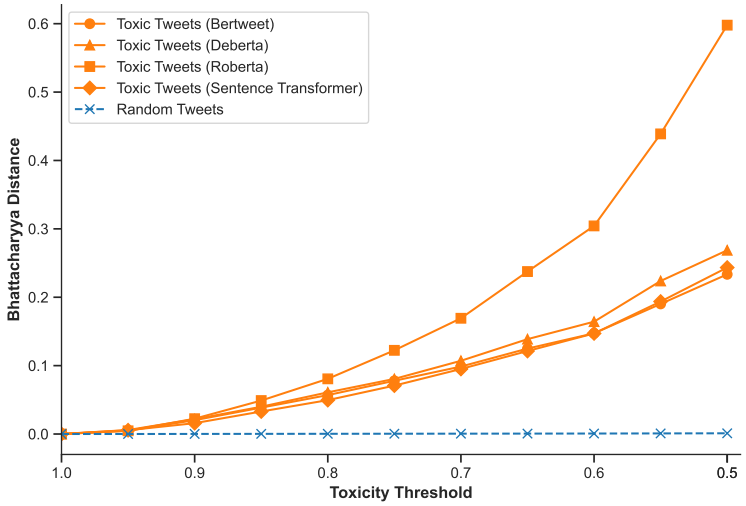


Notes: The figure plots the BCD obtained when each of the topics created by the Top2Vec topic model is removed from the data one by one. We also report the topic's share among all Tweets in brackets after the topic label.

Alternative Embeddings

As a robustness check, we reproduce our findings using the alternative embeddings described in Appendix B.1. This test rules out that our findings are driven by the particularities of the specific transformer model we have chosen, even though BERTweet is one of the standard choices for the analysis of English-speaking Twitter data. For this robustness exercise, we created new embeddings based on the RoBERTa, DeBERTa, and DistilBert-Base-Multilingual-Cased-V2 models and reconstructed the BCD based on these embeddings. The findings in Figure C.3 highlight that the findings are remarkably similar not only with regards to the overall patterns but also the magnitudes of the content-moderation-induced increases in the BCD.

Figure C.3: Robustness: Alternative Embeddings



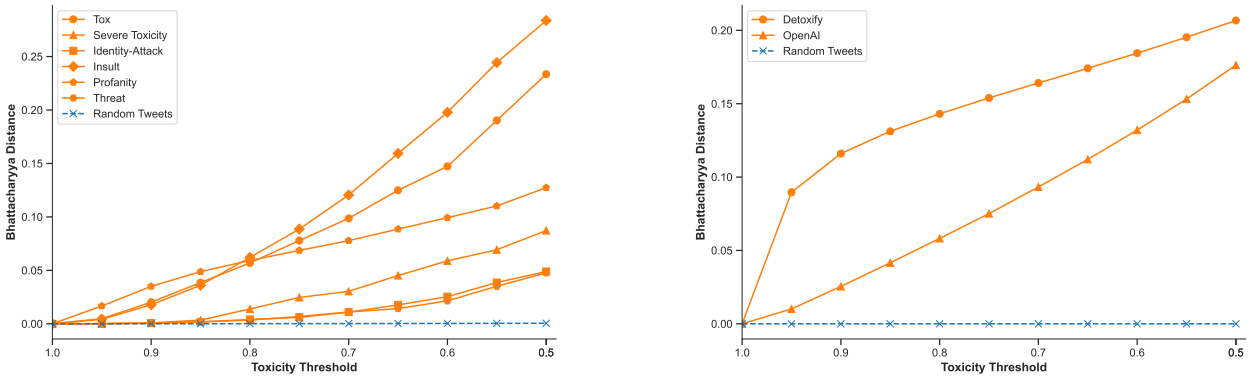
Notes: The figure shows the BCD computed using embeddings generated by DeBerta and RoBerta, and DistilBert-Base-Multilingual-Cased-V2 after the exclusion of toxic and random Tweets from the data.

Alternative Toxicity Measures

As another robustness check, we use the alternative Toxicity dimensions from the Perspectives API (see Figure C.4a) as well as other toxicity scores based on the classifiers from Detoxify (Hanu and Unitary team, 2020) or OpenAI's Moderation API (OpenAI, 2024) (see Figure C.4b). We find that the BCD is increasing independently of the toxicity measure that we are using.

