Elon Musk's Purchase of Twitter and its Effect on Hate Speech Impressions

Ryan Keane

Received July 13, 2024
Accepted November 20, 2024
Electronic access November 30, 2024

Elon Musk's purchase of Twitter has stirred controversy over the platform's moderation policies and ability to effectively combat hate speech. While previous studies have explored hate speech on Twitter following Musk's purchase of the platform, a longitudinal comparison between the two eras of ownership has not yet been conducted. This study implemented a quantitative, comparative analysis to examine the change in the prevalence of hate speech across anti-LGBTQ+, anti-immigrant, xenophobic, and misogynistic speech. An existing lexicon of hate speech was used to implement a binary system of classification: a tweet would be considered hate speech if it contained one or more words present in the dictionary. Through analysis of a sample of the 50,000 most-liked tweets flagged as hate speech, it was discovered that the sample of tweets released after Twitter's change in ownership (after October 27, 2021) constituted 54% of the dataset. The remaining 46% of tweets were thus from the earlier time period; the difference between these proportions was determined to be statistically significant. These results suggest an upwards trend in hate speech on Twitter that is possibly correlated to Musk's ownership of the platform. Policies under Musk's administration therefore may have contributed to the trend, or his personal political views may have influenced the user base. Future research is encouraged to corroborate these findings by improving on this study's methodological limitations (i.e. more sophisticated hate speech detection) and to evaluate a more concrete relationship between changes in policy and politics and hate speech trends on Twitter.

INTRODUCTION

Twitter, also known as X, has had a tumultuous political history throughout its lifespan, receiving frequent condemnation from both the left and right concerning its Terms of Service, algorithm, and administrative principles. The issue worsened on October 27, 2022, when Elon Musk sparked global controversy by purchasing Twitter.

Musk stated that he intended to make sweeping changes to Twitter over the course of his tenure as owner, including reinstating accounts that were banned for violating the platform's Terms of Service and promoting his belief in "free speech absolutism." Though Musk did not clarify his personal belief in the meaning of the term, a free speech absolutist is typically a proponent of unregulated free speech regardless of its potential consequences. Supporters of Musk have argued that his ownership of the platform has ushered in a new era of productive discourse wherein individuals across the political spectrum can freely debate their ideas. Some, however, have countered that Musk's permissiveness of all forms of speech has created a climate ripe for far-right radicalism, which typically manifests in the form of hate speech. The purpose of this study is to investigate if this conjecture has been exhibited in a meaningful, quantitative trend in hate speech on Twitter.

There exists several academic definitions of hate speech, but most tend to share certain common criteria. Going forward, mentions of hate speech will refer to the definition provided by Lewandowska-Tomaszczyk et al.: "The act of humiliating individuals or groups by resorting, usually explicitly, to offensive expressions addressed at communities that are [perceived as] weaker or inferior in terms of cultural/social properties"[1]. Similarly controversial is the definition of the far-right (also referred to as right-wing extremists or the extreme right), but members of this group are typically described as possessing a belief in the superiority of one group over another along racial or ethnic lines and supporting an extreme form of nationalism either for one nation or a group of nations (oftentimes Europe)[2]. The far-right does not refer to a single political ideology but encompasses several similar frameworks, including fascism, ultranationalism, and traditionalism. Though members of groups that are typically considered institutionally or socially marginalized can hold far-right beliefs against majority groups, it is far more common for the inverse to occur. Mentions of the far-right in this paper will thus unilaterally be referring to far-right individuals whose interests align with majority groups.

Since Musk's purchase of Twitter, many changes have been made to the platform's moderation and hate speech prevention policies. Only a few months after the purchase of the platform, Musk disbanded Twitter's Trust and Safety Team, which was responsible for advising the administrators of Twitter on issues

of hate speech[3]. The platform's Terms of Service were also changed to preclude deadnaming as a form of hate speech[4]. Deadnaming is the act of referring to a transgender individual by a name other than their preferred name (typically the one they held before transitioning). Moreover, the shutdown of the Twitter API in 2023 made data collection on the site significantly more difficult and therefore led to the suspension of over 100 studies. These alterations to the platform's administration all most likely contributed to a rise in hate speech to some extent. Several confounding variables could nonetheless have contributed to any change that may be observed on Twitter. Changes in societal acceptance of hate speech may be more responsible for the change than Musk's ownership. Furthermore, higher regulations of hate speech on other platforms could have pushed more users to Twitter.

