Sexuality & Culture (2021) 25:700–732
https://doi.org/10.1007/s12119-020-09790-w

ORIGINAL PAPER

Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online

Thiago Dias Oliva[1,2] · **Dennys Marcelo Antonialli**[3,4,5,6] · **Alessandra Gomes**[2,7,8]

Accepted: 26 October 2020 / Published online: 6 November 2020

Abstract

Companies operating internet platforms are developing artificial intelligence tools for content moderation purposes. This paper discusses technologies developed to measure the 'toxicity' of text-based content. The research builds upon queer linguistic studies that have indicated the use of 'mock impoliteness' as a form of interaction employed by LGBTQ people to cope with hostility. Automated analyses that disregard such a pro-social function may, contrary to their intended design, actually reinforce harmful biases. This paper uses 'Perspective', an AI technology developed by Jigsaw (formerly Google Ideas), to measure the levels of toxicity of tweets from prominent drag queens in the United States. The research indicated that Perspective considered a significant number of drag queen Twitter accounts to have higher levels of toxicity than white nationalists. The qualitative analysis revealed that Perspective was not able to properly consider social context when measuring toxicity levels and failed to recognize cases in which words, that might conventionally be seen as offensive, conveyed different meanings in LGBTQ speech.

Keywords Artificial intelligence · Content moderation · Toxicity · Hate speech · Queer linguistics · Drag queens

✉ Thiago Dias Oliva
thiago.oliva@usp.br

1 University of São Paulo, São Paulo, Brazil

2 InternetLab, São Paulo, Brazil

3 University of São Paulo Law School, São Paulo, Brazil

4 Stanford Law School, Stanford, USA

5 Bucerius Law School, Hamburg, Germany

6 WHU Otto Von Beisheim School of Management, Vallendar, Germany

7 State University of Campinas (UNICAMP), Campinas, Brazil

8 Federal University of Pará (UFPA), Belém, Brazil

Introduction

Algorithms[1] have been employed from the web's origins to the present. In recent years, they have been playing an increasingly important role in the selection of information deemed relevant to people online, working as 'content curators' at the user level. Algorithms act automatically and with no regular human intervention or oversight, opaquely shaping discourse on the internet. For instance, Gillespie (2014) mentions that algorithms are useful not only for finding information, but also for providing people with tools to organize and classify knowledge, as well as to take part in social and political discourse.

Algorithms scale easily and are employed in the enforcement of content policies by social media platforms like Facebook, Twitter, and YouTube. The algorithms developed by these companies typically rely on machine learning techniques and are customized for each type of content, such as images, videos, audio files, and written text.

Considering they make the organization, processing, and analysis of large-scale amounts of data possible; algorithms are useful in structuring decision-making processes regarding content regulation in social media. In other words, they provide internet platforms with tools to police an enormous and ever-increasing flow of information—which comes in handy in the implementation of content policies established in platforms' Terms of Service (ToS).

Major internet platforms use technology to filter content, at times before it is available online, in order to improve efficiency when enforcing copyright (Keller 2018, p. 6). YouTube's Content ID, for instance, employs algorithms to identify and take down infringing content. YouTube's algorithms search for 'hashes', a unique digital fingerprint that is automatically assigned to specific images and videos, allowing all content with the same fingerprints to be quickly identified and removed. When users try to upload files that match the ones previously provided by copyright owners, the material may be blocked before even becoming available on YouTube.

Hash matching algorithms are also used to counter the dissemination of child pornography (Curtis 2015). Efforts to efficiently remove this kind of content brought together major internet platforms like Facebook, Google, Microsoft, Twitter, and Yahoo. These companies began using a shared database with a large trove of images, gathered by the Internet Watch Foundation, to feed a 'hash list' of content to be identified and deleted. A similar hash-matching technology is employed by Facebook in the detection and automatic removal of non-consensual intimate images (Solon 2017). Content identified as terrorist propaganda is also being taken down by hash matching technologies. In 2016, Facebook, Twitter, Google, and Microsoft announced a shared database of hashes for removing offending content.

As opposed to traditional law enforcement, which encompasses detection, prosecution, adjudication, and punishment performed by state actors with

[1] Algorithms may be defined as 'encoded procedures for transforming input data into a desired output, based on specified calculations' (Gillespie, 2014: 167). They are designed to store and analyze data, apply mathematical formulas to it and come up with new information as a result.

nominal transparency, algorithmic content policing allows private actors to perform all of the above roles. It focusses on early detection and prevention, making transparency even more challenging. Because algorithms are embedded in the architecture of the digital environment, they effectively shape the way people can behave online.

In 'Code and Other Laws of Cyberspace, Version 2.0', Lessig (2006) explains how the development of technological structures enabled behavior regulation over the internet. In this sense, the 'code', which is behind the web's architecture, shapes much of the user's experience when accessing online content. For that reason, behavioral regulation by means of the 'code' might be more efficient than law enforcement, especially when it comes to speech regulation, since the 'code' enables the technical blocking of content, which means it will be immediately and unavoidably removed.

Consequently, choices involving algorithm development help establish and maintain standards of communication online, which means that internet platforms—the ones behind the development of algorithms governing social media—exercise great power over users' speech in their digital environments. In fact, speech that vanishes from the most popular platforms loses much of its impact simply because many potential listeners never come across it. It is worth adding that most platforms reserve great powers in their ToS to remove content posted by users, establishing they may do it at their 'sole discretion' or under the 'belief' that it violates their policies (Suzor 2018). That means intermediaries may greatly impact users' digital rights when algorithms wrongfully remove non-offending content.

Perel and Elkin-Koren (2016) write that copyright enforcement technologies such as Content ID may fail to recognize fair use cases, blocking content even when it does not infringe copyright. Other hash matching technologies also seem to operate in 'context blindness'. Such content blindness appears to be among the reasons why YouTube shuttered the channel of Qasioun News Agency, temporarily taking down over 6,000 videos documenting the Syrian conflict (BBC 2017). Regardless of recent improvements, content moderation technologies still fail to adequately understand context.

This paper focuses on artificial intelligence (AI) technologies developed to measure the 'toxicity' of text-based content, particularly those using natural language processing (NLP) and sentiment analysis to detect harmful text without relying on a set of banned words or phrases. While these technologies may represent a turning point in the debate around hate speech and harmful content on the internet, recent research has shown that they are also still far from being able to grasp context or to detect the intent or motivation of the speaker, failing to recognize specific usages of certain words as socially valuable content.

Mindful of these limitations, this article discusses the performance of an existing AI technology against a background of socially situated speech in order to reflect upon the following question (1) should computers decide what is toxic on the internet?

The use of 'mock impoliteness' has been identified by queer linguistic studies as a form of interaction that serves to prepare members of the LGBTQ community to cope with hostility. Automated analyses that disregard such a pro-social function

might have significant repercussions on the queer community's ability to reclaim words such as 'dyke', 'fag', and 'tranny' and reinforce harmful biases.

With this in mind, the present study measures the levels of toxicity of postings from prominent drag queens in the United States (USA) and compares the results with the levels of toxicity of postings from other prominent Twitter users. To that end, it uses 'Perspective', an AI technology developed by Jigsaw (formerly Google Ideas).[2] Inspired by LGBTQ activists' reports on the biased enforcement of platforms' ToS, and the peculiarity of LGBTQ communication codes, the present study investigates whether Perspective reinforces harmful bias against the LGBTQ community.

In order to better contextualize the research published here, the next sections start by introducing additional information on content moderation in internet platforms. They review studies on the implementation of AI tools for dealing with plain text. Additionally, they summarize existing literature on communicational practices of LGBTQ people, pointing to socially situated speech that current AI technologies could have difficulties analyzing properly.

The section 'Methods' explains the methodology employed in the research, i.e., how the results were obtained. It brings attention to the technical aspects of the study, naming the tools employed in the whole process. It also explains the methodological choices that were made by the authors.

The section 'Results and Discussion' presents the results of both the quantitative and the qualitative aspects of the research, discussing the impact AI tools such as Perspective might have on LGBTQ expression online. In short, the research revealed that a significant number of drag queen Twitter accounts were perceived by Perspective as more toxic than the accounts of Donald Trump and white nationalists. Toxicity levels of drag queens' accounts ranged from 16.68 to 37.81%, on average. White nationalists' averages remained between 19.68 and 33.46%; Trump's reached 21.84%. The qualitative analysis indicated that Perspective was not able to grasp social context when analyzing drag queen tweets precisely because it failed to recognize cases in which usually offensive words conveyed different meanings in LGBTQ speech or were reclaimed by members of the queer community.

The Use of AI for Dealing with Text-Based Harmful Speech and Risks to LGBTQ Expression

In addition to artificial intelligence tools for dealing with audiovisual content, internet platforms are developing technologies for handling text-based harmful speech. In fact, some of the earliest internet filters were programs that matched text content in order to filter spam or pornography. They seemed to be effective for that purpose, but they also resulted in many false positives, blocking legitimate content. Current technology uses NLP and sentiment analysis, in an attempt to detect harmful text without relying on a set of banned words or phrases. Even though this technology

[2] Available at: https://www.perspectiveapi.com/#/.

has evolved a great deal over the years, the accuracy of commercially available tools remains between 70 and 80 percent—i.e., the algorithms take down non-harmful content one time in every four or five (Duarte et al. 2017, p. 5). Recent research (Duarte et al. 2017, p. 3) developed by the Center for Democracy and Technology concluded that:

> Today's tools for automating social media content analysis have limited ability to parse the nuanced meaning of human communication, or to detect the intent or motivation of the speaker.[…] Without proper safeguards, these tools can facilitate overbroad censorship and biased enforcement of laws and of platforms' terms of service.

