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 how content moderation affects online discourse. We propose and validate a methodology for measuring the content-moderation-induced distortions in online discourse using text embeddings from computational linguistics. We test our measure on a representative dataset of 5 million US political Tweets and find that removing toxic Tweets distorts online content. This finding is consistent across different embedding models, toxicity metrics, and samples. Importantly, we demonstrate that content-moderation-induced distortions are not caused by the toxic language. Instead, we show that, as a side effect, content moderation shifts the mean and variance of the embedding space, distorting the topic composition of online content. Finally, we propose an alternative approach to content moderation that uses generative Large Language Models to rephrase toxic Tweets to preserve their salvageable content rather than removing them entirely. We demonstrate that this rephrasing strategy reduces toxicity while minimizing distortions in online content.

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

*We are grateful to Elliott Ash, Luca Bragherie, Sarah Eichmeyer, Ruben Enikolopov, Matthew Gentzkow, Rafael Jimenez, Horacio Larreguy, Debora Nozza, Jacob N. Shapiro, Arthur Spirling, 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. As 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 regulation of online platforms that demand 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). One upshot of this research is that algorithms are susceptible to false positives, often triggered by swear words or otherwise innocent words that happen to often be found in the context of hate speech (e.g., Attanasio et al., 2022). 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 criticism from politicians (e.g., Samples, 2019; Vogels et al., 2020; The Texas Tribune, 2022). As a result, online platforms face a dilemma of seemingly contradictory objectives, which they need to balance in their content moderation efforts. This dilemma of balancing the removal of inflammatory content and the preservation of plurality

[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)

of opinions is exacerbated by the fact that extensive disagreements 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.

The dilemma of content moderation is further complicated 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 measures exist to quantify the effect of content moderation-induced changes in online content. In this paper, we narrow this gap by putting forward and validating a measure for the distortions of online content. We formalize the idea of content distortions based on changes in the statistical moments (mean and variance) of the semantic space.

As a first step, we approximate the semantic space using text embeddings based on Transformer models (Vaswani et al., 2017). These embeddings represent texts as projections into 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. Embeddings have become the de-facto standard text representation in natural language processing (NLP). As of 2024, the original Vaswani et al. (2017) has received over 110,000 citations, and Transformer models have proven highly successful in countless 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 better than previous approaches. 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).

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 similarity 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-*

[2] In a recent paper, Solomon et al. (2024) argue that US partisan differences over what constitutes hate speech might be overstated.

agnostic, i.e., it does not involve any choice of which content is worth preserving.[3] Further, the measure is scalable to large datasets and entire platform ecosystems.

We validate the measure's potential on a representative sample of 5 Million US political Tweets. Our analysis proceeds in four steps. First, we show that removing toxic Tweets leads to measurable distortions in online content. Importantly, no such distortions occur if Tweets are removed at random. This result persists independent of the embedding model, toxicity score, or Twitter sample. To the best of our knowledge, our paper is the first to provide a measure to quantify this trade-off empirically. We can also show that toxicity-based content moderation distorts both the mean and the variance of the semantic space. Additionally, we find that the variance of the semantic space is reduced by increased content moderation.

Second, we show that the content moderation-induced distortions are not driven by the position of toxic language in the embedding space. To this end, we build on the literature on the debiasing of text embeddings (e.g., Bolukbasi et al., 2016), and create projections of the embeddings 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 our results remain virtually identical with these modified embeddings. This finding suggests that it is not the toxic language itself but rather that some topics that are disproportionally discussed in a toxic manner.

Third, we shed more light on the content moderation-induced distortions of online content using a Top2Vec topic model (Angelov, 2020). This allows us to characterize which topics are disproportionally affected by content moderation. Further, to benchmark our findings, we show that removing all Tweets with a Toxicity above a threshold of 0.8 produces distortions in online content comparable to removing 4 out of 67 topics from the data at random.

Lastly, we propose an alternative approach to content moderation that, instead of removing toxic content, rephrases Tweets in a way that strips them of their toxic language while leaving the core message intact. For this analysis, we use the language generation capabilities of large language models (LLMs) (e.g., Argyle et al., 2023). Using the rephrased Tweets, we can show that this alternative form of content moderation can measurably reduce toxicity while mitigating distortions to the semantic space.

Our results break new ground by proposing a measure of content distortion that allows us to quantify the inherent trade-off in content moderation. Given the fundamental importance of this trade-off, our measure is of immediate policy relevance when it comes to the moderation of toxic and fake news content. The economic theory of multitask models (Holmstrom and

[3]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.

Milgrom, 1991; Feltham and Xie, 1994) predicts that if principles have to choose between different objectives, only one of which is measurable, the effort will be focused on the measurable tasks. In other words, in the absence of readily available measures of content distortions, online platforms, and lawmakers are likely to put far greater emphasis on removing hateful content at the "cost" of distorting online content. Hence, our measure represents an essential puzzle piece in the debate on content moderation.