Though research into this subject has analyzed the rates of hate speech on Twitter, no studies have considered a temporal comparative analysis to determine if the platform's change in ownership affected the occurrence of tweets containing hate speech. This paper seeks to fill this gap by answering the question, "how have changes in Twitter's ownership longitudinally affected the prevalence of hate speech on the platform?" To establish causality, this study will compare the number of tweets featuring hate speech from a sample of 50,000 tweets collected over a two-year time span stretching one year before and after Musk's acquisition of Twitter on October 27, 2022. The proportion of tweets in the sample belonging to each time period will then be subject to a statistical analysis. Whether or not a tweet was made while Musk owned Twitter acts as the independent variable. Though it cannot be definitively stated that this is the only factor contributing to the relationship being studied, Musk's acquisition of Twitter is the primary difference between the two time periods and thus any notable change will be attributed to this event.

Should a noteworthy change in the occurrence of hate speech be observed, this study will have confirmed that the results of previous studies held true in a more longitudinal analysis. Furthermore, it could provide useful information to Twitter users and advertisers regarding current and future trends for the platform in terms of demographics.

Literature Review

Online Far-Right Radicalization

The far-right has been active on the Internet since its creation, but in the past decade its presence has become far more widespread[5]. Compared to the far-left, terrorism conducted by the far-right is far more common and is typically more motivated by the Internet[6,7]. Conway et al. identified 2016 as the year with the greatest uptick in far-right content online, attributing the change to the 2016 U.S. election, the Brexit movement, and the Syrian refugee crisis[8]. Berger found that the prevalence of white nationalist Twitter accounts increased sixfold between 2012 and 2016, a datapoint which Greene would attribute to the far-right's use of "new media and ironic or satiric communicative styles"[9,10]. Hagen suggests this is largely due to the Internet's "fast-paced attention economy," which causes Internet users to focus primarily on the most direct, humorous, or entertaining content[11].

Proponents of far-right ideology take advantage of this by using humor that often includes outwardly offensive stereotypes presented as jokes[12]. Many far-right posts also contain "dog whistles"—statements that seem harmless to the broader population but hold a second meaning to other members of the far-right[11]. Several examples of dog whistling are present in the lexicon of the far-right; some far-right extremists, for example, may refer to themselves as "racialist" or "alt-right" to disassociate from the mainstream view of the term "far-right"[2]. Memes themselves may also be dog whistles: one example is the "Pepe the Frog" meme which was conceived on the social media platform 4Chan to promote far-right beliefs[13]. As the meme gained more mainstream popularity, however, it became too large for Internet users to truly associate it with its far-right origins, showing a potential mechanism by which far-right dog whistles can be converted into standard speech.

Disinformation or "fake news" is another element of the extreme right's radicalization strategy[14]. Disinformation is misleading or outright incorrect information that is deliberately spread, often to serve an agenda of some kind[15]. When presented with disinformation, most people typically accept it as truth rather than investigating further to ensure its validity[16]; as such, the far-right will often use disinformation regarding public health and social issues as convincing radicalization material[17].

While much research has inspected the qualitative nature of far-right action online, quantitative investigations into the data have neglected certain factors that are necessary to understand the totality of the far-right's presence. Many investigations have focused on specific subgroups of the far-right that do not provide insight into the trend of the overall movement, such as Díez-Gutiérrez et al.who investigated Latin American far-right movements, Arcila-Calderón et al.who examined Twitter responses to the European migration crisis, or the Center for Countering Digital Hate which examined anti-LGBT hate specifically on Twitter[18,19,20]. Hate speech is intersectional in the sense that it includes disparagement towards a variety of groups . This paper will consider all of these various facets to provide a comprehensive investigation into each aspect of hate speech rather than just one category.