Still according to the research, even though NLP filtering tools are more sophisticated than earlier technologies, they are far from grasping the full complexity of human communication. Meaning is highly context-based, depending on, for instance, the audience, place of communication, speaker, and tone. Besides, translating policy into code may result, unintentionally, in substantial changes in meaning, mostly because computer languages have limited vocabularies compared to those of humans.

In the paper 'Deceiving Google's Perspective API Built for Detecting Toxic Comments', Hosseini et al. (2017) mentioned that minor changes in sentences and their structures—such as inserting spaces between words and typos—can deceive Perspective, Jigsaw's AI, diminishing dramatically the perceived toxicity of harmful speech. The company continuously improves the technology to detect these changes, but it remains flawed—the study 'All You Need is "Love": Evading Hate Speech Detection' published by Gröndahl et al. (2018) mentions, for instance, that if the word 'love' is inserted into the sentences, their perceived toxicity also diminishes.

Another study on automated hate speech detection highlights the challenge it is to separate hate speech from other instances of offensive language. Davidson et al. (2017) employed a crowd-sourced hate speech lexicon (Hatebase.org) to collect tweets with hate speech keywords and had a sample of these tweets labeled by CrowdFlower workers into three categories: those containing hate speech, only offensive language more broadly and those with neither. Thereafter, a multi-class classifier was trained with the labeled data to distinguish between these three different categories. As a result, the study concluded that lexical methods may be effective to identify potentially offensive content but are inaccurate at identifying hate speech—around 40% of hate speech content was misclassified as something else, while 5% of merely offensive content was misclassified as hate speech. Hence, tweets without explicit hate keywords were more difficult to classify. The study also discussed how difficult it is to adopt a clear definition of hate speech, considering 'our classifications of hate speech tend to reflect our own subjective biases'. The researchers mentioned, in this regard, that CrowdFlower workers identified racist and homophobic slurs as hate, but tended to see sexist language as merely offensive.

A more recently published paper that also analyzed the performance of AI tools for handling text-based harmful speech found out that 'African American English tweets are twice as likely to be labelled offensive compared to others' (Sap et al. 2019, pp. 1668, 1669). This disparity led the authors to recommend extra attention

to 'the confounding effects of dialect so as to avoid unintended racial biases in hate speech detection'.

The study attempted to quantify the unintentional racial bias that may arise during annotation processes, investigating the correlation between toxicity annotations and dialect probability. It also discussed how these biases are acquired by predictive models. For that purpose, the study reported differences in rates of false positives between African American English and White-aligned dialect groups for models trained on toxic language datasets used in hate speech detection. Thereafter, it applied these models to two reference Twitter corpora and calculated average rates of reported toxicity, indicating how unintended biases generalize to other data. In short, the work shows how insensitivity to dialect can result in discrimination against minorities, even when social identities are not explicitly mentioned.

These studies indicate that classifiers might miss nuances in language use, what ends up affecting the performance of AI tools. Therefore, the review, by human moderators, of flagged content (whether flagged by users or by AI technologies) remain essential for analyzing context and avoiding mistakes.

Human rights activists claim that platforms are wrongfully silencing their speech. In 2017, for instance, justice organizations contacted Facebook, arguing the company disproportionately removes content produced by minorities (Levin 2017). Even when the first analysis of certain materials is carried out by human moderators, once human errors feed into a filter's algorithm, these errors might be amplified and cannot be corrected promptly, causing mistakes to become the rule and making it impossible for users to post certain images or words—that's the case, for instance, with the use of the words 'dyke', 'fag', and 'tranny' which is restricted by Facebook's speech policies (Lux and Mess 2019). This 'context-blind' ban on the platform overlooked LGBTQ people's attempts to reclaim these words as means for self-expression or for 'building a thick skin'.

The Pro-social Function of the Language and Discourse of Drag Queens for The LGBTQ Community

Queer linguistics and studies on communication styles of LGBTQ people have long acknowledged the playful use of potentially impolite utterances by members of the community. Murray (1979) and Perez (2011) have, for instance, discussed how gay men engage in ritual insults, while Jones (2007) and Johnson (1995) identified interactional practices such as playful putdowns. Heisterkamp and Alberts (2000) reported LGBTQ people tease each other as a means to humorously build in-group solidarity. In a recently published study that encompassed empirical analysis of interactions codes of drag queens, McKinnon (2017, p. 90) identified the use of 'utterances, which could potentially be evaluated as genuine impoliteness outside of the appropriate context, [but] are positively evaluated by in-group members who recognize the importance of 'building a thick skin to face a hostile environment'.

McKinnon labeled such utterances as 'mock impoliteness' because they involve interactions in which the talk or conduct are potentially opened to evaluation as impolite by at least one of the participants in the interaction, and/or as non-impolite

by at least two participants. Frequent topics that are used in these interactional practices are related to sexual roles in relationships and the visibility of gayness and sexual promiscuity, subjects usually explored by those who aim to verbally attack gay men.

In one of the examples provided by McKinnon in his study, which uncovers drag queen cultural practices in backstage talk, one of the drag queens is being called out by others due to her alleged sexual behavior. She is deemed 'thirsty'—a slang term used to describe her longing for sexual encounters. Instead of feeling offended, she collaboratively engages with the other drag queen's observation, adding that 'she needs to go to the ocean because she's so thirsty'—a humorous and creative self-evaluation that combines the literal and metaphorical use of the word 'thirsty' (McKinnon 2017, pp. 108, 109). Furthermore, in the same conversation, the same drag queen is said to have 'chlamydia's brain', meaning not only she is sexually promiscuous but also that her brain was affected by the sexually transmitted disease. In her collaborative response—'I can't have all this and brains toooooo'—she implicitly assumes she has many sexual partners, but because she is beautiful, countering another possible attack against her sexual behavior (McKinnon 2017, p. 120). Even though LGBTQ people discuss sexual behavior often, this topic could also be explored by non-community members wishing to verbally attack them.

Another recurrent topic is the visibility of gayness, as attested by Heisterkamp and Alberts (2000): LGBTQ people often have perceptions about 'how queer' they look or behave, especially in public places—perceptions that become part of their self-identities. In the teasing interactions analyzed by Heisterkamp and Alberts, participants attempted to create specific identities for themselves as either gay/lesbian-looking or straight-looking. For instance, Alex, a gay man, stated people could not realize he was 'queer' just by opening a door and seeing him. Liz, a lesbian woman, answered in form of a tease (Oh no, Alex), as suggested by her exaggerated and emphasized utterance. When questioned whether Liz could be perceived as lesbian, Alex answered 'The first time I saw her at a meeting over in City Heights I thought dyke, dyke'—followed by Liz's laughter. Heisterkamp and Albers (2000, p. 396) concluded that 'these teasing interactions also serve to form a relational identity in which these participants are bonded in their mutual recognizability as gay men or lesbians'.

It is worth adding that in-group/out-group status may help create contextual conditions that predispose particular experiences of language. A word that is experienced as a slur when hurled by an outsider can be experienced as a joke when used by an in-group member. LGBTQ people reclaim slurs by using them within the community. A word that might normally convey malice here conveys solidarity. Such usages 'boost the armor' of those engaging in the interaction and provide an opportunity to hone verbal skills for future defense against harmful uses and malicious actors.

That is the case, for instance, in one of the interactions between drag queens McKinnon analyzed in his work. When talking about the television series *The Facts of Life*, Melinda assigns Eva the role of Natalie, a large-figured character. After Eva's playful protest, Melinda responds with the personalized negative assertion 'you're fat and gross' and vocative 'you dumb bitch', which could be understood as

potentially offensive. Despite seeming self-conscious about her weight, Eva's laughter and her exaggerated vocalization after Melinda's assertion indicates she positively evaluated the interaction as mock impoliteness (McKinnon 2017, p. 113).

Therefore, McKinnon considers this kind of interaction has a pro-social function: to prepare members of the LGBTQ community to face a hostile world outside of it. According to Murray (1979, p. 218), gay men

> are likely to encounter degrading remarks made by other gay men as well as those made by participants in the dominant culture. A sharp tongue is a weapon honed through frequent use, and is a survival skill for those who function outside genteel circles […] [they have] been using perception and quick formulation to demand acceptance – or to annihilate any who would deny it. Such [in-group] play is quite literally, self-defense.

Among drag queens, this practice is called 'reading' and was already reported in the early 90s documentary 'Paris is Burning' as 'the real art form of insult'. According to McKinnon (2017, p. 97), the primary function of *reading* is to 'verbally arm members to combat instances of genuine impoliteness of the drag queen community, which makes use of factual statements for its ammunition a fundamental aspect of this practice'. Eva, a drag queen interviewed by McKinnon (2017, p. 104–105) on communication practices of the community, explained that

> reading, is not about being mean and awful […] is about finding something that you know the other person is kind of self-conscious about, and picking on that. In a playful way […] it's never meant to be mean.[…] I think it's also a form of building a thick skin. We do it to each other, in jest, because somebody, because we're going to enter this world, where we're over feminized as people.