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).[4]. 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.[5] 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. We then finetune a BERTweet model (Nguyen et al., 2020) for the classification of political Tweets (see Appendix A for details). We use the tuned model to classify all Tweets as either political or apolitical.

We create two samples for our analysis. Our main analysis is based on a sample of 5 million randomly-drawn *political* Tweets.[6] 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); Lupo et al. (2024).

For our content moderation analysis, we assign three toxicity scores to each Tweet based on different models. Our baseline model is Google's Perspectives API (??, Per). The Perspectives API has become one of the standard tools for toxicity analysis and is used by several platforms for content moderation (e.g., NYT, 2016; Forbes, 2019; Delgado, 2019). For each Tweet, the API returns six scores measuring different toxicity dimensions (toxicity,

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

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

[6]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.

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 Toxicity scores.[7] 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 of these 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. Appendix Table A.3 provides examples of toxic Tweets in our data (Warning: The examples contain offensive language.)

2.2 Measuring Distortions to Online Content

As described in the introduction, our measure of content distortions is built around the idea of the semantic space. To build intuition, imagine that the semantic content of a text can be represented as a vector in 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 measure to what extent the removal of toxic content changes the mean and the variance of the semantic space. The measure will be small if the semantic space after the removal of toxic 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.

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).[8] 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.[9]

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

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

[8] We also provide robustness checks for embeddings created by the RoBerta and DeBerta models.

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

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 embeddings matrix $\mathbf{X}$. Our measure is based on the Bhattacharyya distance (BCD), a commonly used metric for comparing probability distributions. As the embeddings are approximately normally distributed (see Appendix Figure B.1), we use the formula for Bhattacharyya distance for multivariate normal distributions (e.g., Abou-Moustafa and Ferrie, 2012). 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:

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

where $\Sigma = \frac{\Sigma_1 + \Sigma_2}{2}$. The Bhattacharyya intuitively 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 $\mathbf{X}$.[10] 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.[11]

In this way, BCD provides a unidimensional measure that summarizes the overall distortions of the embeddings 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.

[10] 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 approximate is very close to the analytical solution of the variance-covariance matrix.

[11] 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.

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.[12]

3 Results

In the following, we use the BCD to establish four sets of results. First, we analyze how the content is distorted when we remove highly toxic Tweets from the data. This analysis allows us to simulate the effect of more stringent content moderation. Second, we demonstrate that the content moderation-induced distortions are not driven by the toxic language of the Tweets. Third, we investigate what underlies the changes in the BCD. Fourth, we propose an alternative approach to content moderation that reduces toxicity while preserving the content of Tweets.

3.1 Removal of Toxic Content and Distortions of Online Discourse

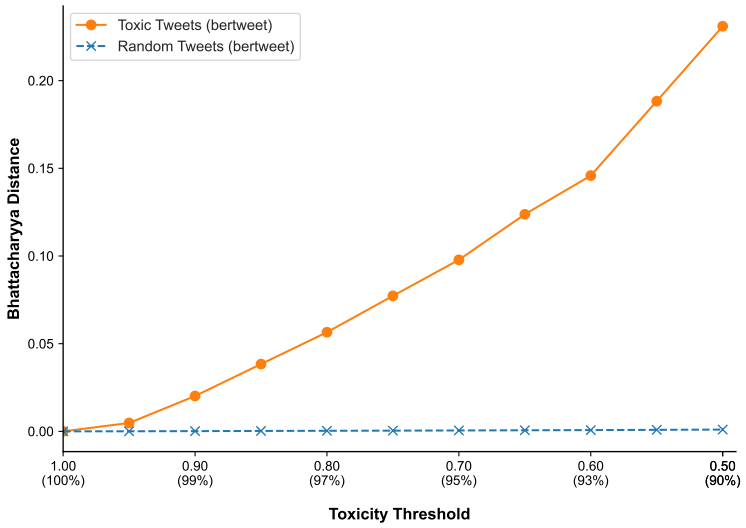
We simulate the effects of more stringent content moderation by removing toxic Tweets based on varying toxicity thresholds from our data. We then analyze if removing toxic Tweets leads to distortions of online content as measured by the BCD. We will compare these content-moderation-induced changes in the semantic space to changes in the BCD if we instead remove the same number of Tweets from the data at random. Note that this analysis intentionally abstracts from any form of behavioral reaction on the side of the users, as we are interested in the direct effects of content moderation. This analysis should provide 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).

The results from this 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 distortions

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

to the semantic space (orange line). The findings indicate that distortions 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 driven by the changing composition of online discourse due to content moderation. We interpret this result as evidence that removing toxic distorts online content as measured by 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.