An understanding of hate speech on the Internet as a whole is necessary to be able to more specifically dissect how it might manifest on Twitter. Though hate speech on Twitter may not necessarily reflect the same hate speech trends as the rest of the Internet, it is more than likely to adopt certain characteristics

of it. For a more detailed analysis of hate speech specific to Twitter, a review of the platform's past and current policies on the subject is provided below.

Twitter's Management and Views on Hate Speech

Prior to 2022, Twitter's policy on far-right content was relatively strict and primarily concerned with preventing the spread of disinformation. On December 18, 2017, Twitter banned many accounts associated with the far-right, including large public figures like Richard Spencer who was one of the founders of the extremist "alt-right" movement20. Even still, various governmental organizations including the British House of Commons Home Affairs Committee insisted that Twitter take "significantly more action to remove illegal and extremist content…" after the bans[21]. Amidst the COVID-19 pandemic and 2020 U.S. presidential election, Twitter began refining its policy towards disinformation[22]. Discourse surrounding COVID-19 vaccines exploded in prominence during this time and remained a large issue on the platform thereafter[23]. While Twitter left many posts containing disinformation online, they did introduce the "Birdwatch" feature, now called Community Notes, to create a crowdsourced method of combating misinformation by allowing moderators to leave messages containing context under questionable Tweets[24].

In Fall of 2022, Elon Musk purchased Twitter and instantly made clear his intentions to rework the platform's moderation policy. He stated that he considered himself to be a "free speech absolutist," a term used to refer to someone who supports any kind of speech so long as it does not amount to harassment or otherwise threatening speech against specific individuals[25,26]. Musk also announced his intentions to unban several accounts previously banned for offensive speech[25]. The Center For Countering Digital Hate found that many of the large unbanned accounts propagated bigotry on Twitter and, in so doing, generated considerable ad revenue for Twitter[20]. They also later found that anti-LGBT rhetoric was prominent in the months following the change in ownership[20].

Musk has drifted further right throughout his tenure as CEO. In November 2023, he replied to an anti-Semitic post which stated that Jewish communities were pushing "dialectical hatred" against whites by stating "You have said the actual truth"[27]. He has also repeatedly agreed with conspiracy theories regarding the Democratic party[28]. These relatively recent developments in Musk's political stances may reflect a substantial shift in the quantity of hate speech on the platform in combination with his original purchase of Twitter. More generally, no longitudinal analysis has been conducted on Twitter's density of hate speech following Musk's purchase of the platform with the goal of correlating it to the change in ownership.

METHODOLOGY

The preliminary hypothesis for this research process was that a substantial increase in the amount of hate speech would be found following Musk's purchase of the platform. In order to best determine the accuracy of this statement, a quantitative comparative analysis was conducted on a sample of tweets. The most effective way to collect a large number of posts for analysis was through the use of a data scraper, which is simply any tool that can gather a large set of data from a website fairly quickly. Initially, tweets were to be collected using a data scraper called SNScrape, which returns a list of all tweets matching a given search query. Changes to Twitter's servers during April 2023 required that an account be created to view tweets[29], however, rendering SNScrape unusable as it accessed tweets without the use of an account. Fortunately, another project based on SNScrape titled TWScrape was released shortly after this issue arose and functions similarly, although less efficiently. TWScrape uses accounts provided by the user to view and store tweets with a similar setup to SNScrape.

For this study, ten accounts were created to use TWScrape to collect tweets from October 27, 2021 to October 27, 2023. Accounts were created using a naming convention that followed the pattern apresearch0001, apresearch0002, etc. This date range was chosen because it extends one year before and after Musk's purchase of the platform on October 27, 2022[25], providing an equal period of time both before and after. It is long enough to supply an accurate representation of the hate speech climate before and after the purchase of the platform.

Distinguishing hate speech from other forms of offensive language is a complicated subject considering no definite boundaries existing between different forms of speech[21]. Papcunová et al. suggest a relatively elaborate system of classification is necessary in order to properly determine the extent to which a statement can be classified as hate speech, which is reflected in many articles' use of hate speech detection algorithms rather than manual determinations[30]. Alkomah & Ma outline various types of hate speech detection methods, the most prominent types being lexicon-based models that refer to a set of key terms associated with hate speech and machine learning models that use neural networks to artificially create hate speech indicators[31]. One fact of many hate speech detection models is that they use a binary classification, considering a sample to be either hate speech or neutral[32]. Though this may be considered a weakness in some cases, it provides a delimitation to allow for concrete, quantitative analysis.