The widely popular American reality competition series 'RuPaul's Drag Race', a show that documents the search for 'America's next drag superstar', frequently displays *reading* challenges, in which competitors must make insulting observations about one's peers for comedic effect.

Methods

The research focused on Perspective, an AI technology developed by Jigsaw (formerly Google Ideas), that measures the perceived levels of 'toxicity' of text-based content. Perspective defines 'toxic' as 'a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion'. Accordingly, the model was trained on crowdsourced data. Annotators were asked to label internet comments on a scale from 'very toxic' to 'very healthy'. The levels of perceived toxicity indicate the likelihood of a specific content to be considered as 'toxic'.

This study uses Perspective's API to measure the levels of toxicity of prominent drag queens in the USA and compare them with those of other prominent Twitter users in the USA, especially far-right figures. The drag queens' Twitter

profiles were selected out of former participants of the reality show 'RuPaul's Drag Race', popular in the USA and abroad, particularly among the LGBTQ community.

In order to develop the research, access to both Perspective's and Twitter's API was requested by the authors. After the request was granted by both Jigsaw and Twitter, we proceeded to the collection of tweets of all former participants of RuPaul's Drag Race (seasons 1 to 10) who had verified accounts on Twitter (in May 2019) and who post content in English, amounting to 80 drag queen Twitter profiles. The research was carried out with Perspective's production version 6 dealing with 'toxicity'. Only content posted in English was submitted for analysis—tweets in other languages were excluded.

Tweets of prominent non-LGBTQ people (Michelle Obama, Donald Trump and white nationalists) were also collected. These Twitter accounts were chosen as control examples for less controversial or 'healthy' speech (Michele Obama) and for extremely controversial or 'very toxic' speech (Donald Trump and white nationalists).

The 10 white nationalist accounts analyzed in the research were selected from a list published by HuffPost USA (O'Brien 2019) that included 62 Twitter users who identify themselves as white nationalists and share content supporting white supremacy. The profiles selected were the first ten—in the order listed by HuffPost's journalist—to have at least 100 tweets posted on 30 September 2019: Richard Spencer, David Duke, Stefan Molyneux, Kevin MacDonald, James Edwards, Steve King, Faith Goldy, Jean-François Gariépy, Lana Lokteff and Henrik Palmgren. Nick Fuentes was excluded because he does not publicly identify himself as white nationalist, despite sharing views that often align with the ideology. Other accounts were already suspended by Twitter.

In total, 127,137 tweets were collected, and 124,148 tweets were analyzed (after exclusions). Retweets were deliberately not collected because they were not written by the people behind the Twitter accounts under analysis. That was also the case for tweets answering to other tweets (in threads, for instance) since their tones and messages are framed by ongoing conversations with other users.

Two indexes were derived with the results returned by Perspective's API: the 'average toxicity level of accounts' and the 'median toxicity level of accounts'. The first was calculated by adding the toxicity level of all tweets (a number between 0 and 100% returned by Perspective's API) from the same account and dividing it by the number of tweets of that same account. The latter was obtained by finding, between the results returned by Perspective' API for all the tweets of a same account, the value such that a number is equally likely to fall above or below it. Those indexes were used as basis for comparing perceived toxicity levels in connection with drag queen and white nationalist speech. In addition to the quantitative analysis described above, the research also involved a qualitative analysis of tweets posted by drag queens and white nationalists to identify eventual discrepancies in the results provided by Perspective's API.

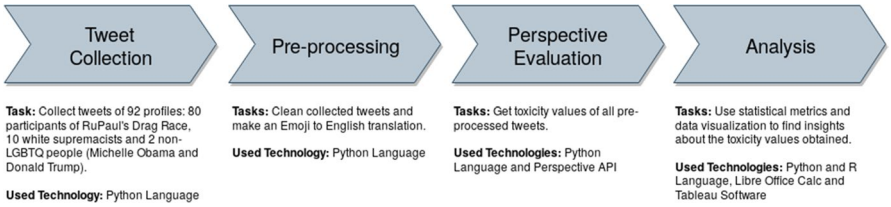


Fig. 1 Research development steps

Technical Aspects of the Research

Tweets from the selected profiles were automatically collected and submitted to Perspective's API. The whole process consisted of four steps: (1) tweet collection; (2) pre-processing of tweets; (3) Perspective's evaluation; and (4) in-depth analysis of the results returned by Perspective. Figure 1 illustrates these steps:

In the first step, tweets of the selected profiles were collected and stored in a CSV dataset. Only three attributes of the tweets were stored: the author's name, the tweet's time information and its content. A tweet identifier (an ID auto increment number) was also created and stored. It is worth mentioning that Twitter's API returned not only tweets but also retweets and replies of the selected Twitter profiles. Due to the reasons explained in the previous section, retweets and replies were ignored, what means only regular tweets were stored in the dataset.

It is important to note that Twitter's API returns a maximum of 3,200 tweets per request. Considering only one request per profile was presented, it is possible that the API did not return all tweets posted by the selected accounts. Additionally, Twitter does not provide information on the criteria for returning relevant tweets.

We built a script in Python that uses the Tweepy library[3] to collect tweets. A pseudo-code of the algorithm developed is presented below:

Step 1: Pseudo-code used in the collection of tweets from Twitter's API

```
get twitter authorization
get twitter users' ids
FOREACH twitter user id IN list of all twitter users' ids
        create a csv file for the current user
        WHILE (current user has tweets) DO
                get current user tweet
                IF (tweet is not a reply) THEN
                        save tweet data in the current csv file
                END
        END
END
```

[3] Available at: https://www.tweepy.org/.

The following step involved the cleaning and preprocessing of the content, which included the removal of links in general, as well as special characters (such as', ",", …), tweets that were not written in English, and the translation of emojis into English text. Considering Perspective is not able to analyze images, they were also removed in the cleaning process. The authors made the methodological choice not to employ optical character recognition (OCR) techniques to convert images into text since these techniques could greatly impact the outcome of the research. Emojis, on the other hand, were included in the analysis because their meanings are less contested and are more efficiently translated into plain text.

We built a script in Python that uses the Emoji library[4] in the translation of emojis into English. The pseudo-code of this step follows below:

Step 2: Pseudo-code used in the cleaning of tweets

```
get twitter users' ids
FOREACH twitter user id IN list of all twitter users' ids
        create a new csv file for the current user
        get current user csv file
        FOREACH tweet IN current user csv file
                remove links from the current tweet
                remove special characters from the current tweet
                translate to English the emojis of the current tweet
                save cleaned tweet data in the new csv file
        END
END
```

In the third step we submitted the preprocessed tweets to Perspective's API in order to obtain their toxicity probability. The probability values returned were stored in a new CSV dataset together with the tweets' content and their respective unique IDs. With this new dataset, it was possible to know how 'toxic' each of the analyzed tweets was according to Perspective's evaluation.

We built a script in Python that uses the Requests library[5] to submit the selected tweets to Perspective's API. The pseudo-code of this step follows below:

[4] Available at: https://pypi.org/project/emoji/.

[5] Available at: https://requests.kennethreitz.org/en/master/.

Step 3: Pseudo-code used to collect toxicity values from Perspective's API

```
get twitter users' ids
FOREACH twitter user id IN list of all twitter users' ids
        get current user csv file with the cleaned tweets
        create a new csv file for the current user toxicity values
        FOREACH tweet IN current user cleaned tweets csv file
                submit tweet through Perspective API
                get toxicity value
                save toxicity value in the current user toxicity file
        END
END
```

The Python source code of the algorithms employed in the three steps described above is available at GitHub and may be accessed on the following link: https://github.com/internetlab-br/ai_content_moderation.

In the final step toxicity values were analyzed. Here, we applied exploratory statistics with the intention of confirming/denying our hypothesis and discovering possible unknown patterns or useful information. For this purpose, we used basic techniques of data manipulation with Libre Office Calc, Python and R programming language to explore statistical functions. In addition, we resorted to data visualization with the software 'Tableau'. For this step no specific software scripting was developed.

Due to Twitter's Developer Policy,[6] which provides rules and guidelines for developers who interact with Twitter's applications and content, the authors decided not to publish the CSV dataset. The document sets forth several restrictions for that matter, limiting what could be disclosed in downloadable datasets. Additionally, it provides that any third party with access to the dataset would have to adhere to Twitter's ToS, Privacy Policy, Developer Agreement, and Developer Policy—the authors would not be in a position to guarantee this if the dataset was publicly available.

