Measuring the impact of a new content moderation strategy on the incidence of LGBTQIA+ toxic language on Twitter

George Perrett[1], Christopher Praley[1], Amanda Morelli[1], Xiaoxuan Zhang[1], Duyi Liu[1], Verónica Mesalles[2], Qianhe Zhou[1], Joseph Marlo[3], Bilal Waheed[4], Alex Chohlas-Wood[1], and Daphna Harel[1]

[1]Department of Applied Statistics, Social Science, and Humanities
Steinhardt School of Culture, Education, and Human Development
New York University

[2]Department of Applied Psychology
Steinhardt School of Culture, Education, and Human Development
New York University

[3]Lander Analytics

[4]Fund for Public Health

Author Note

Correspondence concerning this article should be addressed to George Perrett. Email: gp77@nyu.edu; 246 Greene Street, New York, NY, 10003.

ORCIDs:

Qianhe Zhou: 0009-0009-9057-860X

Christopher Praley: 0009-0002-9449-9924

George Perrett: 0000-0002-1930-2581

Verónica Mesalles: 0000-0002-4408-4950

Alex Chohlas-Wood: 0000-0002-8279-6270

Daphna Harel: 0000-0001-7015-5989

Abstract

Certain demographic groups can be the target of hateful language on social media platforms. Stakeholders—including social media platforms themselves—have attempted to limit the spread of hateful language through various approaches, including via crowdsourced reports. In July of 2013, Twitter replaced an obscure web-based abuse reporting form with a “Report Tweet” button native to mobile devices. This new content moderation policy was touted as an easier way to facilitate the ability for a user to report hateful tweets. Yet any impacts from this change in policy are not publicly known—including any impacts on hateful language toward the particularly vulnerable lesbian, gay, bisexual, transgender, intersex, asexual, and related (LGBTQIA+) communities. To study the impact of this policy change, we collected and analyzed a quasi-random sample of 8,513,876 tweets posted on Twitter between July 2012 and July 2014. To assess tweets at this scale, we leveraged large language models using a pair of prompts validated to identify the number of tweets during this period that appeared to contain toxic language directed at the LGBTQIA+ communities. Using both classical and Bayesian variations on interrupted time series, we found that the introduction of the “Report Tweet” button led to a roughly 50% decrease in the number of tweets containing hateful language referencing LGBTQIA+ identities within one year of the policy change. Our findings suggest that this content moderation approach improved the experience of LGBTQIA+ individuals on social media platforms, and our approach highlights the considerations when using large language models to understand social phenomena at scale, including on social media.

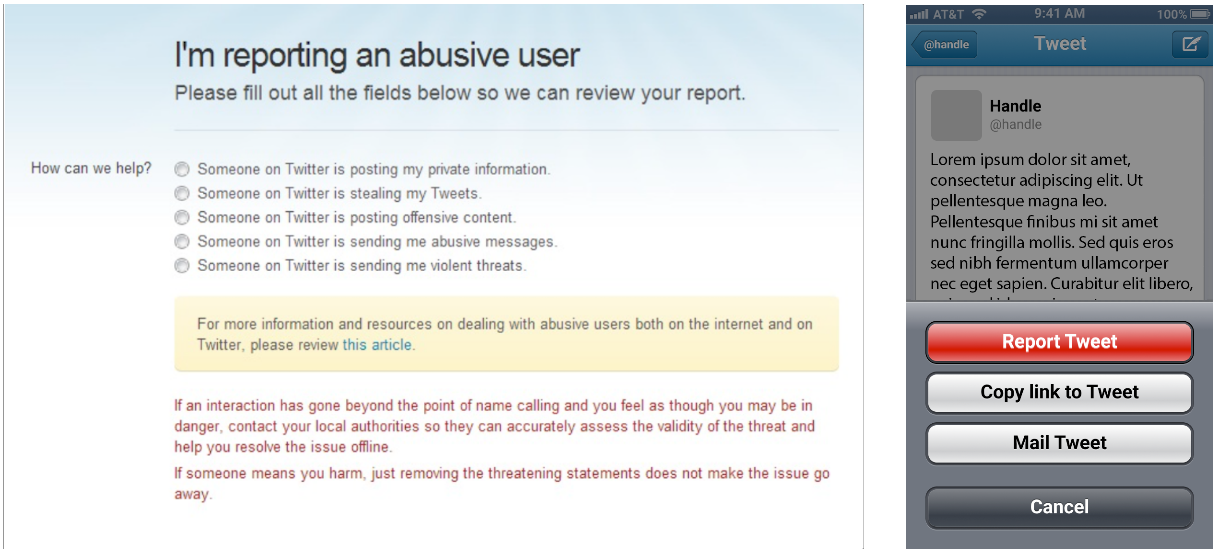
Keywords: Large language models; interrupted time series; social media; hate speech; LGBTQIA+

Measuring the impact of a new content moderation strategy on the incidence of LGBTQIA+ toxic language on Twitter

Introduction

Since the inception of social media, advocates, policymakers, and social media companies themselves have lobbied for a variety of approaches to limit the spread of hateful language online. Most strategies used by social media companies to limit hateful language follow a typical workflow: content is detected as potentially inappropriate, content is reviewed, a decision is made on whether standards are violated, and then the company choose whether to take action (Stackpole, 2022). When potential hateful language is identified, some platforms employ in-house or outsourced content moderators who review sensitive posts (Confessore & Frenkel, 2021). Other platforms use automated approaches that allow companies to review potentially offensive content and remove it. For example, YouTube currently uses automated systems to identify and remove videos that violate community guidelines (YouTube, 2024), while Instagram uses keyword filters to block hashtags associated with self-harm or misinformation (Instagram, 2019). Social media companies have also made use of user-facing interventions that preemptively discourage potential bad actors before they post inappropriate content (Stackpole, 2022).

The available evidence suggests that when companies decide to act, they can reduce the amount of hate on their platforms. For example, Reddit attempted to reduce the amount of hateful content on its platform by removing subreddits where hateful content was highly prevalent, leading to an up to 80% reduction in hateful language (Chandrasekharan et al., 2017). Using automated approaches to content moderation, Meta reduced the prevalence of hate speech on Facebook by 50% to a prevalence of only 0.05% of all posts viewed (Meta, 2021). This may be due to the faster response times that automated tools provide (Schneider & Rizoiu, 2023). Experiments have shown that nudging users to reconsider hateful language can result in short-term reductions


I'm reporting an abusive user

Please fill out all the fields below so we can review your report.

How can we help?

○ Someone on Twitter is posting my private information.
○ Someone on Twitter is stealing my Tweets.
○ Someone on Twitter is posting offensive content.
○ Someone on Twitter is sending me abusive messages.
○ Someone on Twitter is sending me violent threats.

For more information and resources on dealing with abusive users both on the internet and on Twitter, please review this article.

If an interaction has gone beyond the point of name calling and you feel as though you may be in danger, contact your local authorities so they can accurately assess the validity of the threat and help you resolve the issue offline.

If someone means you harm, just removing the threatening statements does not make the issue go away.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut pellentesque magna leo. Pellentesque finibus mi sit amet nunc fringilla mollis. Sed quis eros sed nibh fermentum ullamcorper nec eget sapien. Curabitur elit libero,

Figure 1

Screenshot of pre-July 2013 (left) and a mockup of the post-July 2013 (right) reporting systems.

of hate speech, on the order of 5-10% (Katsaros et al., 2022; Yildirim et al., 2023).

This study analyzes the impact of a moderation feature on iOS and mobile web browsers that Twitter (now X) first introduced on July 8th, 2013, in a change we refer to as the "Twitter 2013 policy change". The feature allowed mobile users to flag tweets for review by Twitter's content moderation team through an easily accessed "Report Tweet" button that was placed near where the "Report Spam" button had already been. Previously, users had to find and complete a separate web-based form to flag general patterns of abuse, which required searching and locating the form and flagging an entire user's account (Jeong, 2016). Figure 1 shows the former abuse reporting system alongside the new mobile-friendly button that was introduced by Twitter's 2013 policy change.

The initial release of this feature was limited to iOS and mobile web browsers. But an online petition—created following the targeted abuse of a British activist, and signed by over 100,000 individuals—began circulating on July 27, 2013 and led to this feature being expanded to Android and web browsers on July 29, 2013 (Twitter UK, 2013). Although these changes were first introduced in July of 2013, familiarity with this new feature likely grew over time as updates were rolled out to users' devices and users learned about the new reporting method. While Twitter's policies have changed since 2013, a similar "report post" button still exists on the current version of the website, where users can report a post for reasons such as hate, abuse and harassment, or violent speech. However, the question remains: was Twitter's 2013 policy change effective at reducing toxic language on the platform?