Robustness

We conduct several robustness exercises for this finding. First, we reproduce our finding using alternative embeddings based on the widely-used RoBerta and DeBerta models (see Figure C.1). Second, 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.2). 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.3).

We discuss all of these robustness checks in greater detail in Appendix C. To summarize, none of these changes make 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.

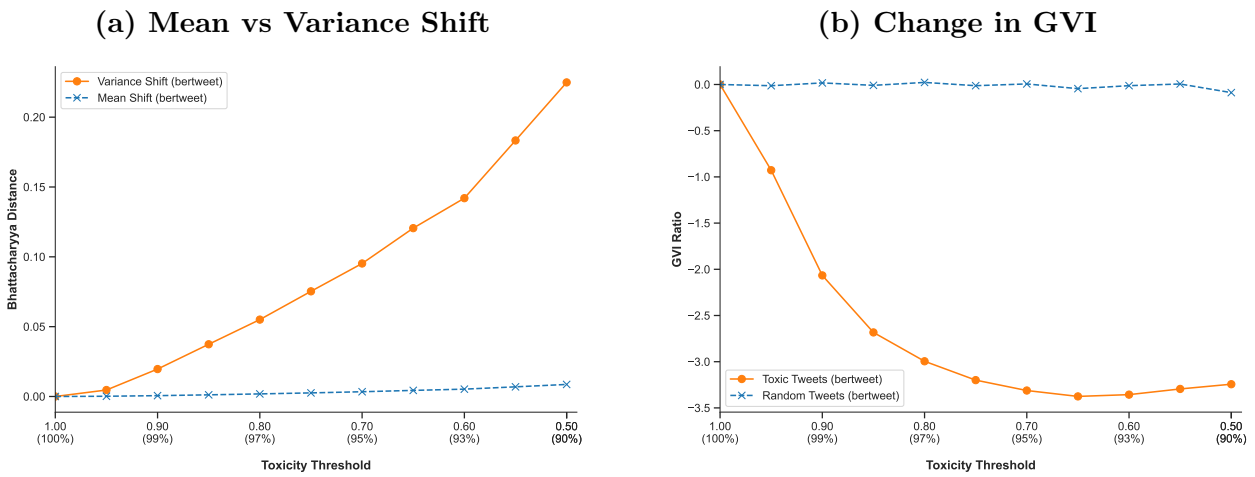
3.2 What Drives Distortions to Online Content

We next aim to shed light on the factors that drive the content-moderation-induced distortions of the semantic space we have documented. First, we investigate the individual components of the BCD and show that both the mean as well as the variance of the embedding space get distorted by content moderation. Second, we show that the increase in the BCD is not driven by the use of toxic language itself. Third, we provide further evidence for the content-moderation-induced distortions of the topics of online debate.

Decomposing the Bhatacharya Distance

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 to what extent 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 2. We observe that while both the first and the second component contribute to the increase in the BCD, the shift in the variance explains around 96% of the increase in the BCD, while the shift in the mean only explains 4%.

Figure 2: 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 2). 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.

Accounting for the Toxicity Dimension of the Embedding Space

The content moderation-induced changes in the semantic space could be driven by two factors. On the one hand, the toxic language itself could occupy a particular part of the semantic space. Removing toxic content would then mechanically increase the BCD, but this would not necessarily indicate a "real" distortions to online content. On the other hand, toxic Tweets might discuss specific issues and topics that are underrepresented online, independently of their toxicity.

We investigate this question by removing 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. Our baseline approach uses linear regressions of the following form:

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

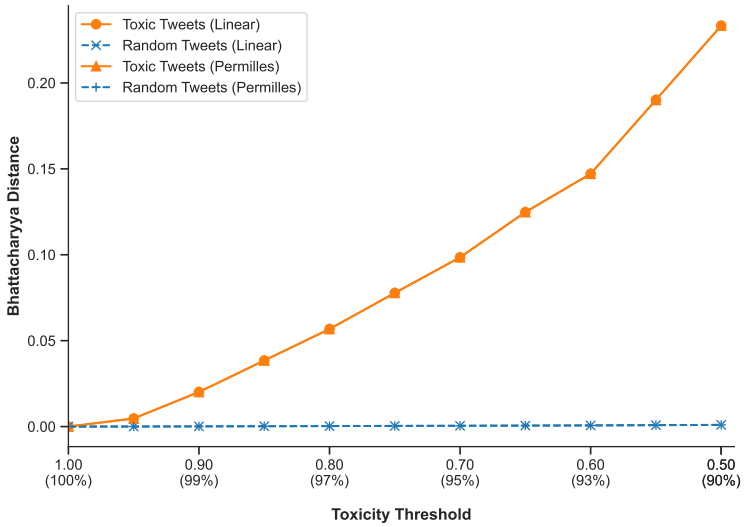
where $x_d \in \mathbf{X}$ is one of the D embedding dimensions. $\mathbf{Tox}$ is a matrix containing 1000 indicators for the permilles of the toxicity distribution.[13] We estimate these regressions for all embedding dimensions $d \in D$ and replace the embedding dimensions with the regression residuals. This leaves us with a residualized matrix $\tilde{\mathbf{X}}$, which is orthogonal 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 Tweets' content independent of their toxicity.