Figure C.4: Robustness: Alternative Toxicity Measures

(a) Alternative Toxicity Dimension **(b) Detoxify and OpenAI Scores**

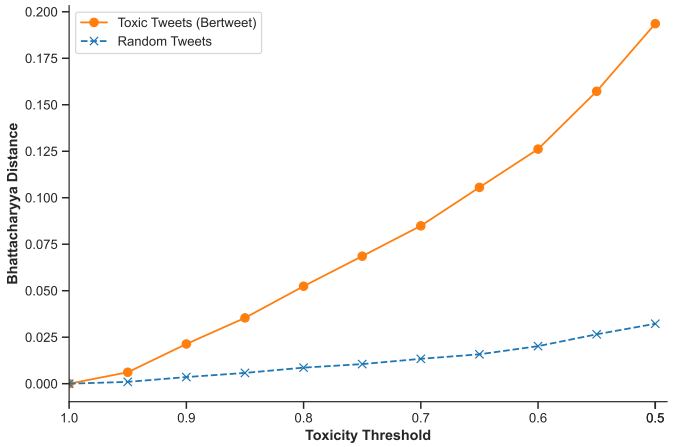


Notes: Panel (a) shows the BCD following the exclusion of Tweets characterized by high levels of Severe Toxicity, Profanity, and Insult as identified by the Perspective API. Conversely, Panel (b) shows this measure after the removal of toxic Tweets, using toxicity scores generated by Detoxify and the OpenAI Moderation API.

Engagement Weigting of Content

As an additional robustness test, we weight the Tweets in our data by their engagement as measured by the number of Retweets when calculating the BCD. This allows us to arguably better account for the frequency with which users would encounter the toxic tweets. The results from this analysis are shown in Figure C.5. We find that weighting Tweets by their engagement has no bearing on our results and leads to very similar magnitudes for the BCD.

Figure C.5: Robustness: Engagement Weighting



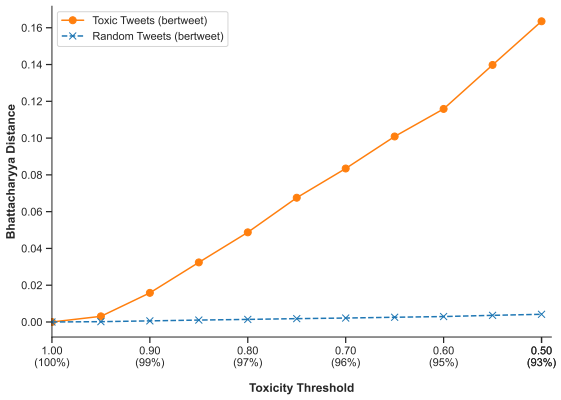
Notes: The figure shows the BCD after excluding Tweets with a toxicity score exceeding the threshold shown on the x-axis. We weight Tweets by the number of retweets when calculating the BCD. The blue line illustrates the BCD when an equivalent number of Tweets is excluded from the dataset at random.

Alternative Samples

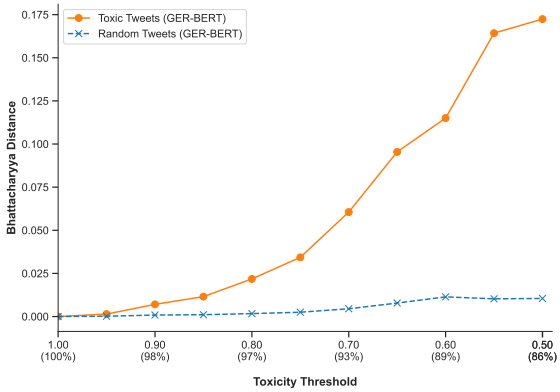
As a last robustness check, we repeat our main analysis based on three alternative samples of Tweets. First, we use 1 million randomly drawn Tweets from our English data without filtering for political content. Second, we use 1 million randomly drawn Tweets from a sample of politically interested German Twitter users. Third, we use 1 million randomly drawn Tweets from a sample of politically interested Italian Twitter users. To create embeddings for the German and Italian samples, we use the "bert-base-german-uncased" and "twitter-xlm-roberta-base" models, respectively. Toxicity in all cases was coded using the Perspectives API. For each of these samples, we repeat the analysis from Figure 1. The results are presented in Figure C.6. We find that independent sample content moderation appears to introduce distortions to the semantic space as measured by the BCD.