Hateful language on social media is disproportionately directed towards marginalized groups (ADL, 2021). In this paper, we focus on a specific subset of hateful language that concerns the lesbian, gay, bisexual, transgender, queer, intersex, asexual, and related (LGBTQIA+) communities. These groups are disproportionately targeted by those who post hateful language (Saha et al., 2019; Silva et al., 2016), and such hateful language can trigger particularly detrimental impacts on mental health (McConnell et al., 2017), as well as increased risks of real-world violence (Lupu et al., 2023; Tsesis, 2002). The concentration of toxic language against LGBTQIA+ communities, and the heightened consequences of such speech, make any reductions in this speech particularly impactful—and thus motivate our focus on toxic language of this type.

To understand whether Twitter's 2013 policy change impacted toxic language concerning LGBTQIA+ communities, we collected and analyzed a quasi-random sample of 8,513,876 tweets posted over a two-year period surrounding the policy change. Due to the scale of this dataset, we could not rely on traditional approaches that enlist human annotation of tweets as hateful or toxic language. Instead, we leveraged large language models (LLMs) to classify whether each tweet contained toxic language concerning LGBTQIA+ communities. We then estimate the impact of the policy change using both classical and Bayesian implementations of an interrupted time series (ITS) causal identification strategy. We estimate that Twitter's 2013 policy change resulted in roughly

50% reductions in English language hateful or toxic language concerning LGBTQIA+ identities within one year of the policy change.

In the following sections, we discuss the methods we used to collect our sample of tweets, the prompt engineering strategy used to elicit ratings from LLMs, how LLM ratings interact with our causal identification strategy, and our evaluation of Twitter's 2013 policy change on hateful language concerning LGBTQIA+ identities. Finally, we conclude with the implications of our findings and potential future directions for research.

Data collection

We collected our corpus of tweets through Twitter's application programming interface (API). First, we identified a quasi-random sample of Twitter users, and collected all available tweets for each randomly selected user. Collecting tweets in this way (via random users, instead of random tweets) reduces cross-user variation in our sample, likely increasing the stability of our estimates on LGBTQIA+ directed toxic language before and after Twitter's 2013 policy change. Similar to random digit dialing, we identified this quasi-random sample by generating random integers for potential user IDs. While most people refer to Twitter users by their self-chosen handle, all Twitter users have a numeric ID that is associated with their account.[1] We submitted each randomly generated potential ID to the Twitter API to check if a user existed with that ID. If so, we downloaded their tweet history, up to a limit of 3,000 tweets per user.[2] We also aimed to restrict our sample to U.S.-based users by relying on the user's self-described open-text location field and only considered accounts where the location matched a US city, state, or other locale.

After 25 days of querying the Twitter API in September and October of 2020, the resulting dataset consisted of approximately 92 million tweets from nearly 160,000 unique

[1] In recent years, there have been slight changes to how user IDs are generated on Twitter, but these changes occurred after the time period we analyze in this study.

[2] Approximately 8% of users in our sample had 3,000 or more tweets.

accounts.

Because we study a policy change that occurred in July of 2013, we restricted our analysis sample to tweets that were posted within a year of this policy change, between July 2012 and July 2014. We downloaded over 20.5 million tweets in this time frame. For computational feasibility, we took a simple random sample of a smaller final analysis set of 8,513,876 tweets from the total number in our corpus. This number was selected based on sample size calculations for the methods described below.

Measuring toxic language on Twitter

To understand the impact of Twitter's 2013 policy change on toxic language directed at the LGBTQIA+ communities, we need to know whether each tweet in our sample contains toxic language concerning LGBTQIA+ identities. Traditionally, this labeling task has been completed by either training research assistants to provide ratings or by hiring workers through crowdsourcing platforms like Amazon Mechanical Turk (MTurk). These approaches can have high inter-rater reliability and are often highly correlated with expert-provided ratings (Benoit et al., 2016), but would require a substantial budget to execute at the scale considered here. To reduce these costs, researchers have begun to examine automated approaches to labeling. Traditionally, such automated approaches have taken the ratings obtained from researchers or MTurk workers, and then using the resulting annotations to train machine learning classification models (Fortuna & Nunes, 2019).

The advent of LLMs has provided another novel automated strategy for labeling unstructured text at scale. These models—which have been trained on much of the public internet— can be used to surface nuance and context in natural language. This ability is a notably useful feature for subtle classification tasks like distinguishing between toxic and innocuous language, both of which may use similar words, but in very different ways (Gilardi et al., 2023; Rathje et al., 2024; Ziems et al., 2024). Using an LLM to classify tweets makes it feasible to create large sets of labels at relatively low cost, and avoids the

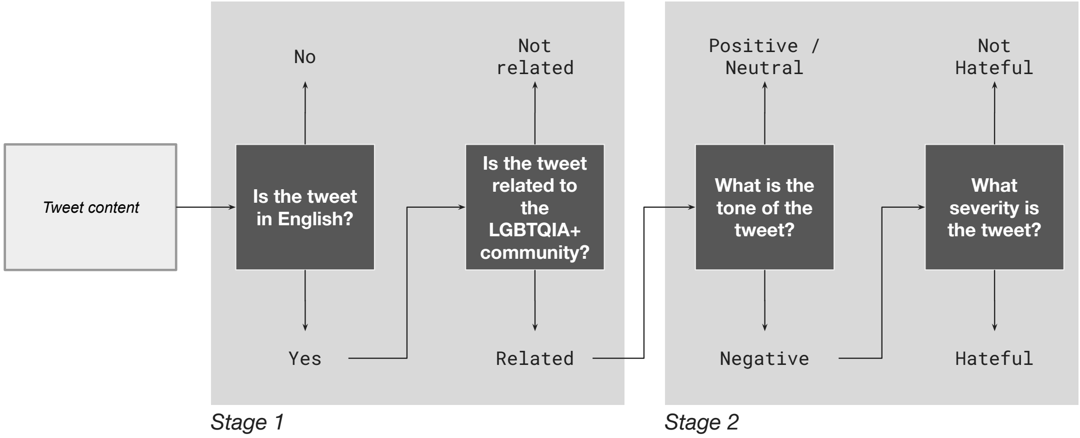


Figure 2
The four-step process we developed to identify whether a tweet contained English language toxic language concerning LGBTQIA+ identities. Tweets are gauged against each criterion in sequence, and a tweet must meet all four criteria before it is considered hateful language concerning LGBTQIA+ identities.

challenges of managing a team of human raters, the disadvantages of exposing human raters to toxic language, or training a custom model for classification from a smaller pool of crowdsourced labels.

Despite these advantages, using LLMs to label unstructured text introduces a different set of considerations, including choice of prompt and replicability. In the standard setup for LLM annotation, a researcher must create a set of instructions, known as a "prompt", that instruct the LLM on how to approach the task. Critically, the performance of LLMs on labeling and annotation tasks is substantially impacted by the specific wording of this prompt; subtle differences across prompts may have consequential implications (Barrie, Palaiologou, & Tãķrnberg, 2024; Li et al., 2024). Moreover, labels generated from LLMs may not be reproducible because details of the labeling process are often omitted from published work or due to the inherent variability in LLM outputs (Barrie, Palmer, & Spirling, 2024).

In our project, challenges with prompt sensitivity are closely related to the challenge of creating a precise a definition of hateful or toxic language. The meaning of hateful or toxic language varies across different cultures, time periods, and social contexts (Fortuna & Nunes, 2019; Lee & Gilliland, 2024; Tontodimamma et al., 2021; Vergani et al., 2024). We iteratively built a suitable definition of toxic language for our study by considering 125 tweets that contained explicit slurs against LGBTQIA+ communities, working off of the conceptual framework of hateful language outlined by Hietanen and Eddebo (2023). During this iterative process, we had high levels of disagreement among the study team when we attempted to classify each tweet with a single "hate speech" label. In contrast, we found that we were able to agree across our different understandings of hate speech only if we clearly disambiguated the different elements of toxic speech. This new classification scheme was inspired by "task decomposition" approaches, in which a complex activity is broken into multiple, simpler steps (AlKhamissi et al., 2023). For example, we realized it was not useful to consider the severity of a tweet if the tweet appeared to be supportive or make a neutral comment about LGBTQIA+ identities; on the other hand, the question of severity was highly relevant for tweets that expressed negative sentiment about LGBTQIA+ identities.

Our resulting definition contains four sequential criteria that we deemed sufficient to determine whether a tweet contained hateful language concerning LGBTQIA+ identities (Figure 2). Our definition first focuses on whether each tweet was written in English. We limited our focus to English-language tweets given that all the authors are fluent in English and able to understand tweets in this language. Next, our definition considers whether a tweet clearly refers to LGBTQIA+ identities in some way—either through explicit reference to an associated identity, or by implicit association from a topic that is strongly associated with these communities (e.g., drag performance). Then our definition prompts an assessment of the sentiment of the tweet, and screens out tweets which are neutral or positive in tone. Finally, our definition asks whether the language of negative tweets expresses this sentiment in a casual or flippant way, or whether the language is severe enough in tone to be considered hateful.