[13]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.

We then repeat our previous analysis based on the residualized embedding matrix $\tilde{\mathbf{X}}$. Figure 3 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 find that removing the toxicity component from the embedding space hardly makes any difference to our findings. The overall patterns are very close to our original results. The BCD sharply increases once toxic Tweets are removed from the data. This suggests that the observed changes in the BCD do not appear to be driven by the toxicity of the Tweets but rather by their content.

Figure 3: 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.

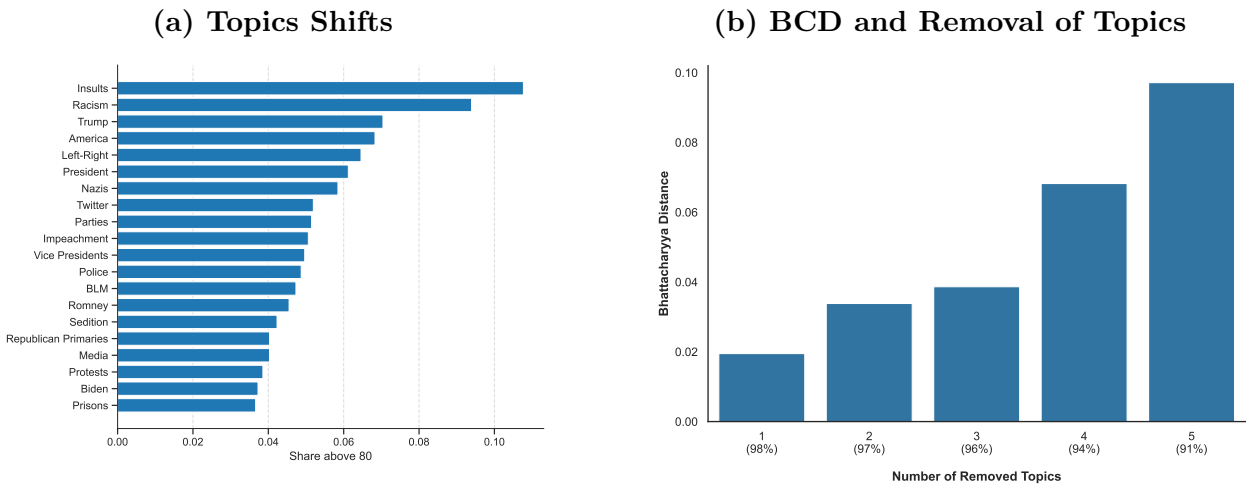
Content-Moderation-Induced Topic Shifts

As the last piece of evidence on the content-moderation-induced shifts in online content, we analyze how the topic composition of Tweets is shifting with increased content moderation. Additionally, we provide evidence that, in line with our original motivation, the removal of topics is associated with increases in the BCD. The latter result also allows us to benchmark the content-moderation-induced increases in the BCD relative to the direct removal of topics from the embedding space.

In Panel (a) of Figure 4, we show the topics that would be affected most by content moderation. Perhaps unsurprisingly, we find that Tweets containing insults against political

figures are most affected by content moderation.[14] Moreover, we also find that Tweets discussing the topics of "Racism", "Trump", or the "Black Lives Matter (BLM)" movement are heavily affected by content moderation, thereby distorting the topic composition of online debate. It is also important to highlight that our results do not hinge on shifts in any one topic, i.e., the results are very virtually identical if we do not consider Tweets containing insults (see Appendix Figure C.4). Further, we can show that content moderation distorts the semantic space even within topics (unreported).

Figure 4: Content Moderation and Topic Shifts



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

In Panel (b), we lastly show that removing topics leads to increasing changes in the BCD. For this figure, we calculate the BCD for the cases in which we remove 1, 2, 3, 4, or 5 topics from the data at random. For this exercise, 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.

We find that the BCD increases progressively with the removal of additional topics. This exercise also allows us to benchmark 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 ca. 0.7. It is also worth highlighting that removing four deletes twice as

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

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 removing content based on toxicity.

Taken stock, the previous exercises suggest that the reduction in BCD are not driven by the toxic language itself, but rather the fact that removing Tweets with toxic language as a byproduct 1) shifts the mean of the embeddings space, 2) reduces the variance of the semantic space, and 3) removes topics from the online debate. In the last part of the paper, we propose an alternative approach to mitigate some of the content-moderation-induced distortions.