Figure C.6: Bhattacharyya Distance Alternative Samples

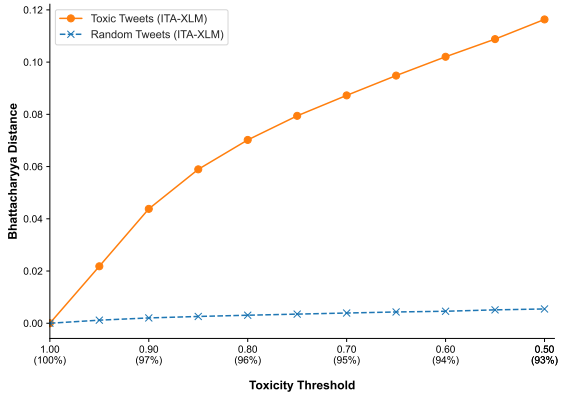
(a) Non-Political Tweets



(b) German Tweets



(c) Italian Tweets

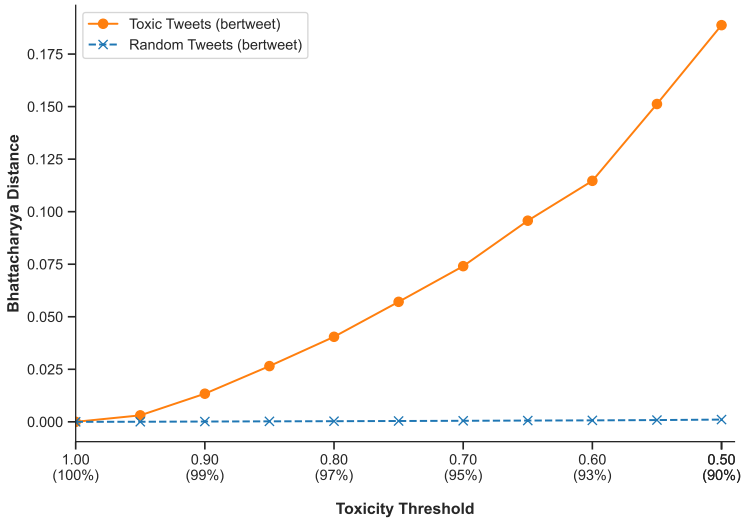


Notes: The figure shows the BCD after the exclusion of toxic and random Tweets from the data. Panel (a) uses a sample of 1 million English Tweets without filtering for political content, Panel (b) uses a sample of 1 million German Tweets, and Panel (c) uses a sample of 1 million Italian Tweets.

Excluding Tweets from the "Insult" Topic

The figure below reproduces Figure 1 when we remove all Tweets belonging to the "Insult" topic in advance. The results are virtually identical, suggesting that content moderation also introduces distortions of the semantic space beyond simply removing uncivil content.

Figure C.7: Content Distortions and Removal of Toxic Content (No Insults Topic)