This four-step process formed the basis for our prompts. Given the relative scarcity of LGBTQIA+ related tweets among the full spectrum of discussions that happened on Twitter in 2013, we split this four-step process into two stages. In the first stage, a prompt instructed OpenAI's `gpt-4o-mini` to complete the first two steps in our definition, resulting in a subset of 0.6% of all tweets that were determined to be both in English and to contain references to LGBTQIA+ identities or topics. In the second stage, we fed the English LGBTQIA+ related tweets to `gpt-4o-mini` with a prompt that instructed the LLM to assess the tone and severity of the tweet. This two-stage process substantially reduced computing costs associated with processing, as we avoided the costs of rating most tweets on tone and severity.

We repeatedly revised each prompt until it adequately flagged tweets as directed. At the end of both prompts, we included a handful of example tweets—including examples it had mislabeled on past rounds—and the correct corresponding labels, drawing on a prompt engineering process known as few-shot learning. To validate the first prompt, we paid particular attention to whether the prompt and LLM identified the small subset of English-language tweets that touched on LGBTQIA+ related themes. Drawing on a random validation sample of over 300k tweets, we first measured recall by isolating all tweets that contained at least one LGBTQIA+ related term. These terms ranged from those like "bisexual", "lesbian", "queer", and "transgender" that are strongly associated with LGBTQIA+ identities, to terms like "rainbow" and "ally" that are associated with LGBTQIA+ identities but also often used in non-LGBTQIA+-contexts. Our prompt recovered 100% of tweets in the sample that included strongly associated terms like "bisexual" and "lesbian", and (correctly) flagged fewer tweets when they contained weakly-associated terms like "rainbow" and "ally", flagging 34% and 5% of these tweets, respectively. To measure precision, we examined a random sample of 100 tweets rated as both English and LGBTQIA+ related, determining that precision in this sample was around 74% (95% CI: 65%–83%). Many of the false positives in this sample were tweets associated with romantic, sexual, or bodily themes, but without specific information indicating the tweet referred to these themes in an LGBTQIA+ context.

In the second stage, many of the tweets automatically tagged as "hateful" contained the f-slur, likely indicating specific and intense contempt for LGBTQIA+ individuals. On the other hand, many of the non-hateful tweets used the word "gay" in a context which seemed negative but not overtly aggressive or hostile (even though its use in this way could be inappropriate and could hurt some LGBTQIA+ individuals). To validate the performance of our automated process at this stage, we manually reviewed three hundred tweets, randomly selecting one hundred tweets from each of the following three categories: first, those rated as positive or neutral, and of those rated as negative, those rated as hateful and non-hateful. Of the 100 randomly selected tweets in this sample that our automated process labeled as negative and hateful, we determined that 72% of these tweets indeed appeared to match these criteria upon manual review. Many of the 28% of tweets that appeared to be misclassified appeared to be either vigorous anti-LGBTQIA+ speech or in-group speech (i.e., the use of slurs by an apparent member of the LGBTQIA+ community in a non-derogatory manner). In addition, of the 300 tweets we reviewed that our human rater determined to be negative and hateful, 93% were also labeled by our automated process as negative and hateful. Given these functional precision and high recall rates, we determined that classification performance was sufficient to proceed to analyze the impact of the 2013 policy change.

Five example tweets and their corresponding classifications are shown in Table 1. Each example was selected to represent examples of tweets identified from the corpus. We include more information on these prompts—including the prompts themselves,

Text	**Language**	**LGBTQIA+ related**	**Tone**	**Severity**
me enferma ver gente opinando del racismo, pero ajá el wey súper machista y encima homofobico tantita ptm hijos	not English	-	-	-
good! rt @wsfa12news breaking: harvey updyke hit with federal charges in #toomer's corner tree poisoning case... http://ow.ly/4ww6d	English	not related	-	-
rt @adamoceano: gay marriage vs straight marriage, who the fuck cares, fuck what the bible says or these peice of shit politicians. open...	English	related	neutral	-
man i'm stuck at school my class went on ft i'm split up gay ass teacher lookin.lik the grinch	English	related	negative	not hateful
@wtfsylviaaa hes a fag dont cry over him or that stupid bitch	English	related	negative	hateful

Table 1

Classification of tweets based on language, LGBTQIA+ relation, tone, and severity. The five tweets presented here were chosen as examples to represent the general idea of each classification step. We include 20 randomly chosen examples from each category in the appendix.

hyperparameter settings, and ten randomly selected examples from each step of our classification process—in the Appendix under "LLM setup" and "Classification examples".

Statistical analysis

We used an interrupted time series (ITS) design to estimate the causal effect of the 2013 policy change, because—unlike other causal inference methods—ITS does not require the existence of a separate comparison group. Based on the Rubin causal model, we formalize our causal identification assumptions (Holland, 1986; Rubin, 1978) and borrow the adapted potential outcomes notation from Miratrix (2022) for ITS.

Notation

Let t represent an index for any given week in our sample, t_0 represent the week where the policy change began, t_{min} represent the first week in our inferential window (the week including July 1, 2012), and t_{max} represent the last week in our inferential window (the week including July 31, 2014). Let Y_t represent the incidence of toxic language concerning LGBTQIA+ identities during week t. $Y_t(0),\ \{t = t_{min}, \ldots, t_{max}\}$ represents what the incidence of LGBTQIA+ hateful language on Twitter would be in the absence of Twitter's 2013 policy change, although these values were not observed in practice. Likewise, $Y_t(1),\ \{t = t_{min}, \ldots, t_{max}\}$, represents what the incidence of LGBTQIA+ directed toxic language on Twitter would be in the presence of Twitter's 2013 policy change.

We assume that $Y_t(1) = Y_t(0)$ for all $t < t_0$, meaning that there is no effect of the policy change prior to its implementation in July of 2013. The causal effect of the policy change at time point t is defined as $\Delta_t \equiv Y_t(1) - Y_t(0)$. After the beginning of the policy change, $t \geq t_0$, all observations are values of $Y_t(1)$ (the incidence under the policy change), and we do not observe any values of $Y_t(0)$ because the policy was implemented.

Classical implementation of ITS

ITS uses observed values of $Y_t(0)$ when $(t < t_0)$ to create a forecast of what $Y_t(0)$ would have been after the beginning of the policy change $(t \geq t_0)$ because these values are not directly observed. To estimate the causal effect of the policy change $(\widehat{\Delta}_t)$, the forecast of $Y_t(0)$ is compared against observed incidence under the policy change $(Y_t(1))$. ITS assumes the pre-intervention trend would have continued in the absence of the intervention in the post-intervention time period. The classical ITS analysis specifies an ordinary least squares regression model as follows:

$$Y_t = \beta_0 + \beta_1(t - t_0) + \delta_0 I(t \geq t_0) + \delta_1 I(t \geq t_0)(t - t_0) + \epsilon_t$$

Under this specification, $\widehat{\Delta}_t = \delta_0 + \delta_1(t - t_0)$ for $t \geq t_0$.

Following Gelman and Hill (2007), we centered the time variable t so that its coefficient, β_1, and the subsequent interaction term, δ_1, are interpretable. While many classical ITS analyses do not model autocorrelation across adjacent weeks, we applied the Newey and West (1986) method to adjust our standard errors. This approach, as described in Bottomley et al. (2019), extends the Huber-White heteroskedasticity-consistent standard errors approach (Huber et al., 1967; White, 1980) to account for potential temporal dependencies in ITS analyses. We implemented this correction with the `vcovHAC` function in the `sandwich` R package (Zeileis et al., 2020).

Bayesian time-series simulation ITS

A Bayesian implementation of ITS proposed by Miratrix (2022) explicitly models autocorrelation with a time-series model and again creates a forecast of the unobservable $Y_t(0)$ values for $t \geq t_0$ with data simulated from this time-series model. This method uses Bayesian inference and extends the simulation approaches described in Gelman and Hill (2007) to the ITS approach. It begins by fitting a lagged outcome model to pre-policy data specified as:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 Y_{t-1} + \epsilon_t \ \text{ with } \ \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

In our analysis, we fit this model with the `stan_glm` function in the `rstanarm` package with the default weakly informative prior distributions (Goodrich et al., 2024). Our first unobserved $Y_t(0)$ value is at t_0, the first week of the intervention. To simulate a plausible value of $Y_{t_0}(0)$ we draw a vector of $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*)$ values and a scale value of σ^* from the posterior distribution of our model.

For t_0, we can still observe Y_{t-1}, thus our simulated value is a draw from the posterior predictive distribution:

$$Y_{t_0}^* \sim N(\beta_0^* + \beta_1^* t_0 + \beta_2^* Y_{t_0-1}, \sigma^{*2})$$

The simulated value of $Y_{t_0}^*$ is then used as the lagged term to simulate the plausible value of $Y_{t_0+1}^*$. For $Y_{t_0+1}^*$, a new draw of $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*)$ and σ^* is taken from the posterior distribution and then a value of $Y_{t_0+1}^*$ is simulated from the posterior predictive distribution:

$$Y^*_{t_0+1} \sim N(\beta^*_0 + \beta^*_1 t_{0+1} + \beta^*_2 Y^*_{t_0}, \sigma^{*2})$$

This process is continued for subsequent weeks until a value of $Y^*_{t_{\max}}(0)$ is simulated and we have simulated a single auto-regressive time-series of $Y_{t \geq t_0}(0)$. To fully propagate uncertainty in both the parameters and predictions, we simulated a posterior predictive distribution of $Y_{t \geq t_0}(0)$ with 4,000 draws from the posterior distribution.