3.3 An Alternative Approach to Content Moderation

We propose an alternative approach to tackle the issue of online toxicity that has the potential to mitigate some of the distortions in online discourse. Instead of removing toxic content outright, content moderators could instead use the language generation capabilities of large language models and rephrase the message of Tweets using less toxic language. This makes it possible to reduce the toxicity of online content while keeping the original content as far as possible intact.[15]

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. Our argument rather is that there exist significant amounts of content that violate the content standards of social media platforms (e.g., grave insults) but, at the same time, might contain political expression or information that might be 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).

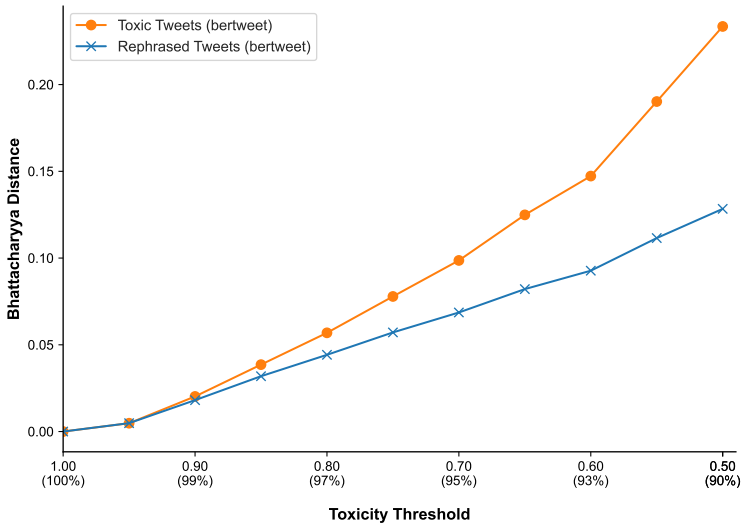
We test the potential of such an approach using GPT4o-mini to rephrase the toxic Tweets. We provide additional details on the rephrasing prompt in Appendix B.2. We also show examples of rephrased Tweets in Table A.3. Overall, GPT4o-mini can rephrase Tweets in a significantly more civil manner while keeping the main message intact. Rephrasing reduces the toxicity of Tweets above the 0.7 threshold to 0.26.

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. 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 than outright

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

removal. Only once we consider content that contains other content independent of the toxic language, rephrasing achieves its intended goal of preserving the original content. This suggests that such an approach can reduce the toxicity of online content with fewer distortions to online content.

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.

4 Conclusion

This paper proposes and validates a new methodology to measure content-moderation-induced distortions to online content. This new methodology enables us to, for the first time, document the impact of the removal of 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 measure captures the distortions in the embedding space of the content that is 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, nor is it the goal of this paper, provide a universal measure of online content and free speech. Given the complex and multifaceted nature of the concept of free speech that encompasses different philosophical and legal standpoints (see for example Warburton, 2009, for a review), no one measure could ever capture all facets of

online discourse. In the same way, despite extensive efforts by the research community, no measure of hate speech could ever capture the importance of cultural context and changing societal norms (e.g., Brown, 2017).

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

References

Perspective api. https://www.perspectiveapi.com/.

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

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.

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.

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*.

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*.

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.

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.

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.

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`.

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

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].

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*.

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.

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.

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*.

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

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.

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*.

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].

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*.

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*.

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*.

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.

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.

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.

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.

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.

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

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.

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.

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. Association for Computational Linguistics.

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*.

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

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 affords us the opportunity 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 an insightful 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.

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 0 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				

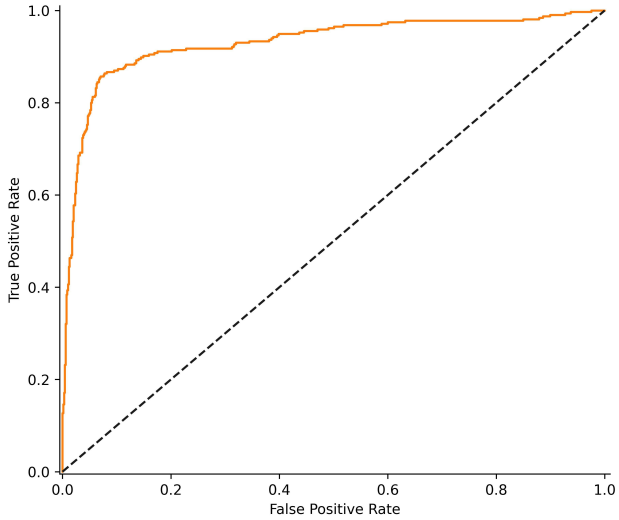
Notes: This table reports the confusion matrix for the political Tweet classifier.

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.