The tweets posted by drag queen accounts, Michelle Obama, Donald Trump, as well as by Richard Spencer, David Duke, StefanMolyneux, and Faith Goldy, were collected and submitted to Perspective's API in May 2019. In the second stage of the research, the number of white nationalists' accounts included in the analysis was increased to ten. The tweets posted by these accounts (Kevin MacDonald, James Edwards, Steve King, Jean-François Gariépy, Lana Lokteff, and Henrik Palmgren) were collected and submitted to the same API version model (Toxicity 6) in September 2019. After testing, in September 2019, a random sample of 100 tweets collected in May to find out whether the results would be different, we identified that their variation rates remained between 0 and 3.79%—except for one of them which varied in 8.96%. These minor changes

[6] Available at: https://developer.twitter.com/en/developer-terms/agreement-and-policy.html (accessed 17 September 2019). Link on Perma.cc: [https://perma.cc/RZM2-4LYW].

were probably caused by the retraining of the API's model, which occurs often according to Perspective's developers.[7] Considering these small variations do not impact the research findings, the data collected in September 2019 was included in the analysis.

Results and Discussion

Research results indicate that a significant number of drag queen Twitter accounts were considered to have higher perceived levels of toxicity than Donald Trump and white nationalists. On average, drag queens' accounts toxicity levels ranged from 16.68 to 37.81%, while white nationalists' averages ranged from 19.68 to 33.46%; Trump's lies at 21.84%, as indicated in the graph below (Michelle Obama in green; drag queens in blue, Donald Trump in orange, white nationalists in red) (Fig. 2).

The range of median toxicity levels was slightly different: 11.72% to 32.14% for drag queen accounts, while white nationalists' medians ranged from 13.97 to 28.25%; Trump's lies at 18.70%, as indicated in the graph below (Michelle Obama in green; drag queens in blue, Donald Trump in orange, white nationalists in red) (Fig. 3).

Taken individually, 5,912 tweets (5.59% of the total amount of 105,701 tweets analyzed) from drag queens were considered to have perceived levels of toxicity that are equal or higher than 70%—i.e., were deemed very likely to be toxic by Perspective' API. When looking into white nationalists' statistics, only 3.33%—491 out of the 14,712 tweets analysed—were considered to have perceived levels of toxicity equal or higher than 70% (Fig. 4).

The distribution of toxicity 70% + tweets posted by drag queens (Fig. 5) is considerably homogeneous across the range 70–100%: (1) 35.97% between 70 and 80%; (2) 29.33% between 81 and 90%; (3) 34.70% between 91 and 100%. There are two modes: 76% and 87%.

The distribution of toxicity 70% + tweets posted by white nationalists (Fig. 6) is considerably uneven across the range 70–100%, with a great deal of them with toxicity levels closer to 70%: (1) 65.76% between 70 and 80%; (2) 28.35% between 81 and 90%; (3) 5.88% between 91 and 100%. The mode is 71%.

The data presented above suggests the AI tool under analysis might in fact consider white nationalist narratives less toxic than LGBTQ speech. However, it is important to consider white nationalists could also write about other topics not necessarily related to white nationalist narratives—they could be posting relevant content in a small set of tweets. If that were the case, the fact that the average and median values of white nationalists are lower than those of the drag queens does not show per se that white nationalist narratives are not classified as toxic. In order to further investigate that possibility, a text analysis was performed as a means to

[7] For more information, access Perspective's profile at GitHub: https://github.com/conversationai/perspectiveapi/blob/master/api_reference.md.

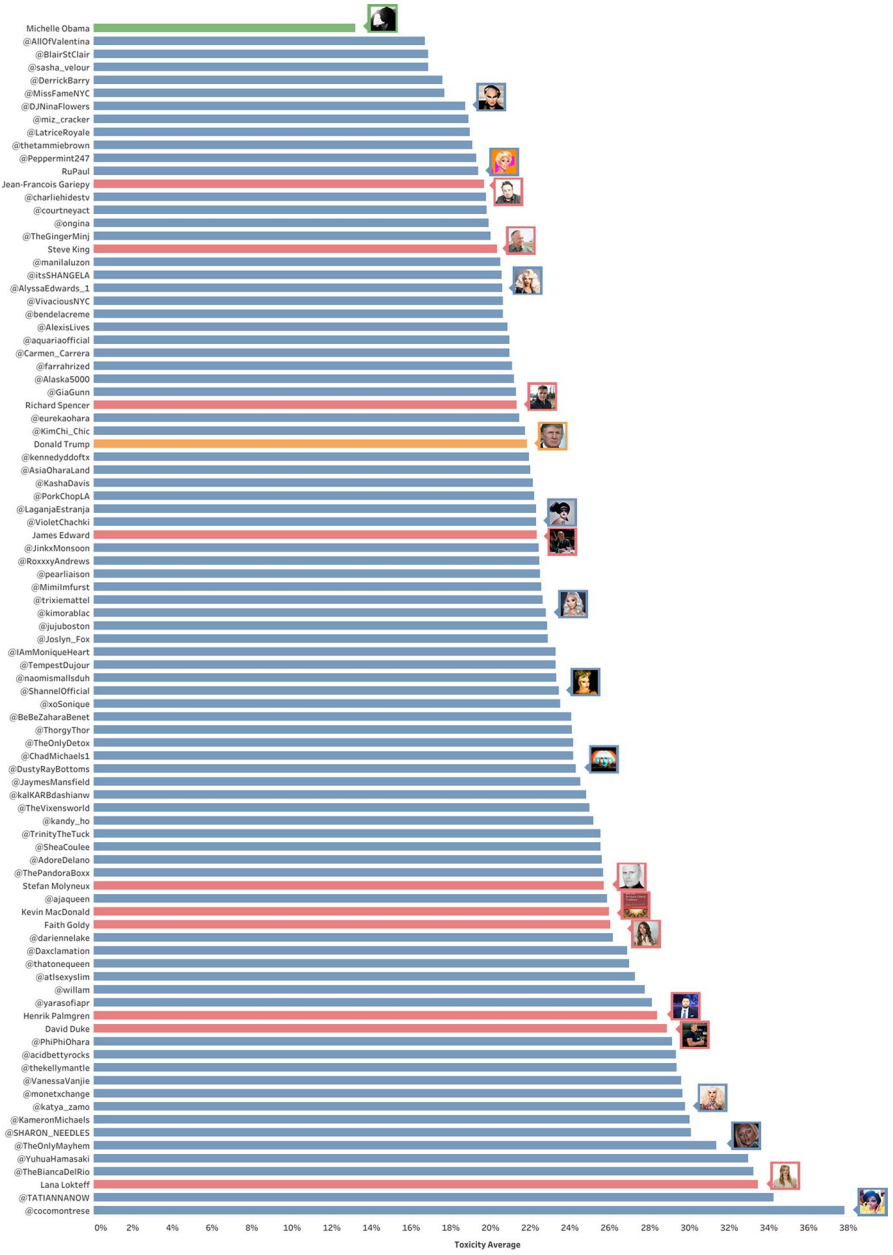


Fig. 2 Average toxicity level of Twitter profiles (Color figure online)

provide more information on the words most frequently mobilized by both groups in tweets with high toxicity levels (+ 70%) according to Perspective.

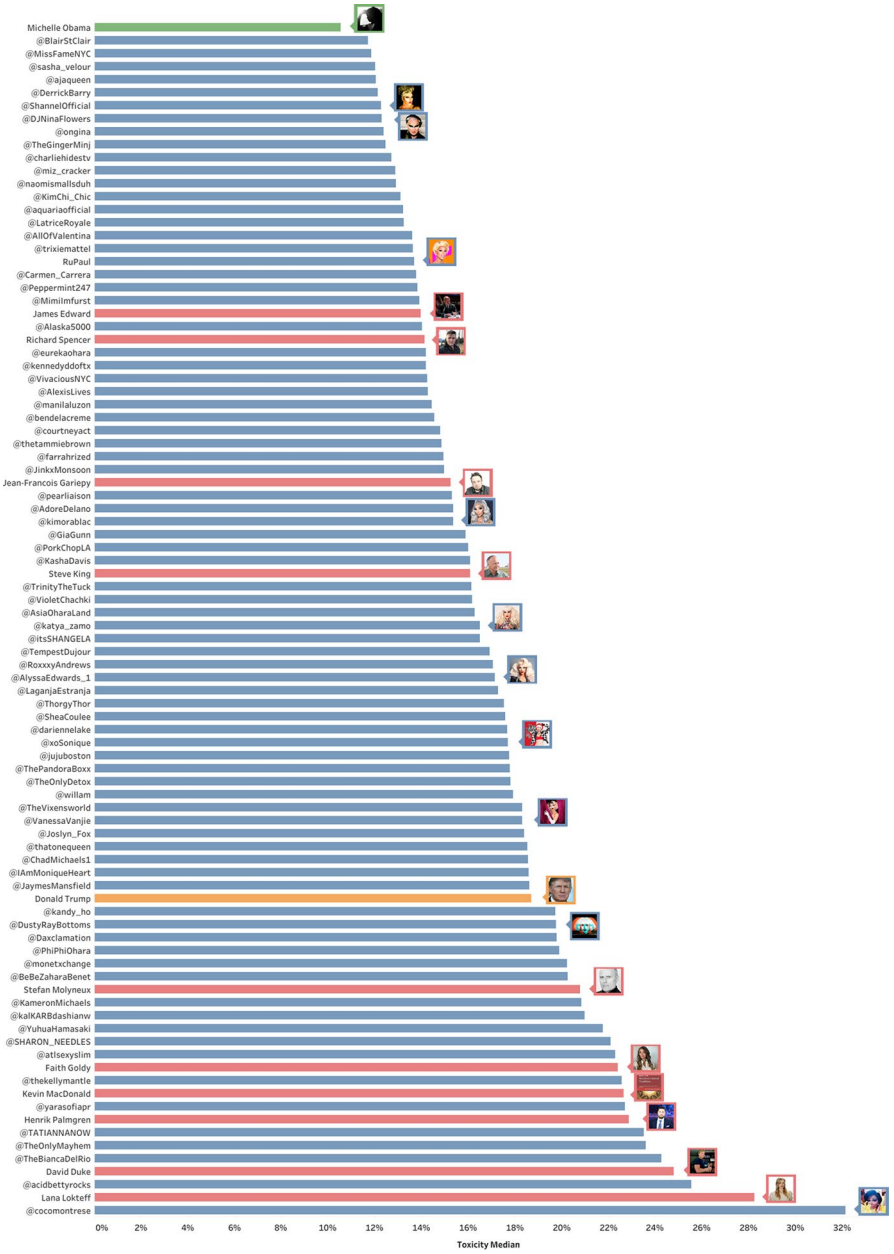


Fig. 3 Median toxicity level of Twitter profiles (Color figure online)