This approach forgoes estimating $Y_t(1)$ because these values are directly observable. Following this approach, estimates for $\widehat{\Delta}_t$ and corresponding credible intervals are obtained by taking the difference between each observed value of $Y_t(1)$ and the posterior predictive distribution of $Y_t(0)$. Therefore, this approach simulates an auto-regressive forecast of what the incidence of hateful language would have been in the absence of the policy change and then estimates causal effects by comparing the predicted incidence against the incidence that was actually observed.

Results

We fit both the classical ITS model with corrected standard errors as well as the Bayesian time-series simulation approach described above. Both analyses considered a time window of one year surrounding the policy change, with incidence measured week-by-week. This yielded a time series of 105 data points, where each value in the time series represents the incidence of hateful tweets per 100,000 tweets in a given week. Both approaches—the classical ITS and the Bayesian simulated ITS—support the claim that Twitter's 2013 policy change led to a substantial reduction in hateful language concerning LGBTQIA+ identities.

The classical ITS model is shown in Figure 3. Table 2 provides regression coefficients, cluster-robust standard errors, and accompanying test statistics for this model as well. Under our parameterization of this model, the coefficient on `treatment` estimates the immediate impact of the policy change, and the coefficient on `treatment:week`

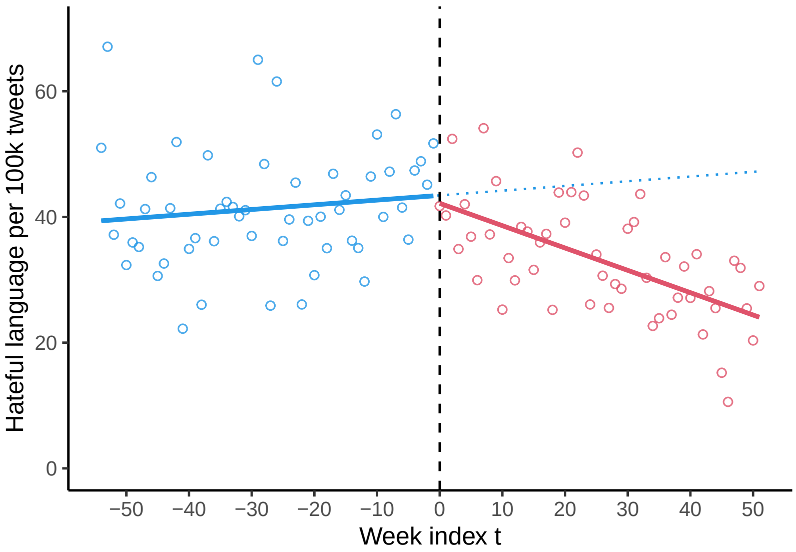


Figure 3

Estimating the impact of the 2013 Twitter policy change using the classical ITS model. Points represent the weekly empirical incidence of hateful tweets concerning LGBTQIA+ identities identified by the LLM, where blue points represent this incidence prior to Twitter's 2013 policy change (vertical black dashed line), and red points represent this incidence following the policy change. Solid lines represent the model fit from the classical ITS analysis. The dotted blue line represents the forecasted counterfactual of what the incidence of hateful tweets would have been in the absence of the policy change. The estimated average causal effect of the policy change at a given week ($\widehat{\Delta}_t$) is the vertical distance between the red solid and blue dotted lines.

represents treatment effect heterogeneity over time. The lack of a sharp discontinuity (and the relatively small size of the coefficient on `treatment`) suggests there was little immediate impact from the policy change on hateful language concerning LGBTQIA+ identities. However, the model estimates that the policy led to a growing reduction in LGBTQIA+ hateful tweets over time (represented by the coefficient on `treatment:week`), ending with a roughly 50% reduction after one year.

Figure 4 presents results from the simulated Bayesian ITS fit with an AR(1) lag, as described in the previous section. Under this approach, we can evaluate the effect of the intervention at the week-by-week level or at the average level across weeks. Observed empirical rates of toxic language that fall outside of the forecasted counterfactual bands

	Estimate	Std. Error	95% C.I.	Pr(> $\|t\|$)
(Intercept)	43.43	2.39	(38.75, 48.11)	<0.0001
treatment	-1.27	2.98	(-7.11, 4.57)	0.336
week	0.07	0.09	(-0.11, 0.25)	0.792
treatment:week	-0.43	0.11	(-0.65, -0.21)	<0.0001
R-squared: 0.3201, Adj R-squared: 0.3001, Residual Std. Error: 8.367				

Table 2
Classical ITS regression coefficients from the analysis of the roughly 8.5 million tweets between July 2012 and July 2014. The coefficient on the interaction between treatment and week indicates the growing impact of the 2013 policy change over time.

signal strong evidence that the policy change led to a reduction in toxic language during the final weeks of our study period compared to this simulated counterfactual. As with the classical ITS model, this analysis suggests that the effectiveness of Twitter's 2013 policy change increased over time. This Bayesian approach also allows easy computation of the likely distribution of the average effect in the year after the policy change (Figure 4, right panel). On average over the year following the policy change, this model estimates that Twitter's 2013 policy change caused an average reduction of 13.20 per 100,000 tweets (95% credible interval -22.52, -4.023) containing hateful language compared to the estimated incidence under the counterfactual.

Effect of the stability of the LLM prompt on the estimated causal effect

As noted in previous sections, LLMs generally achieve equal performance on annotation tasks compared to annotations provided by crowd-sourcing or trained research assistants (Gilardi et al., 2023; Rathje et al., 2024), but LLMs carry unique challenges. If crowdsourced workers or a team of research assistants were asked to replicate annotations for a set of tweets, the variance in annotations between the original set and replicated set may be lower than if this process was completed by an LLM. Barrie, Palmer, and Spirling (2024) demonstrated that in the aggregate, these issues do not affect the accuracy,

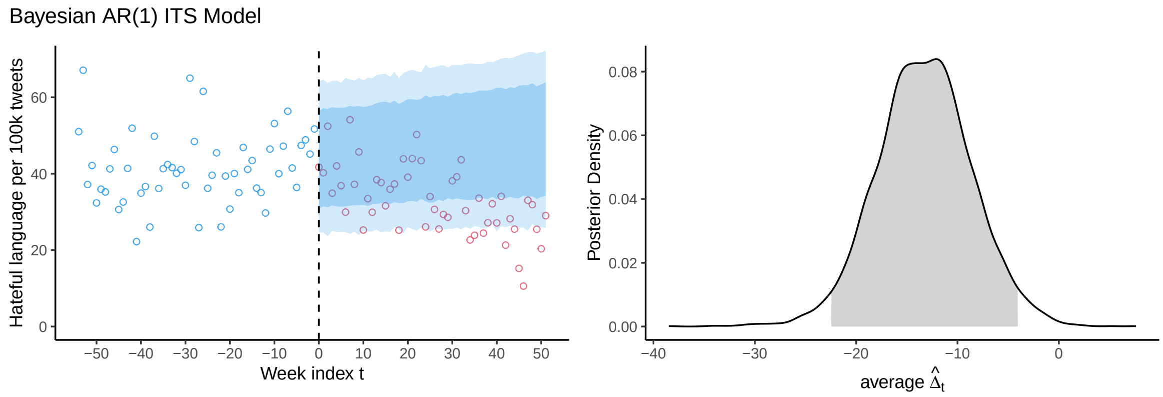


Figure 4

Estimating the impact of the 2013 Twitter policy change using the Bayesian ITS model. The left panel presents results at the weekly level, where points represent the empirically observed incidence of hateful language with blue points occurring before Twitter's 2013 policy change (black vertical dashed line), and red points after. The blue uncertainty bands represent the 80% and 95% uncertainty intervals for the forecasted incidence of hateful language in the absence of the policy change. The right panel shows the posterior distribution for the average estimated causal effect of the policy change over the entirety of the year following implementation with the 95% credible interval depicted with gray shading.

precision, or recall of LLM annotations, but the higher variance of annotations from LLMs may lead to different point estimates, standard errors, and statistical results in reproducibility attempts. To assess our susceptibility to variance in annotations from our LLM, we repeated the second stage from Figure 2 an additional 100 times. Estimates and 95% uncertainty intervals of the interaction term from the classical ITS model and the average $\widehat{\Delta}_t$ from the Bayesian ITS model for all replications are presented in Figure 5. These estimates are substantively stable, and no simulation had uncertainty intervals that crossed zero, indicating that our results are robust to variations in LLM annotations.