Though machine learning hate speech detection methods were considered, this study ultimately implemented a lexicon-based model due to its simplicity and generally accurate performance[21]. This approach involves referring to a pre-made lexicon of offensive terminology and filtering posts to only return those containing terms present in the lexicon. Bassignana

et al. provided four lexicons that were used in this study: a lexicon of anti-immigrant speech, misogyny, and xenophobia, and an additional lexicon of general insults[33]. Any terms included in the insults list that were also present in the more specific hate speech lexicons were removed from the dataset because they could include derogatory language that is not necessarily targeted against social minorities.

Each of these datasets were then transformed into the format of a Twitter search query by inserting "OR" between each term and surrounding each term in quotation marks. Because the Twitter search feature has a limit of 512 characters, the "misogyny" and "xenophobia" queries initially returned no results when entered into the Twitter search bar. Both were manually shortened to fit within the character constraints, which included removing some of the terms. Duplicate terms were removed first, followed by terms that were considered to be "archaic" or "dated" by the Oxford dictionary[34]. The xenophobia query needed to be shortened further after this process was applied, so each term was tested individually in the Twitter search and those that returned the fewest results were removed from the query.

Because the aforementioned study did not provide a lexicon for bigoted terms directed against the LGBTQ+ community, the search query from the Center for Countering Digital Hate's article "Toxic Twitter" was adapted to collect tweets including such language[20]. A section of the query that removed results relating to gay marriage was removed. In the original "Toxic Twitter" study, these tweets were excluded because the researchers were searching only for posts that included hate speech associating the LGBTQ+ community with pedophilia or grooming. For the purposes of this study, including results relating to gay marriage served to provide a more broader analysis of the totality of anti-LGBTQ+ speech. Any slurs in the query were used as an immediate indication of hate speech, whereas previously they would also have to be used in a tweet that contained a term associated with grooming or pedophilia.

Each term in each of the four queries was individually tested to assess how accurate the terms were in returning actual hate speech. Terms that yielded a considerable amount of Tweets that were not manually determined to be hate speech according to the definition provided earlier in this study were omitted. This mostly included double entendre, such as the term "savage" from the xenophobia query. The phrase "NOT (republican or conservative)" was added to each query to remove results that mentioned conservatives because most posts containing a reference to these groups were criticizing bigoted speech, not supporting it. Additionally, a minimum like count of ten was set for the posts collected. Because later analysis would largely be influenced by the collected tweets' like counts, it was beneficial to only consider posts with at least ten likes to produce greater variance in the number of likes for each tweet. Tweets with over ten likes typically are more likely to accumulate a greater number of likes with time as they are viewed by more people, whereas the opposite is generally true of tweets with less than ten likes. The search term "lang:en" was finally appended to the end of each query to only include results in English.

Tweets were gathered over a three-day period from March 7, 2024 to March 10, 2024. Initially, a single search over the two-year period of analysis was to be used for each category in the study. It later became apparent however that Twitter only allows a date range of up to seven days. For this reason, a set of the top 40 posts for each search query was collected for each day between October 27, 2021 and October 27, 2023. Due to inaccuracies in the TWScrape tool, a range of tweets was collected each day rather than a specific number, totalling to about 30-50 posts per day. Additionally, TWScrape inadvertently produced some duplicate tweets that were removed from the dataset.

Because of the aforementioned limitation of TWScrape, the second time period initially had 8059 more tweets than the first time period. To account for this, 8059 posts were removed at random from the dataset. After this change, there was an equal distribution of posts from before and after the change in Twitter's ownership.