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



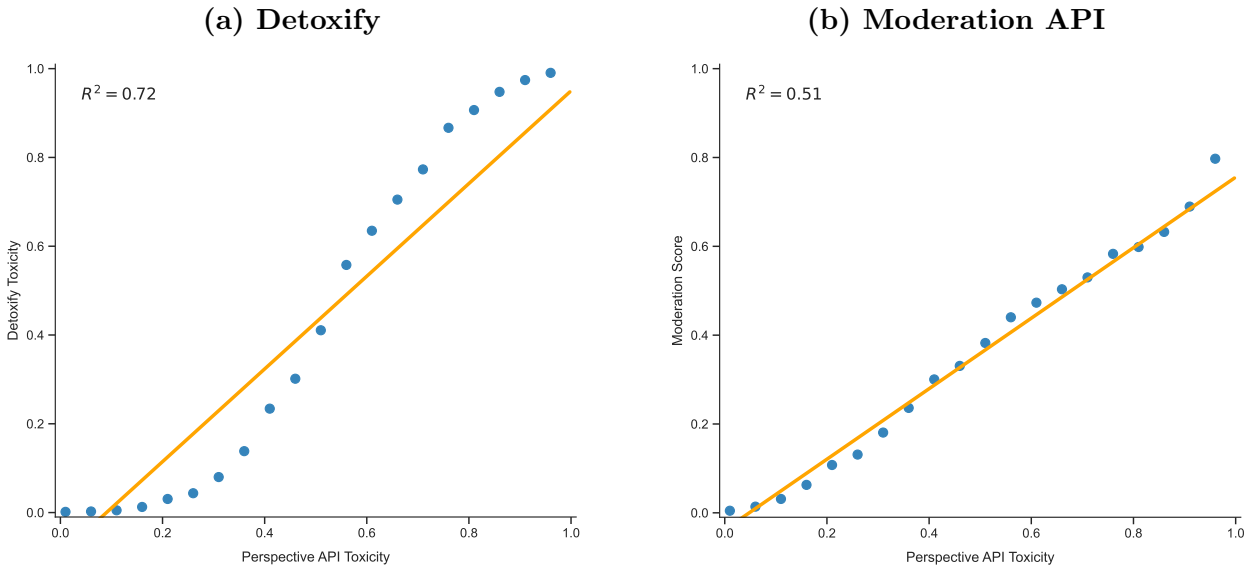
Notes: The figure reports the ROC curve for the political Tweet classifier.

The scores can be approximately interpreted as the probability of a randomly chosen user classifying 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, unsurprisingly, is 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 A.3. 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.[16]

[16]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.

Additional Details on Topic Model

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 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 creates a unique assignment between texts and topics. The topic words are then

derived based on the word vectors that are closest to the topic centroid. We report the most important topics words as well as well as the assigned topic labels in the table below.

Table A.3: Examples Toxic Tweets

Tweet Text	Toxicity Scores	Rephrased Tweet
What is wrong with people!! I don't get what burning the US flag proves. Side note: moron in the cowboy hat is not a real cowboy	(0.71, 0.97, 0.77)	What is wrong with people!! I don't get what burning the US flag proves. Side note: the person in the cowboy hat is not a real cowboy
Again, to the 60+ million so-called Americans who voted for this racist hateful bullshit: FUCK YOU.	(0.96, 1.00, 0.95)	Again, to the 60+ million so-called Americans who voted for this racist hateful nonsense: I strongly disagree with you.
@user You vote for Stupid, you get Stupid. No Republican has any redeeming social values any longer.	(0.95, 1.00, 0.91)	Hey @user You vote for unwise choices, you get unwise outcomes. No Republican seems to have redeeming social values any longer.
@user @JoeBiden You didn't say a damn thing about it when he was having his little rallies did you, you hypocritical bitch!	(0.97, 1.00, 0.91)	@user @JoeBiden You didn't say a thing about it when he was having his little rallies did you, you hypocritical person!
@SenWarren @USProgressives Democrats will lie, cheat, and steal to get your crybaby way. It's disgusting. Go ahead and do it. The GOP will add that many more. It's that tit for that thing to keep you crooks in your place.	(0.77, 0.81, 0.95)	@SenWarren @USProgressives Democrats will bend the truth and manipulate to get their way. It's disappointing. Go ahead and do it. The GOP will respond in kind. It's that back and forth to keep you in check.
@realDonaldTrump We need to stop having school shootings how bout you focus on that actually try to make a difference and take Obama dick out your mouth	(0.93, 0.96, 0.97)	@realDonaldTrump We need to stop having school shootings. How about you focus on that, actually try to make a difference, and move on from the past?
If you're a Christian offended by foul language but not the rejection of refugees, you're not a fucking Christian.	(0.95, 0.98, 0.73)	If you're a Christian offended by foul language but not the rejection of refugees, you're not a true Christian.

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

Table A.4: 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.