Several of the words that appear in Fig. 7 are connected to the LGBTQ context (i.e., 'gay', 'rupaulsdragrace', 'dragqueen', 'lgbt', 'drag', 'queen'), while others are common swear words, such as 'fuck', 'ass' or 'shit'. 'Bitch', which is the word that

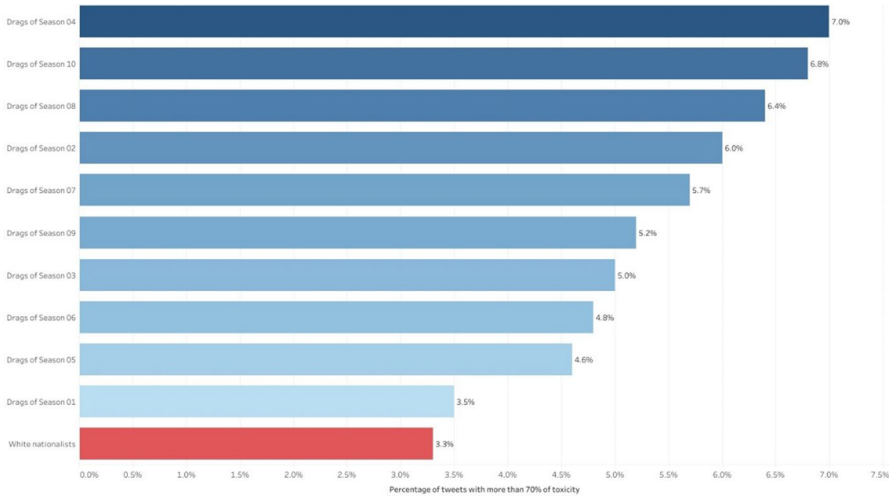


Fig. 4 Percentage of + 70% tweets

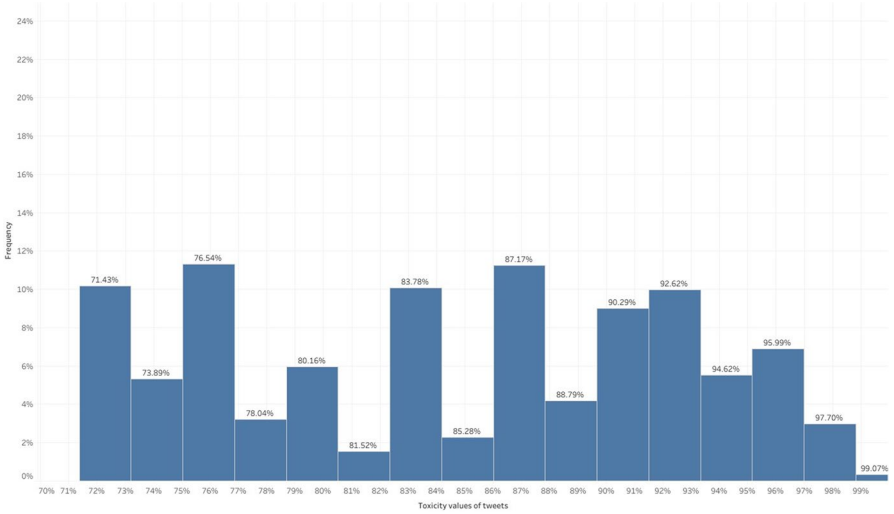


Fig. 5 Drag queens' + 70% tweets distribution

appears the most frequently among toxicity 70% + tweets, is used over 1000 times and accounts for 1.68% of the total word count of the sample. This is relevant considering 'bitch' is mostly used by drag queens to convey a neutral or positive meaning, which will be further discussed in the qualitative analysis. The appearance of some of these words—which could be considered offensive depending on the context—in the word cloud suggests the AI tool may misclassify LGBTQ content as toxic due to their high prevalence. This is similar to what Davidson et al. (2017)

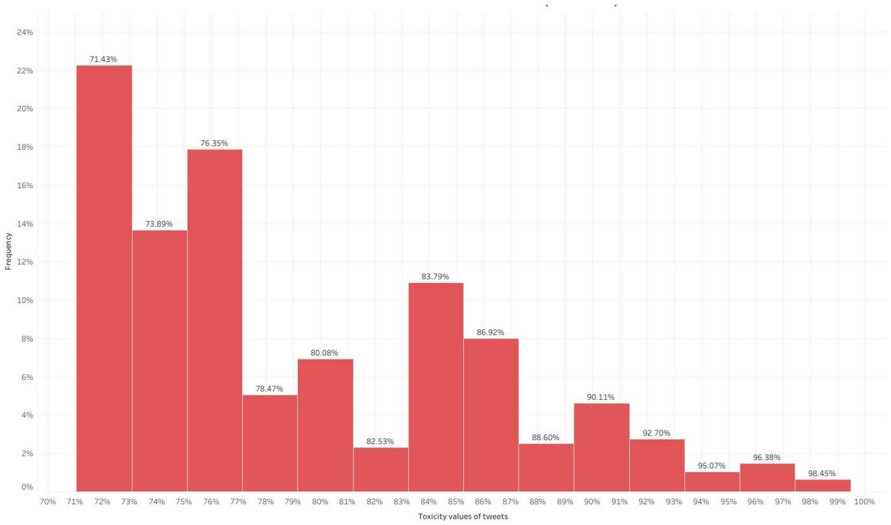


Fig. 6 White nationalists' + 70% tweets distribution

observed when analyzing the performance of their multi-class classifier, concluding 'misclassification appears to be caused by the presence of potentially offensive language' which is, in fact, used by many Twitter users without toxic intent in their everyday communications. It is also worth adding that worlds such as 'love' and 'heart' (in Emoji form) also appear in the word cloud.

We also ran tests measuring the levels of toxicity of words commonly found in the tweets from drag queens. They were individually (i.e. alone, without additional text) submitted to Perspective's API in May 2019. Some of these words alone had significantly high levels of toxicity: BITCH (98.18%); FAG (91.94%); SISSY (83.20%); GAY (76.10%); LESBIAN (60.79%); QUEER (51.03%); TRANSVESTITE (44.48%).

That means that, regardless of context, words such as 'gay', 'lesbian' and 'queer', which should be neutral, are already taken as significantly 'toxic'. This alone points to important biases in Perspective's tool. Additionally, even though other words such as 'fag', 'sissy' and 'bitch' might be commonly perceived as 'toxic', their use by members of the LGBTQ community serves a completely non-toxic purpose—building in-group solidarity and self-defense skills.

Figure 8, on the other hand, reveals common swear words less frequently. However, it indicates, considering the context of white nationalism supporters, a highly divisive speech, revealing words like 'white', 'black', 'slave', 'Jewish', 'Christian', 'Muslim', 'terrorist', 'hate', 'kill', 'murder', among others. The most frequently used word is 'white', accounting to 2.07% of the total word count of the sample, followed by 'people' (1.19%). The most frequent pair of words are 'white people' and 'white men'. The fact that these words, more closely associated with white nationalist narratives, appeared in the word cloud reveals that those narratives are significantly present in the sample. Even though they appear within the group of

Fig. 7 Drag queens' + 70% tweets world cloud

+70% tweets, they tend to be deemed less toxic then LGBTQ speech by the AI tool, considering tweet distributions discussed in Figs. 5 and 6. In this regard, Davidson et al. (2017) mentioned that such tools are more likely to misclassify offensive content if it does not contain any curse words or offensive terms.

In addition to the word clouds presented above, a more in-depth analysis was needed to rule out the possibility that white nationalists were not posting supremacist-related content often. In order to verify whether that was the case, we randomly selected a sample of the tweets we collected from each white nationalist account. Considering the differences in the number of tweets from each account, we decided to apply the sampling technique to the set of tweets we collected from each account separately.

