Identifying Online Hate Speech Literature Review and Project Proposal

Hannah Koizumi
School of Data Science
University of Virginia
Charlottesville, USA
hek3bm@virginia.edu

Lorinda Couch
School of Data Science
University of Virginia
Charlottesville, USA
xkr9pe@virginia.edu

Thomas Lever
School of Data Science
University of Virginia
Charlottesville, USA
tsl2b@virginia.edu

***Abstract*—The rise in online hate speech over the past 10 years has led most social media companies to update their policies to explicitly condemn and remove hate speech and ban its perpetrators. However, the constant evolution of hate speech and the tremendous amount of content to analyze are significant barriers to detection. There are many machine and deep learning models created and tested by researchers; however, lack of transparency of the models and the data create a lack of reproducibility. In our experiment, we propose testing and possibly improving pre-trained existing binary classifiers on Reddit data before and after 2020, when Reddit made significant changes to its policies to address hate speech. In this way, we can contribute more information on the available models' performance. We are hoping to detect trends in online hate evolution.**

***Index Terms*—natural language processing, language models, text analysis, language parsing and understanding, neural models, ethical/societal implications**

I. Introduction

The ever-expanding corpus of online hate speech remains a relevant topic of public discourse around free speech and what responsibility social media companies have to monitor content on their platforms. In the United States, this increase in online hate speech led some social-media companies to tighten bans on certain types of speech by changing their user policies. However, there remains one particular vexing issue for United States based social media —the lack of a "blanket definition of hate speech under American law, which is generally much more permissive than other countries because of the First Amendment to the US Constitution" [1].

On the flip side of online regulation is the European Union. Its Digital Services Act (DSA), enacted in October 2022 and fully implemented on February 17, 2024, creates a comprehensive framework of regulatory standards to:

- ”provide a safe digital environment free from illegal content
- enhance transparency and accountability on the service provider side
- strengthen the protection of fundamental European and consumer rights
- strengthen enforcement (esp. in the cross-border situations)" [2]

Most major social media platforms have banned hate speech and use automated hate speech detection [3]. For example, in 2017, Twitter implemented new policies to remove hateful and abusive speech from their platform [4]; however, research has shown an increase in hate speech on Twitter (now X) after Elon Musk purchased the company [1]. In 2020, Reddit updated their content policy in an attempt to discourage inappropriate speech, specifically citing the desire to reduce hate speech [5]. It is this shift in policy that is the basis of our project —identifying online hate speech.

Complicating moderation of hate speech is algospeak. Algospeak rose as "... a communicative practice in reaction to experiencing content moderation on a platform" and to "circumvent... algorithmic content moderation [6]. Algospeak is "commonly understood as abbreviating, misspelling, or substituting specific words... when creating a social media post with the particular goal to circumvent a platform's content moderation systems" [6]. Algospeak has its roots in a backlash to Google Jigsaw's Conversation AI. Introduced in 2016 as a tool to "... use machine learning to automatically spot the language of abuse and harassment – with, Jigsaw engineers say, an accuracy far better than any keyword filter and far faster than any team of human moderators" [7].

A group, ultimately self-named Operation Google, was formed on 4chan, two days after Google introduced the model [8]. Word substitutions that were voted on and accepted included pepe for alt-right, car salesman for liberals/democrats, and reagans for conservatives. Included in the list are also substitutions for more offensive terms related to race, religion, etc. [8]. Reddit (r/technology) has a thread dedicated to this topic [9] from 2016 which was not removed in the 2020 content policy update purge. This thread was likely not removed due to the use of algospeak throughout —see post by TheFAPnetwork (account suspended) in this thread. The limits of automatic hate speech detection is a challenge. Our research for this project has not yet specifically identified deep learning models that have been adapted to understand algospeak.

To support our project proposal, we reviewed scholarly literature related to the detection of online hate speech through deep learning methods. We also identified potential sources for labeled and unlabeled Reddit data to be used in conjunction with pre-trained language models. Our goal is to examine the performance of and possibly improve specific pre-trained hate speech or other Natural Language Processing (NLP) deep learning models on Reddit submissions and comments prior to and after Reddit made these policy changes.