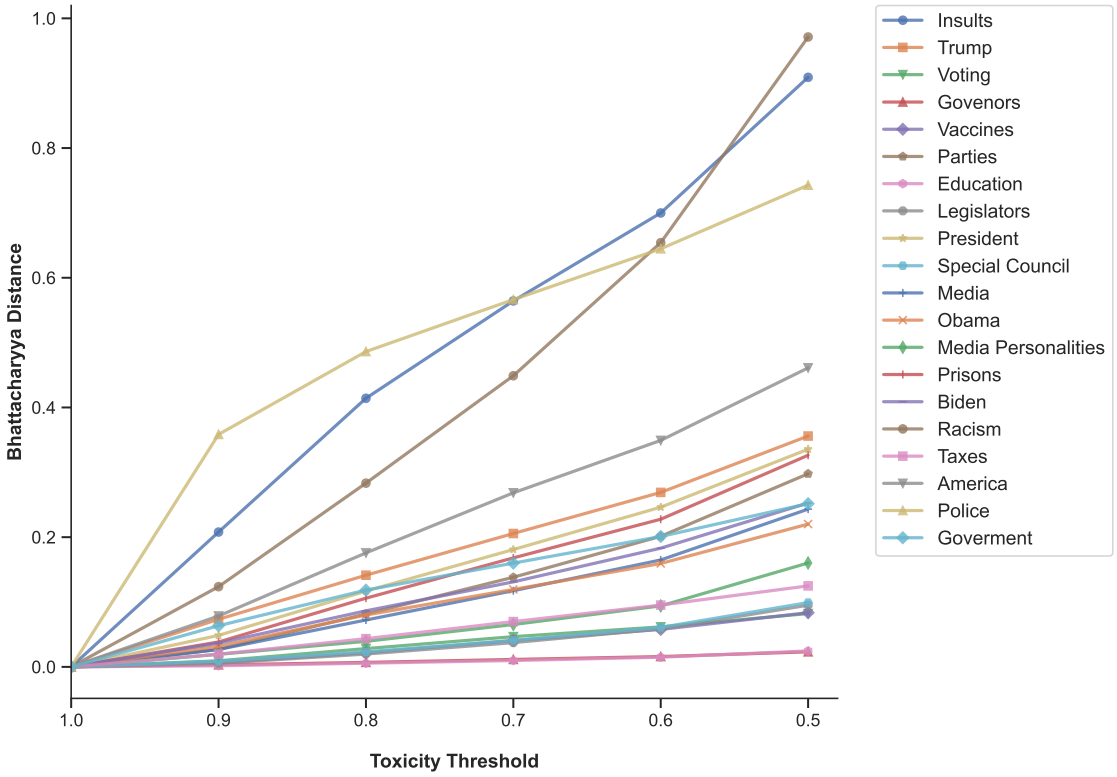


Notes: The figure shows the BCD after excluding Tweets with a toxicity score exceeding the threshold shown on the x-axis. For this figure, we do not consider Tweets belonging to the "Insult" topic. The blue line illustrates the BCD when an equivalent number of Tweets is excluded from the dataset at random. The percentages in parentheses on the x-axis represent the proportion of Tweets retained relative to the original sample size.

Additional Evidence Rephrasing

Figure C.8 shows the degree of semantic distortion, measured by Bhattacharyya distance, resulting from replacing toxic tweets with their rephrased versions. The results highlight a trade-off between toxicity reduction and semantic preservation that varies by topic. Socially polarized topics with prevalent toxic content (see Figure A.4)(e.g., Racism, Police) suffer the greatest distortion, implying that toxicity is embedded in these specific debates. In contrast, broad political discourse (e.g., Trump, Government) shows moderate distortion, while policy-specific topics (e.g., Education, Voting) are largely unaffected. This suggests that while automated rephrasing is a viable strategy for policy debates, it poses a higher risk of semantic distortion for sensitive social discourses.

Figure C.8: Rephrasing Distortion across Top Topics



Notes: The figure shows the Bhattacharyya distance resulting from rephrasing tweets when conducted within the 20 most frequent topics as created by the Top2Vec topic model. The x-axis indicates the toxicity scores above which Tweets were rephrased.

Alternative Method to Account for the Toxicity Dimension of the Embeddings

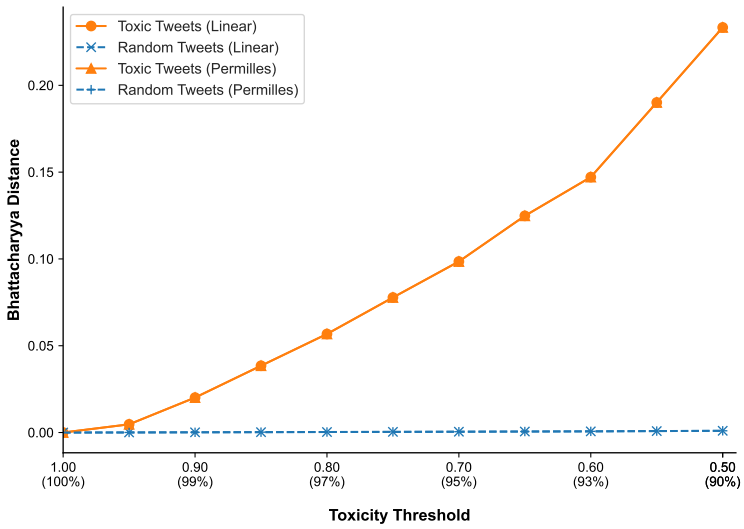
Our alternative approach for the orthogonalization of the embeddings with respect to the toxicity scores uses linear regressions of the following form:

$$x_d = \alpha + \mathbf{Tox}'\beta + \epsilon_d$$

where $x_d \in \mathbf{X}$ is one of the D embedding dimensions. **Tox** is a matrix containing 1000 indicators for the permilles of the toxicity distribution.[25] We estimate these regressions for all embedding dimensions $d \in D$ and replace the embedding dimensions with the regression residuals. The resulting residualized matrix $\tilde{\mathbf{X}}$ is orthogonal to the toxicity scores.

We then repeat our previous analysis based on the residualized embedding matrix $\tilde{\mathbf{X}}$. Figure C.9 visualizes the two different approaches to account for a Tweet's toxicity. More specifically, we use regressions and either residualize the embeddings using permilles or the linear toxicity score. We again find that removing the toxicity component from the embedding space has little impact on our findings. The overall patterns are close to our original results. The BCD sharply increases once toxic Tweets are removed from the data.

Figure C.9: Controlling for Toxicity



Notes: The figure shows the BCD, derived from toxicity-debiased Tweet embeddings, after the exclusion of toxic and random Tweets from the sample. Linear and Percentile Residualization adjust each dimension of Tweets' embeddings by using the residuals of the regression of the embedding values against the toxicity score and toxicity permilles of Tweets.

[25] We chose a non-parametric transformation of the toxicity scores to flexibly account for any potential non-linearities. As we show in a robustness check, the findings are almost identical if we instead residualize linearly with regard to the toxicity score.