Discussion

In this paper, we explored the impact of Twitter introducing a "Report Tweet" button on mobile devices in July 2013 to ease crowdsourced reports of toxic language. We used an LLM to identify tweets that contained hateful language concerning LGBTQIA+

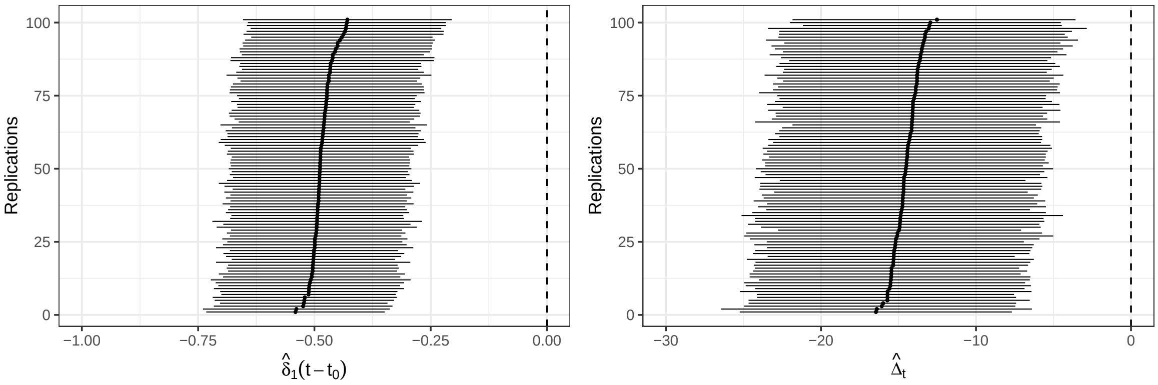


Figure 5

Estimates and 95% uncertainty intervals for 100 replications of the second stage from Figure 2. The y-axis is an ordered index based on the point estimate, the x-axis is the value of the estimate and uncertainty intervals. These replications are substantively stable, indicating that our results are robust to stochasticity in LLM ratings.

identities from a quasi-random sample of 8.5 million tweets posted between July 2012 and July 2014. We then leveraged two versions of an ITS approach to estimate the causal impact of this intervention. Our results suggest that the 2013 policy change on Twitter led to a substantial reduction in hateful language concerning LGBTQIA+ identities, which appeared to increase in impact over time, likely as the updates rolled out to mobile devices and became familiar to users.

Our study has many strengths, but also has some limitations. While annotations from LLMs are not identical to annotations from human raters (Barrie, Palaiologou, & Tãķrnberg, 2024; Barrie, Palmer, & Spirling, 2024), what constitutes a gold standard or absolute rating of toxic language remains up for debate, due in large part to the myriad criteria by which it can be measured (Hietanen & Eddebo, 2023). A growing body of literature comparing labels provided by LLMs to those provided by MTurkers or trained research assistants shows that LLMs have equal or improved performance in many annotation tasks common in the social sciences (Gilardi et al., 2023; Rathje et al., 2024; Ziems et al., 2024). However, scholars have also noted the potential for racial bias in the context of automated hate speech detection (Kim et al., n.d.; Resende et al., 2024; Sap et al., 2019), as well as bias against those in the LGBTQIA+ communities (Dias Oliva et al., 2021). Careful consideration of biased outputs and strategies for mitigation would be necessary to any thoughtful approach to LLM usage in real-world content moderation. Moreover, although our sensitivity analysis demonstrates that our results are robust to the variance in annotations from the labeling process we devised, LLMs may have more difficulty annotating hateful language compared to other classes of annotation (Davidson et al., 2019; Kiela et al., 2020; Li et al., 2024). Further, our results may have differed if our labels had been obtained through a different LLM, different prompt, or a different method.

Therefore, careful consideration of the potential influences of measurement error on our findings is warranted. Measurement error—the systematic under or over-estimation of tweets labeled as toxic—would be problematic if the goal of this work were solely to assess the prevalence of toxic language towards LGBTQIA+ communities on Twitter. Fortunately, this is not the main aim of our study, as our causal estimand of interest is the relative change in toxic language towards LGBTQIA+ communities following the introduction of the 2013 policy change. This subtle distinction has meaningful implications for sensitivity to measurement error. Based on the assumptions made for ITS to be used, the identification strategy used this study is robust against moderate to high measurement error in the labeling of tweets—as long as measurement error is equal before and after the introduction of the policy change, similar to a randomized experiment setting (Gilbert et al., 2024). However, measurement error in our labels would increase residual variance (and therefore the standard errors of the estimated coefficients), resulting in less efficient estimates of causal effects, but such measurement error does not introduce bias in our estimates, so long as the errors do not have temporal dependencies.

ITS models make strong structural assumptions that there are no confounding variables that would explain a change in the trend other than the introduction of the policy change. For example, if a new law or other external event also increased or decreased the incidence of offensive language around the time of the policy change, it would not be possible to estimate the specific effect from the policy change in isolation. Fortunately, related work has found that the incidence of toxic language concerning LGBTQIA+ identities was relatively stable in other time periods, reducing the risk that the changes we estimate here were the result of exogenous forces (Marlo et al., 2020). In another related study measuring hate speech against racial minorities on Twitter, Siegel et al. (2019) also reported similar results, finding no evidence of an influence of current events on the incidence of hateful language.

Our analysis underscores the notion that social media platforms are not simply passive conduits of online speech. Rather, these platforms hold the power to choose which posts to allow and disallow. This ability highlights the political agency held by technology companies today. In recent years, as Twitter has transitioned to X, and has reoriented itself toward weaker moderation in the name of free speech (X, 2024), this change may be imposing substantial, widespread, and harmful impacts on users, including members of vulnerable groups. This toxic language may also harm a wider set of users, e.g., by contributing to an online landscape in which any type of toxic language is permitted, magnified, or even encouraged, and also may extend to other online platforms as well. Open questions remain about who is best positioned to make these high-stakes tradeoffs, and whether the platforms in widespread use today are making the best choice to further the interests of users, stakeholders, and society at large.

In conclusion, this study assessed the impact of a content moderation policy change that Twitter enacted in July 2013 through the introduction of the “Report Tweet” button. We found that this led to a decrease in the amount of toxic language directed at LGBTQIA+ communities on the platform. By using LLMs to rate the tweets used in our analysis, we investigated how to write and test the stability of a prompt meant for a complex labeling task.

References

ADL. (2021). Online hate and harassment: The American experience 2021 [Accessed: 2024-04-11].

AlKhamissi, B., Ladhak, F., Iyer, S., Stoyanov, V., Kozareva, Z., Li, X., Fung, P., Mathias, L., Celikyilmaz, A., & Diab, M. (2023). Token: Task decomposition and knowledge infusion for few-shot hate speech detection. https://arxiv.org/abs/2205.12495

Barrie, C., Palaiologou, E., & Tãķrnberg, P. (2024). Prompt stability scoring for text annotation with large language models. *arXiv preprint arXiv:2407.02039*.

Barrie, C., Palmer, A., & Spirling, A. (2024). Replication for language models problems, principles, and best practice for political science. *URL: https://arthurspirling.org/documents/BarriePalmerSpirling TrustMeBro. pdf*.

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295.

Bottomley, C., Scott, J. A. G., & Isham, V. (2019). Analysing interrupted time series with a control. *Epidemiologic Methods*, *8*(1), 20180010.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.*, *1*(CSCW). https://doi.org/10.1145/3134666

Confessore, N., & Frenkel, S. (2021, August). Facebook taps Accenture for content moderation as scandals mount [Accessed: 2024-11-15]. https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. https://arxiv.org/abs/1905.12516

Dias Oliva, T., Antonialli, D. M., & Gomes, A. (2021). Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture*, *25*(2), 700–732. https://doi.org/10.1007/s12119-020-09790-w

Fortuna, P., & Nunes, S. (2019). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120.

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., & Domingue, B. W. (2024). Estimating heterogeneous treatment effects with item-level outcome data: Insights from item response theory. *Journal of Policy Analysis and Management*.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2024). rstanarm: Bayesian applied regression modeling via Stan. [R package version 2.32.1]. https://mc-stan.org/rstanarm/

Hietanen, M., & Eddebo, J. (2023). Towards a Definition of Hate Speech—With a Focus on Online Contexts. *Journal of Communication Inquiry*, *47*(4), 440–458. https://doi.org/10.1177/01968599221124309

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

Huber, P. J., et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, *1*(1), 221–233.

Instagram. (2019, February). Supporting and protecting vulnerable people on Instagram [Accessed: 2024-11-15]. https://about.instagram.com/blog/announcements/supporting-and-protecting-vulnerable-people-on-instagram

Jeong, S. (2016, January). The history of Twitter's rules. https://www.vice.com/en/article/z43xw3/the-history-of-twitters-rules

Katsaros, M., Yang, K., & Fratamico, L. (2022). Reconsidering tweets: Intervening during tweet creation decreases offensive content. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 477–487. https://doi.org/10.1609/icwsm.v16i1.19308

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, *33*, 2611–2624.