RESULTS

In total, 88,572 tweets from October 27, 2021 to October 27, 2023 were collected. 44,286 of these posts originated from the first year of analysis and the remaining 44,286 came from the second year of analysis. A comparison based on rate of occurrence could not be made between the two samples while they contained an equivalent number of posts. To remedy this, the two datasets were merged into a single list of 88,572 tweets. This list was then sorted by amount of likes, and the 50,000 most liked tweets of the dataset were chosen for analysis. Through this procedure, a comparison could be conducted between the proportion of tweets from each dataset that were considered popular enough for inclusion. In other words, should the resulting 50,000 tweets contain more posts from the second time period than the first to a statistically significant degree, the hypothesis would be confirmed. This analysis thus considers both user engagement and rate of occurrence.

A program written in the Python programming language was used to collect and organize the data shown in Table 1. A two-proportion z-test was used to evaluate the statistical significance of the data shown, with each category of hate speech being expressed as a proportion out of the overall total of tweets. This provided an accurate representation of the number of tweets that were liked enough to be considered relevant for analysis.

The results from Table 2 show the calculated p-values for each of the four hate speech categories analyzed and the total sample of tweets. Note that a p-value less than or equal to $p = 0.05$ indicates a statistically significant difference between the two proportions. The only category to not display a statistically

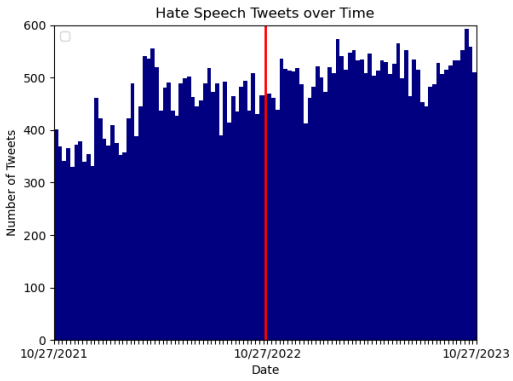

Figure 1. A histogram of the number of tweets included in the final dataset by week. A red line passes through the date October 27, 2022 when Musk purchased Twitter.

Table 1. The results from the sample of tweets analyzed. Note that "Before" and "After" refer to the two time periods of the study. "Total Included" refers to the tweets included in the analyzed dataset.

Category	Number of Tweets	Number of Tweets	Total
	Before	After	
Anti-immigrant	1521	2835	4356
Anti-LGBTQ+	4544	7387	11931
Misogyny	7941	8294	16235
Xenophobia	8871	8607	17478
Total Included	22877	27123	50000
Overall Total	44286	44286	88572

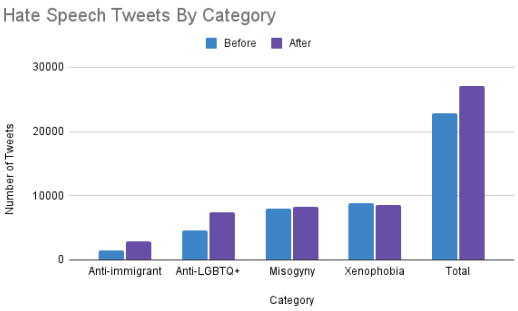

Figure 2. A bar chart representation of the information presented in Table 1.

Category	P-Value
Anti-immigrant	0.00
Anti-LGBTQ+	0.00
Misogyny	0.0010862104
Xenophobia	0.9870896316
Total Included	0.00

Table 2. The results from a two-proportion z-test comparing the proportion of tweets included in the final dataset before and after October 27, 2022. The proportions used can be derived from Table 1 by dividing each category by the corresponding value under the "Overall Total" column.

significant increase was xenophobia, which had a calculated p-value of p = 0.99, suggesting that the change in data was likely due to random occurrence. Every other category, however, including the cumulative comparison, experienced a statistically significant increase in hate speech.

Category	P-Value
Anti-immigrant	0.00
Anti-LGBTQ+	0.00
Misogyny	1.00
Xenophobia	1.00

Table 3. The results from a two-proportion z-test comparing the relative proportions of each category in the included dataset. The proportions used can be derived from Table 1 by dividing each category by the corresponding value under the "Total Included" column.