The analysis was performed by five researchers who were instructed to label as 'white nationalist content' tweets which were found to convey 'divisive messages somehow supporting a white-only State, the belief white people are superior to other races, dehumanizing/attacking vulnerable groups and/or depicting them as enemies/dangerous'. However, it is important to highlight that this kind of analysis is to a certain degree subjective, considering people's understanding about what these divisive messages may entail varies. The tweets analyzed were randomly selected by a Python code using the shuffle function of the random module.[8]

We calculated the sample size using the finite population formula (below). The margin of error is 5%, with a confidence level of 90%:

$$\text{Sample size} = \frac{\frac{z^2 p(1-p)}{e^2}}{1 + \left(\frac{z^2 p(1-p)}{e^2 N}\right)}$$

[8] The source-code is available at: https://github.com/internetlab-br/ai_content_moderation.

Fig. 8 White nationalists' + 70% tweets word cloud

where N: total of tweets (population size), e: margin of error, z: Z value for confidence level (confidence level = 90% and Z value = 1.65, according to Z table).

The final sample resulted in a total of 2,139 tweets, which also represents about 14.5% of the tweets collected from each of the relevant Twitter profiles. The results varied considerably between the accounts, which suggests some of them share content supporting white nationalist narratives more often than others. The proportion of tweets directly related to these narratives varied between 17.07% (Steve King's profile) and 53.48% (Lana Lokteff's profile), as per Fig. 9.

In average, this proportion was 40.63%, what rules out the hypothesis according to which the selected Twitter profiles do not relevantly support white nationalist narratives. This claim is also corroborated by decisions recently taken by YouTube (Alexander 2020) and Twitter (Bogage and Scott 2020) to ban some of the white nationalists operating the Twitter profiles under analysis in our research (Stefan Molyneux, David Duke and Richard Spencer) due to repeated violations of the platforms' hate speech policies.

Qualitative Analysis

The qualitative analysis of individual tweets also illustrates how the biases identified in the research have significant implications for LGBTQ speech. If AI tools with such embedded biases are implemented to remove, police, censor, or moderate content on the internet, they may have profound impacts on the ability of members of the LGBTQ community to speak up and reclaim the use of 'toxic' or 'harmful' words for legitimate purposes, as is shown below:

(1) Cases in which the content should be considered neutral, but was likely considered toxic by Perspective because of a given word, like 'gay' or 'lesbian' (Tweets 1, 2 and 3).

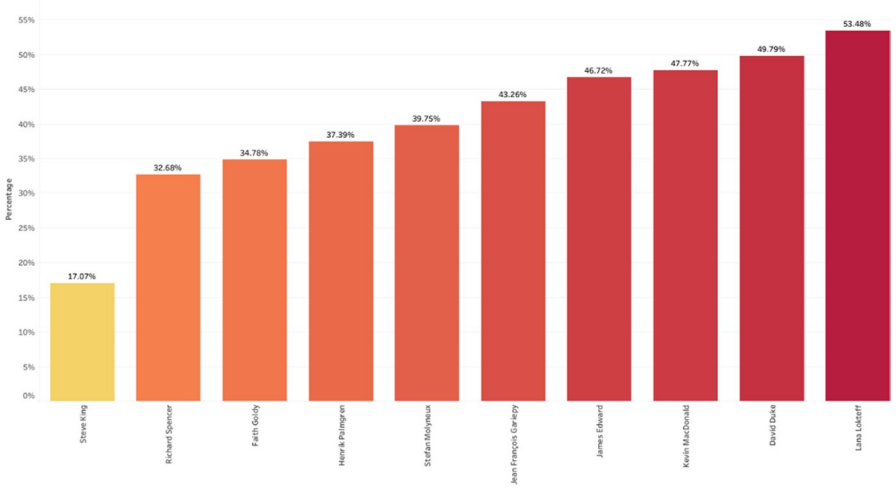


Fig. 9 Percentage of white nationalist content in the tweet sample of each profile

(2) Cases in which a word used could be considered offensive, but could also convey a different meaning in LGBTQ speech—even a positive one (Tweets 4 and 5).

(3) Cases in which words or expressions commonly used to attack LGBTQ people are reclaimed by LGBTQ people (Tweets 6 and 7).

(4) Cases in which drag queens may use strong language, but with a social and political value, often to denounce, or speak up against, homophobia and other forms of discrimination, such as sexism and racism (Tweets 8, 9 and 10).

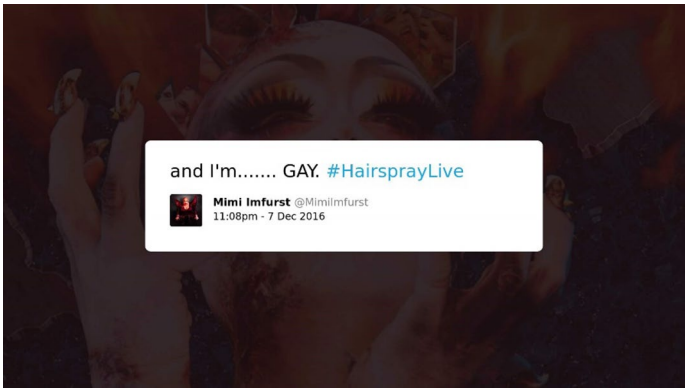
We also brought examples of tweets (Tweets 11 to 20) from white nationalists which, despite promoting discriminatory speech against minorities, were considered by Perspective to have low levels of toxicity.

Often, 'harsh' LGBTQ interactional practices address sensitive topics, like sexual roles in relationships, the visibility of gayness, and sexual promiscuity precisely because these subjects are frequently weaponized by those who aim to verbally attack LGBTQ people. When uttered by members of the LGBTQ community towards each other, these comments may not convey malice, but come from a place of solidarity to 'boost their armor'. The key is to understand that the underlying messages do not promote hate, prejudice, and discrimination. On the contrary, they redeploy 'toxic' words for self-empowerment.

The fight against hate speech is precisely one of underlying message. When toxic words promote hateful and discriminatory ideas, they represent a threat for marginalized and vulnerable groups. By training the algorithm to learn what pieces of content are more likely to be considered toxic, Perspective's tool seems to be giving a lot more weight to individual words rather than their underlying meanings. This was evident in the analysis of toxicity levels of words commonly found in the tweets posted by drag queens, which revealed the words 'bitch', 'fag', 'sissy', 'gay', 'lesbian' and 'queer' had significant high levels of toxicity by themselves—what could endanger LGBTQ expression in case the AI tool is employed for content moderation

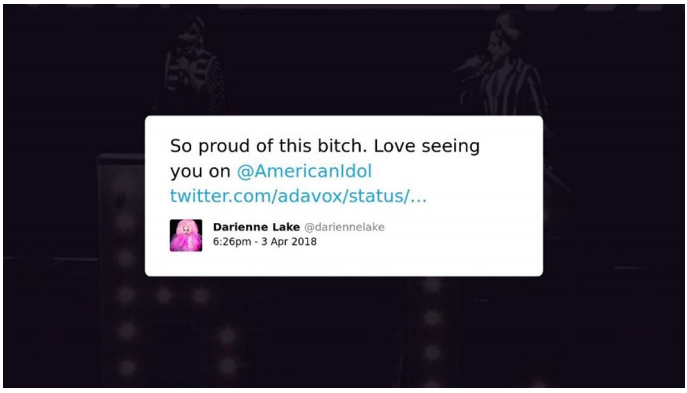
Tweet 1—toxicity level: 90.08%

Tweet 2—toxicity level: 90.85%



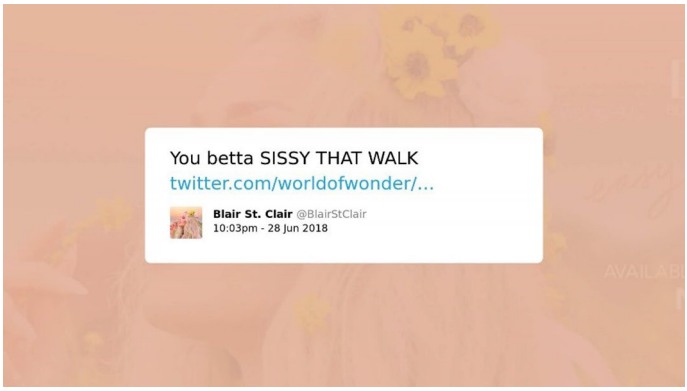
Tweet 3—toxicity level: 92.31%

Tweet 4—toxicity level: 90.75%



Tweet 5—toxicity level: 93.84%

Tweet 6—toxicity level: 94.56%



Tweet 7—toxicity level: 90.15%

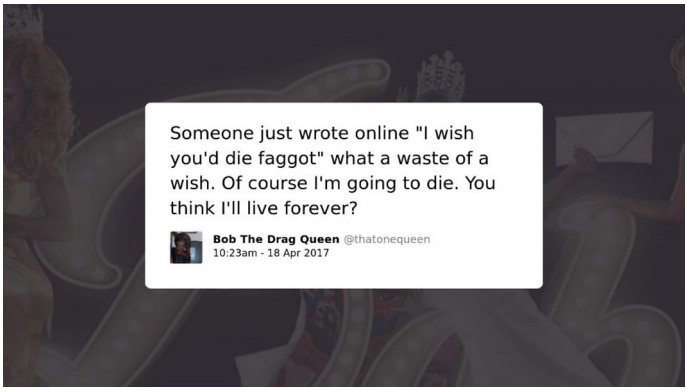
Tweet 8—toxicity level: 90.32%

purposes without human oversight, considering these words are most likely to have a neutral or positive meaning when used by members of the community.