Kim, J. Y., Ortiz, C., Nam, S., Santiago, S., & Datta, V. (n.d.). Intersectional Bias in Hate Speech and Abusive Language Datasets. https://arxiv.org/pdf/2005.05921

Lee, S., & Gilliland, A. (2024). Evolving definitions of hates speech: The impact of a lack of standardize definitions [Series Title: Lecture Notes in Computer Science]. In I. Sserwanga, H. Joho, J. Ma, P. Hansen, D. Wu, M. Koizumi, & A. J. Gilliland (Eds.), *Wisdom, Well-Being, Win-Win* (pp. 141–156, Vol. 14597). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-57860-1_11

Li, L., Fan, L., Atreja, S., & Hemphill, L. (2024). "HOT" chatGPT: The promise of chatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, *18*(2), 1–36. https://doi.org/10.1145/3643829

Lupu, Y., Sear, R., Velásquez, N., Leahy, R., Restrepo, N. J., Goldberg, B., & Johnson, N. F. (2023). Offline events and online hate. *PLOS ONE*, *18*(1), e0278511. https://doi.org/10.1371/journal.pone.0278511

Marlo, J., Perrett, G., & Waheed, B. (2020). Measuring anti-LGBTQ+ language on Twitter. *Working Paper*, (20). https://github.com/joemarlo/hate-speech/blob/main/Writeup_and_presentation/LGBTQ%2B%20hate%20speech%20on%20social%20media.pdf

McConnell, E. A., Clifford, A., Korpak, A. K., Phillips II, G., & Birkett, M. (2017). Identity, victimization, and support: Facebook experiences and mental health among lgbtq youth. *Computers in Human Behavior*, *76*, 237–244.

Meta. (2021, October 17). *Hate speech prevalence has dropped by almost 50% on Facebook* [Meta]. Retrieved March 3, 2025, from https://about.fb.com/news/2021/10/hate-speech-prevalence-dropped-facebook/

Miratrix, L. W. (2022). Using simulation to analyze interrupted time series designs. *Evaluation Review*, *46*(6), 750–778.

Newey, W. K., & West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121.

Resende, G. H., Nery, L. F., Benevenuto, F., Zannettou, S., & Figueiredo, F. (2024, January). A Comprehensive View of the Biases of Toxicity and Sentiment Analysis Methods Towards Utterances with African American English Expressions [arXiv:2401.12720 [cs]]. https://doi.org/10.48550/arXiv.2401.12720

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34–58.

Saha, K., Kim, S. C., Reddy, M. D., et al. (2019). The language of LGBTQ+ minority stress experiences on social media. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–22.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. https://doi.org/10.18653/v1/P19-1163

Schneider, P. J., & Rizoiu, M.-A. (2023). The effectiveness of moderating harmful online content. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(34), e2307360120. https://doi.org/10.1073/pnas.2307360120

Siegel, A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., Nagler, J., & Tucker, J. A. (2019). Trumping hate on Twitter? Online hate in the 2016 US election and its aftermath. *Social Media and Political Participation, New York University.*

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*, 687–690.

Stackpole, T. (2022). Content moderation is terrible by design. *Harvard Business Review.* Retrieved November 16, 2024, from https://hbr.org/2022/11/content-moderation-is-terrible-by-design

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*(1), 157–179. https://doi.org/10.1007/s11192-020-03737-6

Tsesis, A. (2002). *Destructive messages: How hate speech paves the way for harmful social movements* (Vol. 27). NYU Press.

Twitter UK. (2013, July). *We hear you* [Archived at Internet Archive; accessed 8 September 2025]. https://web.archive.org/web/20130801053131/http://blog.uk.twitter.com/2013/07/we-hear-you.html

Vergani, M., Perry, B., Freilich, J., Chermak, S., Scrivens, R., Link, R., Kleinsman, D., Betts, J., & Iqbal, M. (2024). Mapping the scientific knowledge and approaches to defining and measuring hate crime, hate speech, and hate incidents: A systematic review. *Campbell Systematic Reviews*, *20*(2), e1397. https://doi.org/10.1002/cl2.1397

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817–838.

X. (2024, January). *Safeguarding information independence and combating hate speech* [Accessed 8 September 2025]. https://blog.x.com/en_us/topics/company/2023/safeguarding-information-independence-and-combating-hate-speech

Yildirim, M. M., Nagler, J., Bonneau, R., & Tucker, J. A. (2023). Short of suspension: How suspension warnings can reduce hate speech on twitter. *Perspectives on Politics*, *21*(2), 651–663. https://doi.org/10.1017/S1537592721002589

YouTube. (2024). YouTube help - how content is moderated on youTube [Accessed: 2024-11-15]. https://support.google.com/youtube/answer/13304829

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in r. *Journal of Statistical Software*, *95*, 1–36.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291.

Appendix A

LLM setup

These instructions listed below were set as the system prompt for the 2024-07-18 iteration of OpenAI's `gpt-4o-mini`, and then the content of each tweet was passed as the first user prompt. We used structured outputs that required GPT-4o-mini to return its assessment in a pre-specified JSON format, which is included below. As part of this structured output, we required the LLM to first describe its reasoning for its classification *before* providing binary labels, encouraging the model to justify its response before providing a determination. We also set the temperature for the prompt to 0 to reduce expected variation in the outputs generated by GPT-4o-mini. While adjustments to model temperature do not fully ameliorate observed reproducibility issues from LLM-generated annotations, this approach does provide an improvement over higher temperature settings (Barrie, Palaiologou, & Tãķrnberg, 2024; Barrie, Palmer, & Spirling, 2024). As described above, we used two prompts in sequence to determine if a tweet contained hateful language concerning LGBTQIA+ identities. These prompts are included verbatim below.

First phase

Prompt:

```
# Instructions


You will be given a single tweet.


Your job is to determine if the language of the tweet is in english, and if
 ↪  the content of the tweet clearly includes LGBTQIA+ keywords or topics.
```

```
You will be asked to explain your classification first, then follow with your actual labels.

Here are criteria you must follow when making your determinations, followed by concrete examples, and then a step-by-step process for you to follow.

- **explanation**:
    - Write a one-sentence explanation of your classification decisions
    - Be specific about how the content relates or does not
    relate to lgbtqia+ keywords or topics.
    - You should be 100% ABSOLUTELY CERTAIN that the content clearly, objectively, and explicitly relates to LGBTQIA+ keywords or topics.
    - If it simply "suggests" or "can be associated with"
    LGBTQIA+ keywords or topics, **you should decide it is
    NOT LGBTQIA+ related**.
- **english**:
    - You should return "true" if the tweet:
        - Is clearly in English
        - Is in a mix of English and another language
        - Is in a common dialect in English, e.g., African American Vernacular English (AAVE)
    - Otherwise, return "false".
    - **NOTE**: Twitter specific syntax, e.g. "rt" or "@", should not be considered when determining whether a tweet is in English.
- **lgbtqia_content**:
    - You should return "true" if the content of the tweet:
        - *EXPLICITLY* mentions identities, issues, politics, rights,
        ↪  culture, friends, events, slurs, or stereotypes for lesbian,
        ↪  gay, bisexual, transgender, queer, intersex, or asexual
        ↪  (LGBTQIA+) keywords or topics.
        - Mentions something that is 100% likely to be associated with
        ↪  LGBTQIA+ topics.
    - You should return "false" if the content of the tweet:
        - Has nothing to do with LGBTQIA+ keywords or topics
        - "Suggests" or "could be associated with" LGBTQIA+ keywords or
        ↪  topics (we need stronger evidence for this scenario)
        - References romance, sex, or body parts WITHOUT an explicit mention
        ↪  of LGBTQIA+ keywords or topics
        - References identity or gender WITHOUT an explicit mention of
        ↪  LGBTQIA+ keywords or topics
        - References celebrities who *might* be considered LGBTQIA+ icons
        ↪  WITHOUT an explicit mention of LGBTQIA+ keywords or topics

# Examples

- example 1:
    - tweet: Seminario especial de Hip Hop. Todos invitados! No te lo
    ↪  pierdas! http://t.co/4rzyOdQucb
    - correct response: {"explanation": "Tweet is in Spanish and talks about
    ↪  a hip hop seminar without touching on LGBTQIA+ keywords or topics.",
    ↪  "english": false, "lgbtqia_content": false}
- example 2:
    - tweet: me enferma ver gente opinando del racismo, pero ajá el wey
     ↪  súper machista y encima homofobico tantita ptm hijos
    - correct response: {"explanation": "Tweet is in Spanish and touches on
     ↪  homophobia, making it explicitly related to LGBTQIA+ keywords or
     ↪  topics.", "english": false, "lgbtqia_content": true}
- example 3:
    - tweet: man i'm stuck at school my class went on ft i'm split up gay
     ↪  ass teacher lookin.lik the grinch:Right
    - correct response: {"explanation": "Tweet is in English and uses gay as
     ↪  a mild slur, making it explicitly related to LGBTQIA+ keywords or
     ↪  topics", "english": true, "lgbtqia_content": true}
- example 4:
    - tweet: I have a lesbian crush on Ariana Grande and Mila Kunis.  I
     ↪  don't care I don't care I don't careeeee
    - correct response: {"explanation": "Tweet is in English and mentions a
     ↪  lesbian crush, making it explicitly related to LGBTQIA+ keywords or
     ↪  topics", "english": true, "lgbtqia_content": true}
- example 5:
    - tweet: good! rt @wsfa12news breaking: harvey updyke hit with federal
     ↪  charges in #toomer's corner tree poisoning case...
     ↪  http://ow.ly/4ww6d
    - correct response: {"explanation": "Tweet is in English and talks about
     ↪  a federal lawsuit without touching on LGBTQIA+ keywords or topics",
     ↪  "english": true, "lgbtqia_content": false}
- example 6:
    - tweet: Our love is a secret. It wasn't meant to be told to anyone,
     ↪  unless we truly trust them.
    - correct response: {"explanation": "Tweet is in English and discusses
     ↪  love without explicitly touching on LGBTQIA+ keywords or topics",
     ↪  "english": true, "lgbtqia_content": false}
- example 7:
    - tweet: I think about SEX waaaaaay 2 much....this shit gettin way outta
     ↪  hand!!! Lbs
    - correct response: {"explanation": "Tweet is in English and discusses
     ↪  sex without explicitly touching on LGBTQIA+ keywords or topics",
     ↪  "english": true, "lgbtqia_content": false}
- example 8:
    - tweet: Me and my boy ♥♥♥ http://t.co/mn5JODbu6j
    - correct response: {"explanation": "Tweet is in English and discusses a
     ↪  romantic relationship with a boy without explicitly touching on
     ↪  LGBTQIA+ keywords or topics", "english": true, "lgbtqia_content":
     ↪  false}
- example 9:
    - tweet: #PawsUp RT @JennelGarcia: I'm a free bitch, baby!
    - correct response: {"explanation": "Tweet is in English and uses slang
     ↪  without explicitly touching on LGBTQIA+ keywords or topics",
     ↪  "english": true, "lgbtqia_content": false}
- example 10:
    - tweet: @RJL1993 IM BOTH
    - correct response: {"explanation": "Tweet is in English and mentions
    ↪  multiple identities without explicitly touching on LGBTQIA+ keywords
    ↪  or topics", "english": true, "lgbtqia_content": false}


# Process

1. Begin by carefully reading the tweet.
2. Use your best judgment to determine if the tweet is mostly in English and
↪  if its content explicitly mentions LGBTQIA+ keywords or topics.
3. Write an explanation of the labels you plan to use, justifying the
↪  content label with specific detail.
4. Classify the language of the tweet as English or not.
5. Classify the content of the tweet as related to LGBTQIA+ keywords/topics
↪  or not.

***CRITICAL IMPORTANCE: YOU MUST MAKE SURE TO SAY IT IS LGBTQIA+ RELATED
↪  ---IF AND ONLY IF--- THE TWEET EXPLICITLY MENTIONS LGBTQIA+ KEYWORDS OR
↪  TOPICS!***
```