II. Literature Review

A. Summary of Existing Literature

It is widely known that social media companies are not well-equipped to filter through the immense amount of inappropriate speech on their platforms. There are different facets of hate speech online and, as Jahan and Oussalah [10] point out, there are distinctions in definitions of hate speech based on country of origin, sector (academic versus political), and among social media platforms themselves. For the purpose of this literature review, the definition of hate speech will remain flexible enough to focus on the various methodologies used to identify inappropriate speech. Such inappropriate speech can also be defined as hateful, abusive, violent, etc.

1) Online Social Media Review Process: Pre-published and post-published reviews are the two current approaches to monitoring social media posts. Pre-published social media is either human reviewed or put through a simple filter prior to posting online. One study's authors write, "Human moderation of all content is expensive and lacks scalability, while word filters lack the ability to detect more subtle semantic abuse" [11]. Post-published review allows posts to be published online. Once posted, platform moderators or crowd-sourced moderation identify negative posts which, if not removed by this filtering technique, could have significant negative consequences for the intended targets [11]. The lack of reliable and non-human intensive mechanisms which filter and remove hate speech is a main motivation for researchers to build and test machine learning models.

2) Data Availability and Distribution: The volume of data posted on social media platforms is not the only barrier to correctly identifying and subsequently removing hate speech. Imbalanced class distribution with extremely low levels of abusive speech create challenges to models [11], especially models trained on unlabeled data. Nuances in online speech also create detection issues due to the subtle ways that hate speech can be portrayed through sarcasm or stereotypes [12]]; as a result, models may overlook hate speech. Though the work in non-English hate speech has increased, English, because initial modeling work was accomplished with an English corpus, remains dominant. Inroads have been made in other language models and the number of "... non-English hate speech detection toolkits have seen a steady increase" [10].

Data augmentation is one method that can be used to overcome imbalanced data sets in deep learning models. A new technique, Imbalanced Data Augmentation (IDA) can improve the prediction of the minority class (hate speech) by "... enhancing the size of the minority class by replacing uninformative words while preserving the class label, and secondly, augmenting the entire data size by inserting stop words into each data sentence" [13].

3) Preprocessing: Preprocessing is an important step for deep learning models. First, this step allows for data to be cleaned and standardized, thus reducing noise. Second, normalization, as a preprocessing step, standardizes the data. For text data, preprocessing can include converting to all lowercase, stemming, and lemmatization. Tokenization is another key process that breaks the text into words or phrases. Removing stop words such as "and", "the", "or" and "a" reduces the dataset size and can improve processing speed. Lastly, transforming text into an appropriate format such that a model can easily use the data involves conversion of data into numerical vectors via techniques such as Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) and word-embedding techniques such as Word2vec or Global Vectors for Word Representation (GloVe). This conversion allows models to understand and generate patterns. Ultimately, preprocessing steps allow models to train faster and more efficiently, thereby reducing the load on resources [14].

4) Traditional versus Deep Learning Model Performance: Despite challenges discussed in the Data Availability and Distribution section, some studies show that deep learning models outperform traditional machine learning classifiers [15], [16]. Typical machine learning classifiers such as Support Vector Machines (SVM) and Naïve-Bayes classifiers lack the ability to capture syntactic information and also generally ignore word order [11]. SVM models were shown to perform better on smaller rather than large datasets, which could be problematic in this era of big data. Examples were cited where SVM performance was poor with balanced datasets but where performance improved when oversampling methods were applied to the datasets [12].

Deep learning models tend to perform better than traditional machine learning models on imbalanced datasets because of their greater capacity to learn complex patterns from data. As discussed above, data augmentation can expand the minority class to help balance an otherwise imbalanced dataset [11].

Chen, McKeever and Delany [11] compared the performance of a Convolutional Neural Network (CNN) to a Recurrent Neural Network (RNN) in identifying hateful content and found improvement with an ensemble of classifiers rather than a single classifier; others have agreed that ensembles of classifiers are superior to single classifiers [11]. In a survey of hate speech detection, the authors found that CNN, RNN, and BERT (Bidirectional Encoder Representations from Transformers) "... have demonstrated remarkable abilities in detecting hate speech and analyzing sentiment in online content since their introduction [17].