When looking at tweets from white nationalists, on the other hand, even though the ideas they promoted were predominantly discriminatory, they were considered no more, or even less, toxic than empowering tweets from drag queens.

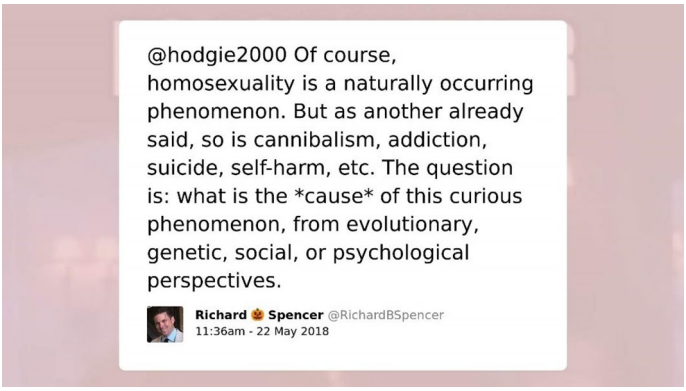
That means that if Perspective's AI tool had to decide which tweets should be removed, the drag queens' ones could be suppressed while those from white nationalists could be left up. In fact, Perspective is already making such evaluative decisions. In March 2019, Jigsaw launched Tune, an experimental browser plugin now available for Google Chrome, which uses Perspective to let users turn a dial up or down to set the 'volume' of online content. At launch, Tune worked across different social media platforms including Facebook, Twitter, YouTube, and Reddit. Users can turn the knob up to see everything or turn it down all the way to a 'hide all' mode, replacing toxic comments with small colored dots. Tune markets itself around

Tweet 9—toxicity level: 90.00%



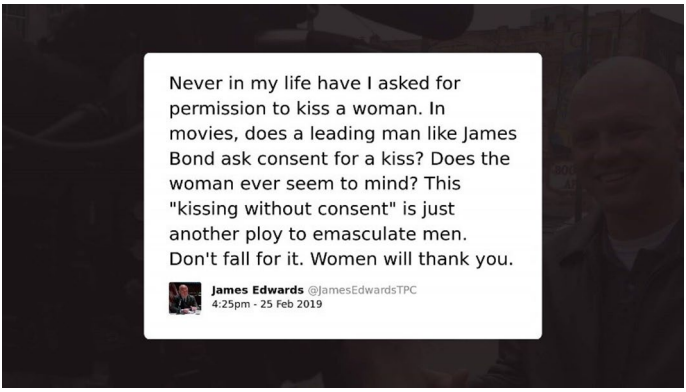
Tweet 10—toxicity level: 91.88%

Tweet 11—toxicity level: 11.71%

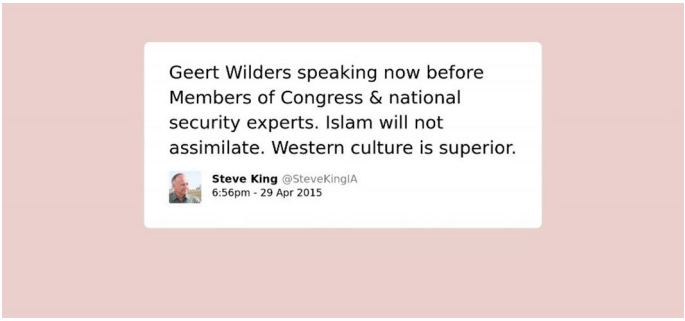


Tweet 12—toxicity level: 13.80%

Tweet 13—toxicity level: 19.01%



Tweet 14—toxicity level: 20.35%



Tweet 15—toxicity level: 23.31%

Tweet 16—toxicity level: 24.42%

Tweet 17—toxicity level: 32.11%

Tweet 18—toxicity level: 33.37%

Tweet 19—toxicity level: 34.99%

Tweet 20—toxicity level: 36.85%

Fig. 10 Yuhua Hamasaki's profile on Twitter with tune on 'hide all' mode

the idea that 'machine learning technology can create new ways to empower people as they read discussions online.' Jigsaw explains, however, that Tune 'still misses some toxic comments and incorrectly hides some non-toxic comments' (Adams 2019) (Fig. 10).

Suggestions on How to Improve the Accuracy of AI Tools Designed to Analyze Text-Based Content

The findings described above support the claim that AI tools developed to analyze text-based content are not able to properly understand context. There are, however, strategies that could improve the accuracy of these tools. García et al. (2019), as well as Hazarika et al. (2018), employed, for instance, a combination of different text mining techniques, while other researchers explored extra textual features, such as hashtags (Nadali et al. 2016), emojis (Felbo et al. 2017), and punctuation (Rangwani et al. 2018).

García et al. (2019) attempted to improve the accuracy of AI tools in the identification of irony in tweets written in Spanish by carrying out an extensive set of experiments which combined different text mining techniques. The researchers were able to demonstrate that the combination of feature extraction techniques with word and document embedding presents the best results for the domain they were working in.

In order to create a tool able to detect and classify emotions—including sarcasm—in tweets, Felbo et al. (2017) trained a model using a dataset of more than 1 million tweets sharing different types of emojis. The researchers reported positive results from this approach, highlighting the good performance was related to the big variety in the use of emojis represented in the dataset. In a similar study, Rangwani

et al. (2018) used emojis to improve irony detection in tweets written in English combining techniques of feature extraction and an external pre-trained emoji model. By adopting this approach, the researchers were able to identify, apart from keywords related to irony, a set of linguistic patterns in ironic writings such as a contrast of feelings and punctuation repetition.

Another study (Nadali et al. 2016) discussed the use of hashtags in tweets to improve sarcasm detection tools. The hashtag '#sarcasm' has been extensively explored for the creation of datasets with sarcastic content; it was employed by Nadali, Murad and Sharef in the identification of patterns of sarcasm in text-based content. Additionally, the researchers were able to indicate a list of other hashtags that can also be used to detect the presence of sarcasm in tweets.

Another resource used to obtain additional context information is stylometry, a technique employed to analyze an author's writing style through various quantitative criteria, identifying patterns. Hazarika et al. (2018) applied stylometry to recognize sarcasm in written text produced by a selected group of people. Considering that people express sarcasm in particular ways, the researchers created custom models for users of a Reddit forum using their individual messages posted on a discussion thread. The authors concluded that this kind of individual modeling presented better results in the task of identifying sarcasm.

In order to improve the performance of tools developed to identify and classify hate speech, researchers have also explored alternative techniques and resources to collect context data. Yang et al. (2019), for instance, noted that researchers usually discard imagens when attempting to identify hate speech on social media. Considering that images may add important contextual information, Yang et al. enriched their textual analysis with image embedding information by using multimodal techniques—and reported a considerable boost in the performance.

Mishra et al. (2019) employed network analysis metrics to gain accuracy when detecting hate speech. The researchers used a dataset of 16,000 tweets previously labeled as racist and sexist, as well as data on the connections between users whose tweets were in the dataset. Mishra et al. represented these connections in an undirected graph to understand how users were connected to each other and how these connections were related to the tweets from the dataset. As their main finding, the researchers reported that users who posted sexist tweets were only a few degrees apart from each other. Additionally, users who posted racist tweets were also likely to post sexist tweets but were isolated users in the graph representing the whole community. Mishra et al. also reported that this approach significantly improved the performance of the hate speech detection tool.

Another approach worth exploring is discussed by Taylor et al. (2017). The researchers observed that many new hate speech related words are produced in extremist online communities—thus not immediately identified by algorithms of automatic detection, what enables these words to gain visibility on social media. With this in mind, Taylor, Peignon and Chen presented a method for automatically learning new harmful words by improving hate speech detection algorithms with messages scraped from extremist communities.

Concluding Remarks

This study corroborates findings discussed in papers reviewed in the first sections of this article: AI tools developed to analyze text-based content are not yet able to understand context. Differently from these other studies, however, this article approaches the issue from the perspective of the LGBTQ community to highlight how content moderation technologies could affect LGBTQ visibility. Given the specific communicational practices of the community, particularly of drag queens, the use of Perspective—as well as other similar technologies—to police content on internet platforms could hinder the exercise of free speech by members of the community.

The most probable explanation for the findings described in the sections above is that machine learning techniques find correlation between input features (words) and target classification (toxicity). When processing datasets, algorithms establish these correlations based on the probability of a given word or expression appearing in content labeled as 'toxic' by the annotators—those who previously classified the whole dataset into two classes: 'toxic' and 'non-toxic'. Considering the words mentioned above are more commonly—i.e., in general speech, outside the LGBTQ community—associated with toxic content, machine learning algorithms infer that using those words makes the text more likely to be toxic. In other words, they will get attached to any features that are relevant for the classification, even if those features only indirectly correlate with it.