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), and DeBERTa (He et al., 2021). These models all built 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

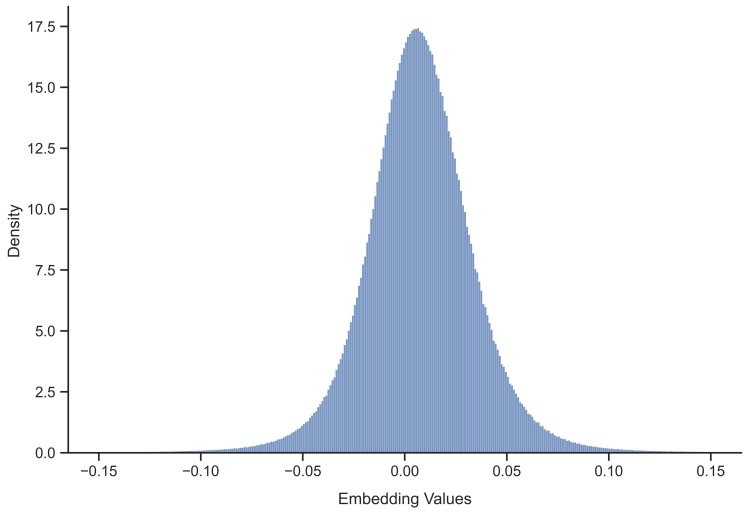
In the rapidly expanding field of Natural Language Processing (NLP), the inception of BERT (Bidirectional Encoder Representations from Transformers) and its subsequent iterations have marked a significant milestone. Introduced by (Devlin et al., 2019), BERT leverages an architecture known as Transformers (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 utilizing 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.

B.2. 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:"

C Additional Results

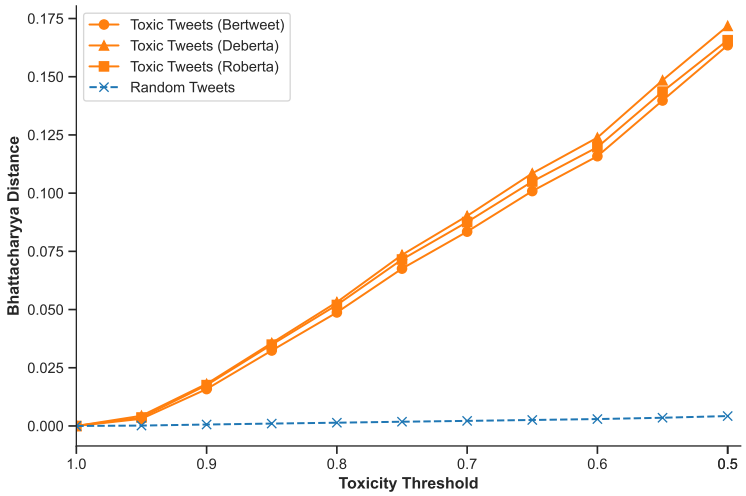
The following section describes the results of additional robustness checks. Specifically, we repeated our main analysis using alternative 1) embedding models, 2) toxicity scores, and 3)

samples. We also show that our results are not driven by Tweets belonging to the "Insult" topic as created by the Top2Vec topic model.

Alternative Embeddings

First, 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 and DeBERTa models and reconstructed the BCD based on these embeddings. The findings in Figure C.1 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.1: Robustness: Alternative Embeddings



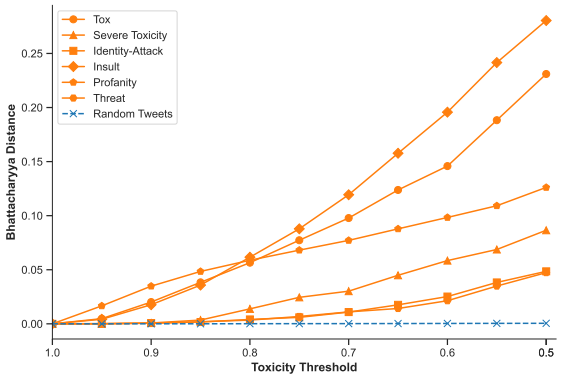
Notes: The figure shows the BCD computed using embeddings generated by DeBerta and RoBerta, after the exclusion of toxic and random Tweets from the data.

Alternative Toxicity Measures

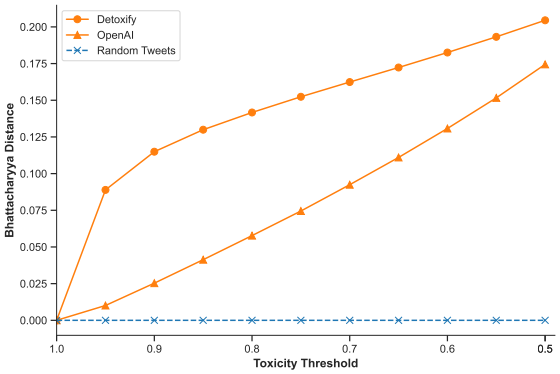
As a second robustness check, we use the alternative Toxicity dimensions from the Perspectives API (see Figure C.2a) as well as other toxicity scores based on the classifiers from Detoxify (Hanu and Unitary team, 2020) or OpenAI's Moderation API (OpenAI, 2024) (see Figure C.2b). We find that the BCD is increasing independently of the toxicity measure that we are using.