Structured output schema:

```
output_schema = {
    "type": "json_schema",
    "json_schema": {
        "name": "tweet_classification",
        "description": "Classify language and content of a tweet.",
        "strict": True,
        "schema": {
              "type": "object",
              "properties": {
                "explanation": {
                  "type": "string"
                },
                "english": {
                  "type": "boolean"
                },
                "lgbtqia_content": {
                  "type": "boolean"
                }
              },
              "required": [
                "explanation",
                "english",
                "lgbtqia_content"
              ],
            "additionalProperties": False
        }
    }
}
```

Second phase

Prompt:

```
# Instructions


You will be given a single tweet.
Your job is to determine if it constitutes some form of toxic language
 ↪  against the LGBTQIA+ community.
You will be asked to explain your classification first, then follow with
 ↪  your actual labels.


It is quite difficult to determine whether a tweet constitutes toxic
 ↪  language against the LGBTQIA+ community. For example, the use of the
 ↪  word "gay" may be used as a general insult without constituting
 ↪  threatening language against the LGBTQIA+ community. Another thing that
 ↪  makes this task difficult is that you should not confuse LGBTQIA+
 ↪  specific toxic language with other forms of toxic language, e.g. toxic
 ↪  language against racial or religious groups.


# Criteria


- **explanation**: Return a one-sentence explanation of your classification
 ↪  decisions on tone and severity.
- **tone**:
    - Determine if the tweet is positive, neutral, or negative in tone.
        - Return "positive" if the tweet contains supportive or affirming
         ↪  tone, using language that shows acceptance or advocacy.
        - Return "neutral" if the tweet is mostly descriptive without a
        ↪  clear positive or negative tone.
        - Return "negative" if the tweet has a dismissive, judgemental,
        ↪  insulting, harsh, or threatening tone.
    - If you find that the tweet has a "negative" tone, proceed to the next
    ↪  step.
    - If not, skip the remaining step.
- **severity:**
    - Return `null` if you already determined that the tweet is positive or
    ↪  neutral in tone.
    - Return one of the options below if you already determined that the
    ↪  tweet is negative in tone:
        - Return "annoying" if the tweet is inappropriate or offensive
        ↪  without aggression or hostility (e.g., sarcastic comments,
        ↪  insensitive jokes).
        - Return "hateful" if the tweet is aggressive, uses slurs or
        ↪  derogatory language, promotes violence, or shows intent to harm
        ↪  LGBTQIA+ individuals and/or their community.

# Examples

- example 1:
    - tweet: man i'm stuck at school my class went on ft i'm split up gay
    ↪  ass teacher lookin.lik the grinch:Right
    - correct response: {"explanation": "The tweet uses the word 'gay' in a
    ↪  negative tone, but not in an aggressive or violent way.", "tone":
    ↪  "negative", "severity": "annoying"}
- example 2:
    - tweet: @wtfsylviaaa hes a fag dont cry over him or that stupid bitch
    - correct response: {"explanation": "The tweet uses the slur 'fag' in an
    ↪  aggressive negative tone.", "tone": "negative", "severity":
    ↪  "annoying"}
```

Structured output schema:

```
output_schema = {
    "type": "json_schema",
    "json_schema": {
        "name": "tweet_classification",
        "description": "Classify tone and severity of a tweet.",
        "strict": True,
        "schema": {
              "type": "object",
              "properties": {
                "explanation": {
                  "type": "string"
                },
                "tone": {
                  "type": "string",
                  "enum": ["positive", "neutral", "negative"]
                },
                "severity": {
                  "type": ["string", "null"],
                  "enum": ["annoying", "hateful", None]
                }
              },
              "required": [
                "explanation",
                "tone",
                "severity"
              ],
            "additionalProperties": False
        }
    }
}
```