5) Pre-Trained Models: Transformer-based deep learning models using attention mechanisms alongside predictions may improve hate speech detection results. "These models are typically pre-trained on extensive text corpora and subsequently fine-tuned for specific downstream tasks, enabling them to deliver exceptional performance even with limited training data" [12]. Among transformer-based models, BERT is rising in popularity with multiple researchers citing it as the superior deep learning architecture [10]. Other recent transformer-based models include BERT variants such as, but not limited to, FinBERT, RoBERTa, ALBERT, and DistilBERT [18].

B. Evaluation of Current Literature

A fair amount of scholarly literature has been written on the proposed project topic, especially if the definition of hate speech is expanded to include online speech that is harassing, violent, or abusive . Most authors agree that the starting point must be to clearly define the type of speech being identified and that there likely will never be sufficient amounts of training data available compared to the amount of online hate speech posted every day.

Zimmerman, Kruschwitz, and Fox [19] point out several critiques of the body of literature on hate speech detection models. The first critique is a lack of consistency in evaluation methods, which makes it difficult to compare results between individual studies [11]. The second is that there exists a low standard of reproducibility because researchers do not include important details related to their models such as weight initialization schemes and generally do not share their code.

The implication of these shortfalls is that researchers may find it difficult to trust these findings without the ability to recreate published results. The inability to replicate these studies can also contribute to poor community consensus building related to these types of deep learning models. Some studies show that CNN models outperform long short-term memory (LSTM) models while other studies find the opposite [10]. Bashar and Nyak [20] compared a CNN model to a LSTM model and found that CNN models were better at discovering larger patterns within a post or comment. This finding conflicts with the findings of Badjatiya, Gupta, Gupta, and Varma [21] who found that LSTM models performed better than CNN models. Overall, this lack of transparency in previously tested models is a disservice to the research and the public at large on hate speech detection.

C. Proposed Future Research

One theme in researchers' recommendations related to future research potential is different hate speech classification methodologies. These methodologies include transfer learning [18], LSTM for modeling sequences of characters [22], and different preprocessing techniques [12]. Our research indicates that the full arsenal of deep learning methods likely have not been tested on hate speech data as only 22% of the algorithms used for hate speech detection are identified as deep learning algorithms [6]. The lack of deployment of deep learning methods could be the result of the relatively recent resurgence of deep learning models.

Also of concern is a shift in certain online speech to algospeak and its ability to evade prior and current moderation whether human or machine. With minimal search efforts, hate speech from 2016 was easily found on Reddit in March 2024 —one may wonder what else lies behind the wizard's curtain yet in plain sight. Jahan and Oussalah build upon this observation with a discussion of annotation quality and the difficulty of standardization due to "loose" grammatical structure and "cross-sentence boundaries" [10]. The authors recommend that datasets for hate speech should be continually updated as language evolves, and that good guidance should be provided to annotators. It was noted a good deal of labeling is done via crowdsourcing [10].

Another set of future research recommendations comes from Zimmerman, Kruschwitz, and Fox [19], who recommend conducting a comparison of existing model weighting schemes in order to provide reproducibility assurances to this field of research. Multiple researchers recommended further testing of existing models on different datasets to prove their robustness [11], [19]. Overall, this research field needs reproducible testing on existing open-source data to further confidence in these models.

Many, if not most, studies on hate speech are grounded in English datasets. Though one study identified hate speech research completed in 21 different languages, opportunities exist for further research grounded in non-English datasets [10]. Lastly, exploration of transfer learning techniques using prior knowledge of hate speech datasets may assist in identification and removal of hate speech with minimal resources [17].

III. Project Proposal

A. Motivation

The motivation behind this project is the belief that deep learning models may have great potential to process and detect hate speech in a large sea of content likely contaminated with hate speech leading up to the 2024 U.S. presidential election. Moderating online content, while respecting the generally recognized First Amendment right to freedom of speech, is and will continue to be one of the greatest challenges of our generation, particularly with the already significant increase in User-Generated Content (UGC) and Artificial Intelligence Generated Content (AIGC).