Figure C.2: 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.

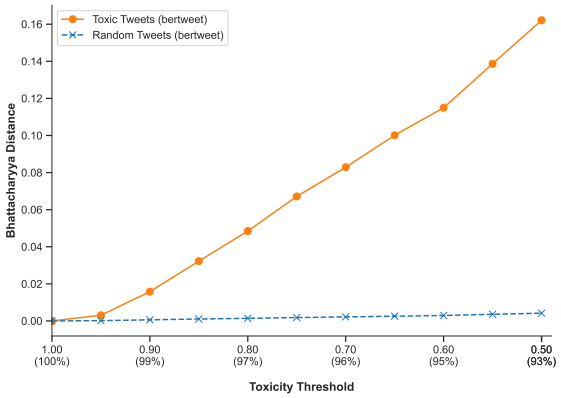
Alternative Samples

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 "German BERT" and "XLM-RoBERTa" models, respectively. Toxicity in all cases was coded using the Perspectives API.

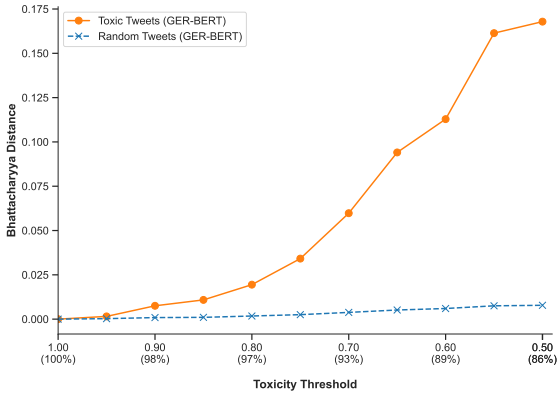
For each of these samples, we repeat the analysis from Figure 1. The results are presented in Figure C.3. We find that independent sample content moderation appears to introduce distortions to the semantic space as measured by the BCD.

Figure C.3: Bhattacharyya Distance Alternative Samples

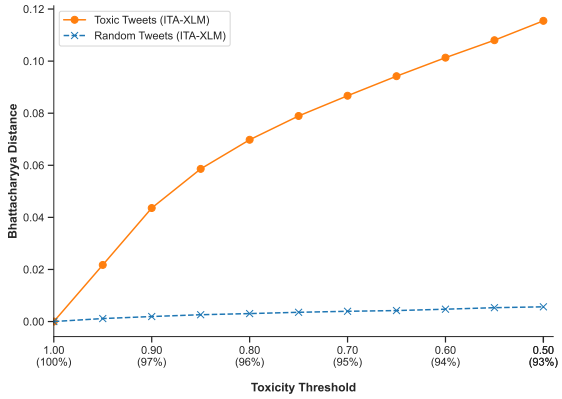
(a) Non-Political Tweets



(b) German Tweets



(c) Italian Tweets

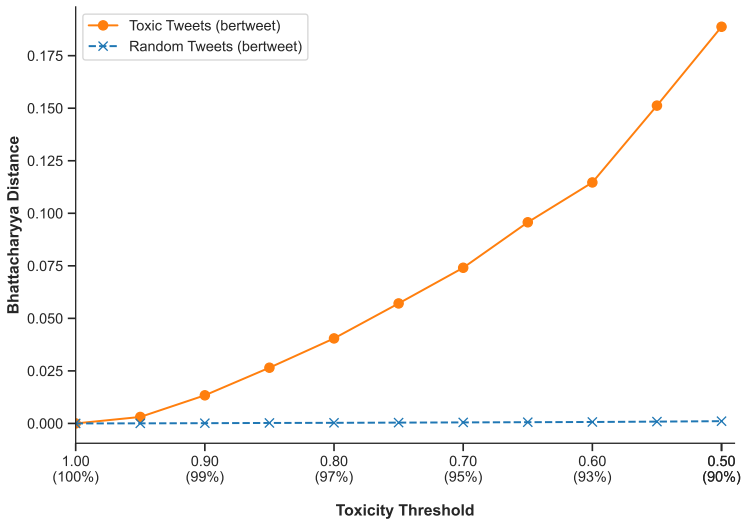


Notes: The figure shows the BCD after the exclusion of toxic and random Tweets from the data. Panel (a) uses a sample of 1 million English Tweets without filtering for political content, Panel (b) uses a sample of 1 million German Tweets, and Panel (c) uses a sample of 1 million Italian Tweets.

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 again that content moderation also introduces distortions of the semantic space over and above purely removing uncivil content.

Figure C.4: 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.