Appendix B

Classification examples

```
=========================

English:  FALSE

LGBTQIA-Related:  FALSE

=========================

 [1] "RT @WOWFakta: Paul Walker 'Fast and Furious' pernah mengajak fansnya
  ↪  lewat twitter utk menyelamatkan Laut Mentawai Indonesia.
  ↪  http://t.co/e…"
 [2] "RT @7ashooor: عندما تُنتهك حُرمات البيوت ؟؟\n\nمن الطبيعي سيكون هناك
  ↪  دهس واعتداء\n\n#صباح_الناصر_تنتفض"
 [3] "http://t.co/haIaTf9l9b"
 [4] "Je suis jaloux. ☹"
 [5] "http://t.co/S6Kgkmzipu"
 [6] "http://t.co/NatDQG9Bum"
 [7] "RT @MichiAle20: La vida no perdona la ignorancia"
 [8] "□Blessed□"
 [9] "@tentanganak @drOei @drtiwi dok anak sy 33 bln, BB: 15 kg, TB: 98cm,
  ↪  malas banget minum susu, tp nafsu mkn bgs.. Solusix gmn?"
[10] "http://t.co/BIZECYJWGz"


=========================

English:  FALSE

LGBTQIA-Related:  TRUE

=========================
 [1] "каталог русского гей порно http://t.co/wXYyK55x http://t.co/0LctILed
  ↪  http://t.co/0AahDpkL"
 [2] "Soy gay, amadme"
 [3] "\"\"@raul_casa: @nachobayuela @4_EverCristian rajado yo??? Rajados
  ↪  vosotros maricones, en cuanto se os gana huis perras;)\"\":-) que
  ↪  dices"
 [4] "RT @lyrajordan1: @farhatabbaslaw Dasar pengacara pengangguran jam
  ↪  segini harusnya kerja bukan malah main tweeter dasar\"\"BANCI\"\""
 [5] "@imkathriaaaaa Ahhh! Si patricia? binusted sya nun eh! XD
  ↪  HAHAHAHAHAHAH. Dami kong twa nun. Langya bakla kse eh!"
 [6] "Sdc malam ini yaaaa @surabayadragcom"
 [7] "RT @_laqueseavecina: ¡SORPRESA! ¡La bollera siempre llama dos veces!
  ↪  #LQSA http://t.co/vyjoWjXc"
 [8] "Aduh ham maneh gestrek ? , geleh anjir :p\"\"@alhamramdhan: Enya apal
  ↪  maneh mah goreng,sabar wenya=D\"\"@Rizkymaulana231: Ke ath ka bencong"
 [9] "RT @KonGre_SO: χθες το πρωί ξύπνησα βαμμένος κομπλέ με μέικ απ...
  ↪  πρέπει να κοιμηθηκα σα πούστης"
[10] "2 pendekar surabaya drag community 201M http://t.co/Wors0YEy"


=========================
English:  TRUE
LGBTQIA-Related:  FALSE
=========================
 [1] "@cm_caroline awh!! I'll miss my beautiful best friend tonight too :( I
  ↪  love you!!!"
 [2] "RT @UMGGAMING: Let's see how this goes. RT for New York Favorite for
 ↪  Texas"
 [3] "The ones that are on welfare and are able to go out and get a job and
 ↪  able bodied, shame on you! Govt handouts suck!"
 [4] "Good morning mentions? Retweet:)"
 [5] "I'm getting kind of hungry"
 [6] "Fuck I really can't sleep"
 [7] "RT @IL0V3J3LLYT0TS: RT FOR SHOUTOUT&lt;3"
 [8] "RT @ThatDudeMCFLY: I'll be done. lol"
 [9] "Glorious Saturn. And You, Too. http://t.co/jeMuKiTBJN"
[10] "@ellen_mair_rae #imprayingforyou"


=========================
English:  TRUE
LGBTQIA-Related:  TRUE
=========================
 [1] "I'm sorry but she's crazy and I fckin luhhhh it.
 ↪  ▯��#womancrushwednesday  #kesha #glitter #bitches
 ↪  http://t.co/D00s7Bjgi5"
 [2] "RT @tomandlorenzo: Butch Charlize Theron is insanely hot. #Oscars"
 [3] "You held your pride like you should have held me."
 [4] "dikes taken over!..kmsl"
 [5] "Can #lesbian #fashion dominate @projectrunway? http://t.co/k6wiZ17g
 ↪  #lgbt #lifetime #television #kors @ninagarcia"
 [6] "I’ve just signed the @ENABLEScotland Anti-Bullying Charter
 ↪  http://t.co/scYKByU6 #OpenYourMind"
 [7] "@evansmcallister lol luv u brah #nohomo"
 [8] "\"\"Pause/No Homo\"\" Rule #37: Never hug a dude longer than
  ↪  ____________."
 [9] "I'm watching Bridegroom (13 others are watching)
  ↪  http://t.co/Fswx0EDlAp #GetGlue @BridegroomMovie"
[10] "@Eric_skeeter_34 fag didn't go to u.s today"


=========================
Tone:  positive
Severity:  NA
=========================
 [1] "\"\"\"\"I'd like forever with you\"\"\"\" - @megan_crandall
  ↪  #betterthananyboyfriend  ������❤"
 [2] "RT @HHSGov: Take the test. Take control. Find #HIV testing sites and
  ↪  care services near you: http://t.co/zPBURBGcaf #NHTD
  ↪  http://t.co/GBnsL…"
 [3] "Go Watch @TrevorMoran Video https://t.co/WHcAJlka0q its really
  ↪  funny!!!!"
 [4] "I just can't contain myself! Soooo happy and excited!!! My cousin is
  ↪  marrying a wonderful girl ! I love em both!! #love #engaged"
 [5] "A VERY special thanks to all of our new customers we met @
  ↪  #LAPride2014!!! You all helped make our… http://t.co/e7bB3DNFRO"
 [6] "MN Publius: Keith Ellison is one of ThinkProgress's \"\"\"\"11 Most
  ↪  Pro-Gay U.S. Representatives\"\"\"\" http://t.co/6MJblzK8"
 [7] "@CVillanOfficial My husband &amp; I are huge Disney fans. We go to Gay
  ↪  Days every year &amp; even got married @Disneyland, Cinderella coach
  ↪  &amp; all."
 [8] "Matthew Curdy is my favorite lesbian"
 [9] "#FF_Special□  my gorgeous wife @KinkyKGrey_FKS  love u baby x
  ↪  http://t.co/uq0FzaKxip"
[10] "Glenn is so Lucky.. Errr.. #girlcrush #thewalkingdead
 ↪  http://t.co/X2catexo1j"


=========================
Tone:  neutral
Severity:  NA
=========================
 [1] "RT @DaReal_MIZZ: \"\"\"\"@Arabella_Rosee: My Girlfriend not allowed to
  ↪  pray silently. I want to know what you &amp; Jesus got going
  ↪  on□□\"\"\"\" □lol"
 [2] "@united Worth noting, @IAMQUEENLATIFAH knows @DanPierson by name."
 [3] "RT @washingtonpost: Showing Michael Sam kiss was not up for debate at
  ↪  ESPN, NFL Network http://t.co/qYDEq2jdSX"
 [4] "THUGGIN #MLP #Rainbowdash #Brony #meanmug http://t.co/iacccz2QYN"
 [5] "Leeeyum!! http://t.co/ny9SJU4lhh"
 [6] "SOMEBODY TOLD ME THAT YOU HAD A BF THAT LOOKED LIKE A GF THAT I HAD IN
  ↪  FEB OF LAST YEAR!!!"
 [7] "New York Times: N.B.A. Center Jason Collins Comes Out as Gay
  ↪  http://t.co/gH442yyEvO"
 [8] "#Beffi���� http://t.co/A0vm748Xfe"
 [9] "MC Lyte Discusses Openly Gay Rappers | ELIXHER: Blood, Sweat and High
  ↪  Hopes: How Black Women... http://t.co/QdK6NiznbK #tlot #tcot #news"
[10] "@opinion8ed_dyke I know!!! #apple"


=========================
Tone:  negative
Severity:  not hateful
=========================
 [1] "RT @ImLeslieChow: Ghetto word of the day: Bishop. My girlfriend fell
  ↪  down, so I picked that bishop."
 [2] "Gay couple finally admits they're the cause of major disasters.
  ↪  \n\nhttp://t.co/JHI07QyE"
 [3] "@OfficialJackM they're all closet cases"
 [4] "RT @maeemousiee: I wish someone would bring up gay marriage. I wanna
  ↪  see Romney squirm. #seriously"
 [5] "@fre3zeyyy @DeanoGriff1 @CSimpss @SimpsonKane all gay retweeting
  ↪  eachother cus it will be funny"
 [6] "@hdnovikova: \"\"\"\"Why is it called a tall boy? That's sexist. Why
  ↪  can't you have a tall girl? Or a tall hermaphrodite.\"\"\"\""
 [7] "@MasonK5  @landaal13 YOURE THE DYKE MASON. HAHAHAHA #RealTalk"
 [8] "RT @WJMFournier: \"\"\"\"The Bible says marriage is between one man,
  ↪  and that man's rib. Nothing gay.\"\"\"\" @ColbertReport @stephenfry"
 [9] "RT @ThickMaster: @therealpinkyxxx is a beast with that strap-on!  RT
  ↪  @PlaceHo69925816: http://t.co/GP0XlrDoUA\"\"\"\""
[10] "Going to see Nana: the only time I hear 7+ second farts and phrases
 ↪  like \"\"\"\"the queers\"\"\"\" come from the same woman."
```

```
=========================

Tone:  negative

Severity:  hateful

=========================

 [1] "@MyMelody_4  FAG!!"

 [2] "@TEYANAGOTFUCKED @TEYANATAYLOR faggot i swear to god"

 [3] "Right-wing pastor: Girl Scouts are wicked and their cookies promote

  ↪  lesbianism via @socialreader"

 [4] "RT @Oliverr_sti: Honda drivers makes me hate Hondas they're all a

  ↪  bunch of faggots that think they're car is all bad when in reality is

  ↪  shi…"

 [5] "Eric is the biggest faggot ever"

 [6] "@QuinlanCain because teh hipsters aren't really cool Quin they just

  ↪  act like on the first impression #fags#snapback#swagfag#deathtorhemall"

 [7] "Men do less than they ought, unless they do all they can.

  ↪  @ibrokeyourheart #buds #gayquote #fagsinplaid #dev

  ↪  http://t.co/4KCvSaMs"

 [8] "\"\"\"\"@CurvedDaily: Really man...? RT" @Durags4Eva: Breh..........

  ↪  RT @JamalFacts: You a faggot If you wouldn't fuck Prince.."\"\"\"\"

  ↪  erm...lmao"

 [9] "RT @FeaturingOso: Pepe Billete: Banning #SameSexMarriage is Tremenda

  ↪  Mariconada http://t.co/bI9PbAnXJM via @cultistmiami"

[10] "@beckyivesxo yesss yess DR FAGGIT mwahh ;)"
```