Our proposed project will attempt to connect data science with social media policy changes. Through this connection, we may be able to trace the evolution of hate speech on the Reddit platform as users reacted to these changes. This approach could offer new perspectives to researchers related to trends in hate and violent speech. Researchers may want to consider these trends when creating new models or updating existing models used to detect hate speech. There is no doubt this type of modeling along with frequent updates to these models will be an ongoing need as the online social media landscape constantly adapts and evolves.

B. Dataset

Jahan and Oussalah, in 2023, write that "there are no commonly accepted datasets recognized as ideal for automatic HS [Hate Speech] detection tasks" [10]. They also discuss how datasets are dissimilar across research efforts due to the requirements of the project. They note that in 69 datasets there were 47 different labels and that most of the datasets were small and imbalanced [10].

The goal of identifying datasets for this project is to be able to use and possibly improve pre-trained models on Reddit data dated prior to and dated after the 2020 policy shift. We have not yet determined the best data compilation strategy. Potential compilation strategy approaches are listed below.

One dataset compilation strategy considered is to use a torrent containing a data dump of Reddit comments and submissions. The site (academictorrent.com) offers Reddit data in two formats — by year and by subreddit. The data files are large with the entirety over 2.5 terabytes. The data are in Zstandard-compressed NDJSON files, which are more challenging to work with than a standard JSON file. Most of the uncompressed files are over 20 gigabytes (GB). Rivanna's upload limit and local storage resources will present a challenge. There are available Python scripts to search the compressed files for particular terms. Data could be labeled or left unlabeled.

Another compilation strategy would be to narrow the window of interest to 2016 and manage data as above. We could identify or create an algospeak model based on a restricted dataset. We could use such a model to identify 2016 posts that are still on Reddit and qualify as hate speech, as described in an above discussion. One research team took a novel approach and, for the first time, built a training dataset of misogynistic hate speech using the Urban Dictionary [15] as its pre-trained data. The team also contributed this dataset to the general public. The Urban Dictionary could be a source of algospeak definitions as this is a crowd-sourced dictionary.

A third strategy would be to access data through the Reddit API. Data could be requested in batches, buffered, processed into more compact and numeric forms, saved, and/or used to train a model. Data could be labeled or left unlabeled.

A fourth dataset compilation strategy would be to access one or more labeled data sets like "A Benchmark Dataset for Learning to Intervene in Online Hate Speech" [23].

At this time, our team is still determining the best approach to compiling datasets. Our team will resolve this issue within the next week after submission of this proposal.

C. *Related Work*

Related work includes various studies that look at aspects of deep learning models. As noted below, one study only found seven papers on the topic of hate speech and deep learning in 2023. This is a difficult field to study given the constantly shifting definitions of hate speech, particularly in the United States, where there is no nationally defined governance for online hate speech. Other related work is included below.

Gambäck and Sikdar [16] used two pre-trained word embedding techniques, Word2vec by Google and GloVe by Stanford to preprocess text. The team had success in hate speech detection with CNN models. Another study noted that, "from a computer science point of view, the scientific study of hate speech is comparatively a new topic, for which the number of review papers in the field is limited. We found only a few survey or review articles during the process of literature review" [10] —in the 2023 study, seven were found.

Another study used one of the methods we propose (Python Reddit API Wrapper (PRAW)) to scrape comments related to five posts surrounding policy changes at Reddit in 2015, 2018, and 2020. The authors wanted to determine if the "attitude towards hate speech moderation" had evolved and "what should and should not be allowed" [22] Because of the aforementioned constraints with the lack of code provided by academics, we also searched the internet for potential models and their performance on specific datasets as references for our own project.

D. *Intended Experiments*

One issue that stymies attempts at consistent modeling with regard to hate speech is the lack of a common definition. Hate speech can be defined, by various entities, as:

- X(formerly Twitter): You may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease [24].
- YouTube: Hate speech is not allowed on YouTube. We don't allow content that promotes violence or hatred against individuals or groups based on any of the following attributes, which indicate a protected group status under YouTube's policy:
  - Age, Caste, Disability, Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration Status,
  - Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran Status [25]
- Meta(includes Facebook,Threads, and Instagram]: We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That is why we don't allow hate speech on Facebook, Instagram, or Threads. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence.
  - We define hate speech as direct attacks against people —rather than concepts or institutions— on the basis of what we call protected characteristics (PCs): race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease.
  - Additionally, we consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants, and asylum seekers from the most severe attacks, though we do allow commentary on and criticism of immigration policies.
  - We define a hate speech attack as dehumanizing speech; statements of inferiority, expressions of contempt or disgust; cursing; and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence. We also prohibit the usage of slurs that are used to attack people on the basis of their protected characteristics [26].
- British Columbia (B.C.), Canada: Both Canada's Criminal Code and B.C.'s Human Rights Code describe hate speech as having three main parts:

- It is expressed in a public way or place
- It targets a person or group of people with a protected characteristic such as race, religion or sexual orientation
- It uses extreme language to express hatred towards that person or group of people because of their protected characteristic [27]

Given the various definitions of hate speech, we decided to use a definition by Castaño-Pulgarín, Suárez-Betancur, Vega, and López [28]: “the use of violent, aggressive or offensive language, focused on a specific group of people who share a common property, which can be religion, race, gender, or sex or political affiliation through the use of Internet and Social Networks”.

Our plan is to choose 2-3 pre-trained models, analyze their performance on identified datasets, and possibly improve them. We then plan to test and possibly improve these models with different architectures, such as BERT or CNN, some of which will be pre-trained on data before and after 2020. Ultimately, we plan to measure and possibly enhance different models' performance before and after the policy change because we believe that hate speech may have adapted to avoid new policies. As a result, models created and trained prior to 2020 may not pick up on hate speech as effectively when used on post-2020 data.

References

[1] ”The Musk bump: quantifying the rise in hate speech under Elon Musk.” Center for Countering Digital Hate. https://counterhate.com/blog/the-musk-bump-quantifying-the-rise-in-hate-speech-under-elon-musk/ (accessed March 7, 2024).

[2] G. Schmid, P. Koehler, and N. Koch. ”Digital Services Act (DSA). What digital intermediaries need to know”, TaylorWessing Webinar, February 20, 2024.https://www.taylorwessing.com/en/insights-and-events/insights/2024/02/dsa-webinar (accessed March 8, 2024).

[3] F. Nascimento, G. Cavalcanti, and M. Da Costa-Abrue. “Exploring automatic hate speech detection on social media: a focus on content-based analysis.”, vol 13, iss 2, 2023. https://doi.org/10.1177/21582440231181311 (accessed March 8, 2024).

[4] D. Lee. “Twitter's hate speech rules are expanded.” BBC. https://www.bbc.com/news/technology-42376546 (accessed March 7, 2024).

[5] “Reddit content policy.” Reddit. https://www.redditinc.com/policies/content-policy (accessed March 5, 2024).

[6] E. Steen, K. Yurechko, and D. Klug. ”You can (not) say what you want: using algospeak to contest and evade algorithmic content moderation on TikTok. Social Media + Society, vol 9, iss 3, 2023. https://doi.org/10.1177/20563051231194586 (accessed March7, 2024).

[7] A. Greenberg. ”Inside Google's internet justice league and its AI-powered war on trolls.”, Wired, September 19, 2016. https://www.wired.com/2016/09/inside-googles-internet-justice-league-ai-powered-war-trolls/ (accessed March 8, 2024).

[8] Lycanroc. ”Operation Google.”, 2016. https://knowyourmeme.com/memes/events/operation-google (accessed March 8, 2024).

[9] aviewfromoutside. ”4chan and /pol/ are launching ”Operation Google.”, 2016. https://www.reddit.com/r/technology/comments/543a1q/ (accessed March 8, 2024).

[10] M. Jahan and M. Oussalah. ”A systematic review of hate speech automatic detection using natural language processing.” in Neurocomputing, vol 546, 2023. https://doi.org/10.1016/j.neucom.2023.126232 (accessed March 7, 2024).

[11] H. Chen, S. McKeever, and S.J. Delany. ”A comparison of classical versus deep learning techniques for abusive content detection on social media sites.” In: S. Staab, O. Koltsova, and D. Ignatov (eds) Social Informatics. SocInfo 2018. Lecture Notes in Computer Science(), vol 11185. https://link.springer.com/book/10.1007/978-3-030-01129-1 (accessed March 7, 2024).