LGTBQ terms indirectly correlate with toxicity, not because of their inherent toxicity but due to their association with toxic content. In this regard, it is important to emphasize that some of these words are frequently used by people in general as swearing words due to the prevalence of heterosexism and cisgenderism in society—which means the 'toxicity' attributed to them have its roots in the same hierarchical system that places the LGBTQ community in a vulnerable position. Since machine learning algorithms make no distinction between direct and indirect correlation, they end up associating LGBTQ terms with toxicity, thus reproducing prevailing ideas about toxicity and the LGBTQ community. This results in the problem that words become associated with toxicity even if they themselves are not toxic.

This is not a flaw, since these algorithms make their decisions based on language alone, irrespective of who says it and in what context. However, it is important to bring this to the attention of non-technical audiences, considering AI tools are frequently mentioned as bulletproof solutions to all sorts of problems—including content moderation—by technocentric narratives.

Since AI technologies such as Perspective do not seem to be fully able to grasp social context or recognize specific uses of 'toxic' words as socially or politically valuable—as it is the case with the LGBTQ 'insult rituals'—it is likely that automated decisions based on AI technologies would suppress legitimate content from LGBTQ people. This is particularly interesting when one of the main reasons behind the development of these tools is to support vulnerable communities by dealing with hate speech targeting such groups. If these tools might prevent

LGBTQ people from expressing themselves and speaking up against what they themselves consider to be toxic, harmful, or hateful, their net impact may be disempowering, rather than helpful.

There are a few strategies worth exploring to improve the accuracy of AI technologies developed to analyze text-based content by adding contextual information. Studies already show that combining different text mining and multimodal techniques in the analysis performed by these tools significatively improves their performance. The same applies to the incorporation of extra textual features, such as hashtags, emojis, punctuation and data on the connections between users.

In any event, risks posed by AI are the reason why algorithms should not be the sole basis for reaching decisions that directly affect communications online. This is not to say AI tools should not be developed and employed by internet platforms for content moderation purposes—they are indeed important to support the huge task of freeing the internet from unwanted content. Nevertheless, while they can assist human moderators by flagging media for analysis, they should not be employed without human oversight.

In the long term, there also needs to be a more profound discussion about how those tools might impact, change, and shape the way we all communicate. If an AI tool considers the word 'bitch' to be 'inappropriate' or 'toxic', does that mean we should stop using it? If computers decide what is 'toxic' on the internet, what does that mean to the future of speech and to the way we decide to express ourselves, on the internet and outside of it?

Acknowledgements The authors are grateful to Timothy Rosenberger for his editing and review support; to Ester Borges, Clarice Tavares and Victor Pavarin Tavares for their research support.

Funding The authors received no financial support for the research, authorship, and/or publication of this article.

Availability of Data and Material Due to Twitter's Developer Policy, which provides rules and guidelines for developers who interact with Twitter's applications and content, the authors decided not to publish the CSV dataset. The document sets forth several restrictions for that matter, limiting what could be disclosed in downloadable datasets. Additionally, it provides that any third party with access to the dataset would have to adhere to Twitter's ToS, Privacy Policy, Developer Agreement, and Developer Policy—the authors would not be in a position to guarantee this if the dataset was publicly available.

Code Availability The Python source code of the algorithms developed to be employed in the research are available at GitHub and may be accessed on the following link: https://github.com/internetlab-br/ai_content_moderation.

Compliance with Ethical Standards

Conflict of interest There are no conflicts of interest.

References

Adams, C. J. (2019). Tune: Control the comments you see. In: *Medium*. https://medium.com/jigsaw/tune-control-the-comments-you-see-b10cc807a171. Accessed 31 Oct 2020.

Alexander, J. (2020). YouTube bans Stefan Molyneux, David Duke, Richard Spencer, and more for hate speech. The Verge. https://www.theverge.com/2020/6/29/21307303/youtube-bans-molyneux-duke-richard-spencer-conduct-hate-speech. Accessed 31 Oct 2020.

BBC (2017). YouTube 'made wrong call' on Syrian videos. https://www.bbc.com/news/technology-41023234. Accessed 31 Oct 2020.

Bogage, J., & Scott, E. (2020) Twitter permanently bans former KKK leader David Duke. Washington Post. https://www.washingtonpost.com/technology/2020/07/31/twitter-david-duke-ban/. Accessed 31 Oct 2020.

Curtis, S. (2015). Facebook, Google, and Twitter block 'hash list' of child porn images. Telegraph. https://www.telegraph.co.uk/technology/internet-security/11794180/Facebook-Google-and-Twitter-to-block-hash-list-of-child-porn-images.html. Accessed 31 Oct 2020.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th international AAAI conference on web and social media*, https://arxiv.org/abs/1703.04009

Duarte, N., Llanso, E., & Loup, A. (2017). Mixed messages? The limits of automated social media content analysis. Center for Democracy and Technology. https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis. Accessed 31 Oct 2020.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, https://arxiv.org/abs/1708.00524.

García, L., Moctezuma, D., & Muñiz, V. (2019). A contextualized word representation approach for irony detection. In *Proceedings of the Iberian languages evaluation forum 2019*, https://ceur-ws.org/Vol-2421/IroSvA_paper_5.pdf.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality and society* (pp. 167–193). Cambridge: The MIT Press.

Gröndahl, T., Pajola, L., Juuti, M., Conti, T., & Asokan, N. (2018). All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, https://doi.org/10.1145/3270101.3270103.

Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., & Mihalcea, R. (2018). CASCADE: Contextual sarcasm detection in online discussion forums. arXiv ahead of print 16 May 2018. https://arxiv.org/abs/1805.06413.

Heisterkamp, B. L., & Alberts, J. K. (2000). Control and desire: Identity formation through teasing among gay men and lesbians. *Communication Studies, 51*(4), 388–403.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's perspective API built for detecting toxic comments. arXiv ahead of print 27 February 2017. https://arxiv.org/abs/1702.08138.

Jones, R. G. (2007). Drag queens, drama queens and friends: Drama and performance as a solidarity-building function in a gay male friendship circle. *Kaleidoscope, 6*(1), 61–84.

Johnson, E. P. (1995). SNAP! culture: A different kind of "reading." *Text and Performance Quarterly, 15*(2), 122–142.

Keller, D. (2018). Internet platforms: Observations on speech, danger and money. Hoover Institution. https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf. Accessed 31 Oct 2020.

Levin, S. (2017). Civil rights groups urge Facebook to fix 'racially biased' moderation system. The Guardian. https://www.theguardian.com/technology/2017/jan/18/facebook-moderation-racial-bias-black-lives-matter. Accessed 31 Oct 2020.

Lessig, L. (2006). *Code and other laws of cyberspace—version 2.0*. New York: Basic Books.

Lux, D., & Mess, L. M. H. (2019). Facebook's hate speech policies censor marginalized users. Wired. https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/. Accessed 31 Oct 2020.

McKinnon, S. (2017). "Building a thick skin for each other": The use of 'reading' as an interactional practice of mock impoliteness in drag queen backstage talk. *Journal of Language and Sexuality, 6*(1), 90–127.

Mishra, P., Tredici, M., Yannakoudakis, H., & Shutova, E. (2019). Author profiling for hate speech detection. arXiv ahead of print 14 February 2019. https://arxiv.org/abs/1902.06734.

Murray, S. O. (1979). The art of gay insulting. *Anthropological Linguistics, 21*(5), 211–223.

Nadali, S., Murad, M., & Sharef, N. (2016). Sarcastic tweets detection based on sentiment hashtags analysis. *Advanced Science Letters, 22*(4), 400–407.

O'Brien, L. (2019). Twitter still has a White Nationalist Problem. *HuffPost*. https://www.huffpostbrasil.com/entry/twitter-white-nationalist-problem_n_5cec4d28e4b00e036573311d?ri18n=true. Accessed 31 Oct 2020.

Perel, M., & Elkin-Koren, N. (2016). Accountability in algorithmic copyright enforcement. *Stanford Technology Law Review, 19*(3), 473–533.

Perez, J. (2011). Word play, ritual insult, and volleyball in Peru. *Journal of Homosexuality, 58*(6), 834–847.

Rangwani, H., Kulshreshtha, D., & Singh, A. (2018). NLPRL-IITBHU at SemEval-2018 task 3: Combining linguistic features and emoji pre-trained CNN for irony detection in Tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, https://www.aclweb.org/anthology/S18-1104.pdf.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, https://www.aclweb.org/anthology/P19-1163.pdf.

Solon, O. (2017). Facebook asks users for nude photos in project to combat 'revenge porn'. *The Guardian*. https://www.theguardian.com/technology/2017/nov/07/facebook-revenge-porn-nude-photos. Accessed 31 Oct 2020.

Suzor, N. (2018). Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media + Society*. https://doi.org/10.1177/2056305118787812

Taylor, J., Peignon, M., & Chen, Y. (2017). Surfacing contextual hate speech words within social media. arXiv ahead of print 28 November 2017. https://arxiv.org/abs/1711.10093.

Yang, F., Peng, X., Ghosh, G., Shilon, R., Ma, H., Moore, E., & Predovic, G. (2019). Exploring deep multimodal fusion of text and photo for hate speech classification. In: *Proceedings of the third workshop on abusive language online*. https://www.aclweb.org/anthology/W19-3502.pdf.