[12] A. Marshan, F.N.M Nizar, A. Ioannou, and K. Spanaki. ”Comparing machine learning and deep learning techniques for text analytics: detecting the severity of hate comments online.” Information System Frontiers. 2023. https://doi.org/10.1007/s10796-023-10446-x (accessed March 7, 2024).

[13] A. Siagh, F.Z Laallam, O. Kazar, H. Salem, and M.E. Benglia. ”IDA: an imbalanced data augmentation for text classification. [Intelligent Systems and Pattern Recognition]. Communications in Computer and Information Science, vol 1940, 2023. https://doi.org/10.1007/978-3-031-46335-8_19 (accessed March 8, 2024).

[14] L. Couch. DS5001 University of Virginia Spring 2024 Course Notes.

[15] T. Lynn, P. T. Endo, P. Rosati, I. Silva, G. L. Santos, and D. Ging, ”A comparison of machine learning approaches for detecting misogynistic speech in urban dictionary.” 2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment, pp. 1-8, 2019. https://doi.org/10.1109/CyberSA.2019.8899669 (accessed March 7, 2024).

[16] B. Gambäck, U.K. Sikdar. ”Using convolutional neural networks to classify hate-speech.”, Association for Computational Linguistic, pp. 85-90, 2017. https://doi.org/10.18653/v1/w17-3013 (accessed March 2, 2024).

[17] M. Subramanian, V.E. Sathiskumar, G. Deepalakshmi, J. Cho, and G. Manikandan. ”A survey on hate speech detection and sentiment analysis using machine learning and deep learning models.” Alexandria Engineering Journal, vol 80, pp.110-121, 2023. https://doi.org/10.1016/j.aej.2023.08.038 (accessed March 7, 2024).

[18] V. Mathur. ”BERT And its model variants.”, May 11, 2023. https://medium.com/aimonks/bert-and-its-model-variants-162bb292611c (accessed March 8, 2024).

[19] S. Zimmerman, U. Kruschwitz, and C. Fox. ”Improving hate speech detection with deep learning ensembles.” In: N.Calzolari et al. [*Proceedings of the Eleventh International Conference on Language Resources*], May 2018. https://aclanthology.org/L18-1404/ (accessed March 7, 2024).

[20] M.A. Bashar and R. Nyak. ”CNN for hate speech and offensive content identification in Hindi language.” CEUR Workshop Proceedings. Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, volume 2517, pp. 237-245, August 2020. https://arxiv.org/abs/2008.12448 (accessed March 7, 2024).

[21] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. ”Deep learning for hate speech detection in Tweets.” [*WWW'17 Companion: Proceedings of the 26th International Conference on the World Wide Web Companion*]. pp.759-760, April 2017. https://doi.org/10.1145/3041021.3054223 (accessed March 7, 2024).

[22] E.Nakajima Wickham and E. Öhman. ”Hate speech, censorship, and freedom of speech: the changing policies of Reddit.”, Journal of Data Mining & Digital Humanities, May 30, 2022, https://doi.org/10.46298/jdmdh (accessed March 9, 2024).

[23] jing-qian. “A benchmark dataset for learning to intervene in online hate speech”. https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech (accessed March 10, 2024).

[24] ”Hateful conduct.”, X[Twitter], April 2023. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy (accessed March 8, 2024).

[25] ”Hate speech policy.”, YouTube, June 2019. https://support.google.com/youtube/answer/2801939?hl=en (accessed March 8, 2024).

[26] ”Hate speech.”, Meta, March 1, 2024. https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed March 8, 2024).

[27] ”Hate speech q&a”, British Columbia's Office of Human Rights Commissioner. https://bchumanrights.ca/hate-speech-qa/ (accessed March 9, 2024).

[28] S.A. Castaño-Pulgarín, N. Suárez-Betancur, L. Magnolia Tilano Vega and H. Mauricio Herrara López.”Internet, social media and online hate speech. Systematic review.” Aggression and Violent Behavior, vol 58, May-June 2021. https://doi.org/10.1016/j.avb.2021.101608 (accessed March 7, 2024).