In addition to tests determining the significance of the increase in hate speech as a proportion of the total number of tweets, the change in relative proportions of each hate speech category was also tested. Table 3 indicates there was a significant increase in the proportion of hate speech tweets from the anti-immigrant and anti-LGBTQ+ categories at the expense of a completely insignificant increase in the misogyny and xenophobia categories.

	Before	After
Mean	1154.718	1374.1
Standard Deviation	12132.002	13898.5

Table 4. The mean and standard deviations of the like counts from the two time periods. The calculated p-value for the change in this data was p = 0.059.

Finally, a comparison between the means and standard deviations of the like counts of the two time periods' tweets was also conducted. The first time period was found to have a mean like count of 1,154.72 and a standard deviation of 12,132.00. The second time period had an increase to a mean of 1,374.10 and a standard deviation of 13898.52. Performing a two-sample T-test on this data yielded a p-value of p = 0.059.

DISCUSSION

The determined p-value for the general occurrence of hate speech shown in Table 2 suggests that the hypothesis proposed earlier in this study was correct. The discoveries made through this research process could not be considered a result of randomness and were almost certainly caused by some change between the two time periods of analysis.

There are several ways in which the platform's change in ownership could have attracted this increase in hate speech. Firstly, as has already been discussed, his personal views may have appealed to far-right individuals and attracted more radical users to the site. Secondly, it is possible that adjustments have been made to Twitter's algorithm to promote more content containing hate speech. This could include recommending accounts that promote hate speech to new users. It is also possible that the algorithmic benefits offered to users who purchase Twitter Blue have allowed far-right activists on Twitter to disseminate hate speech. Their tweets may otherwise have not reached a broad audience, but the payment option allows them to broadcast more efficiently. These accounts' preferential treatment could also decrease the likelihood of their accounts being suspended should they post some form of hate speech that violates Twitter's existing terms of service. Third, Musk's policy of unbanning many accounts that had been suspended from the platform prior to his ownership also likely contributed to the change, considering a large portion of this sample of users had been removed from the platform for hate speech-related offenses.

There nonetheless exists a myriad of other factors that could have contributed to the observed trend. It is possible that societal trends have simply made hate speech more acceptable, and Twitter serves as one outlet for this broader change. Other social media platforms could also have increased hate speech prevention, causing a migration of users to Twitter. In general, additional research is necessary to establish a more direct connection between Musk's ownership and the rise in hate speech.

It is more difficult to explain the changes in each individual category of hate speech. The increase in anti-immigrant and anti-LGBTQ+ speech were anticipated, especially considering the results found by the Center for Countering Digital Hate [20]. Much of far-right activism in the United States is presently focused on anti-LGBTQ+ and anti-immigrant issues, so these two categories likely saw the greatest increase for that reason. The unbanning of many far-right accounts by Musk's administration could thus be a large contributing factor to these areas of hate speech.

Though the increase in misogyny was also considered statistically significant, it was more marginally so. A potential explanation for this could be that misogynistic hate speech is less fundamental to modern far-right ideology, or that women

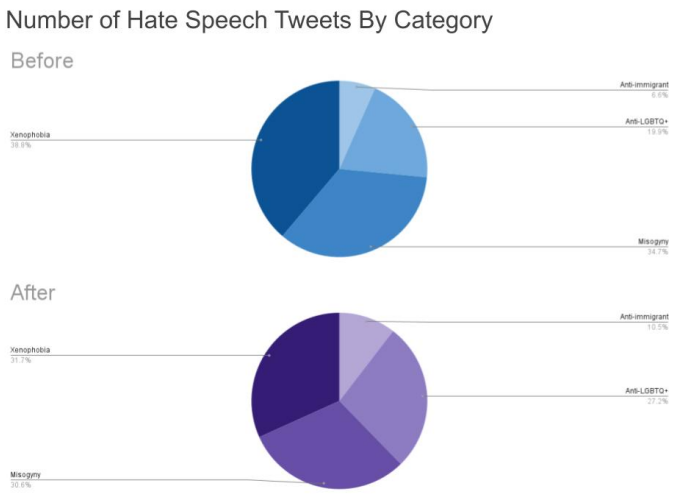

Figure 3. The relative proportions of each category of hate speech. Percentages were obtained in the same manner as in Table 3.

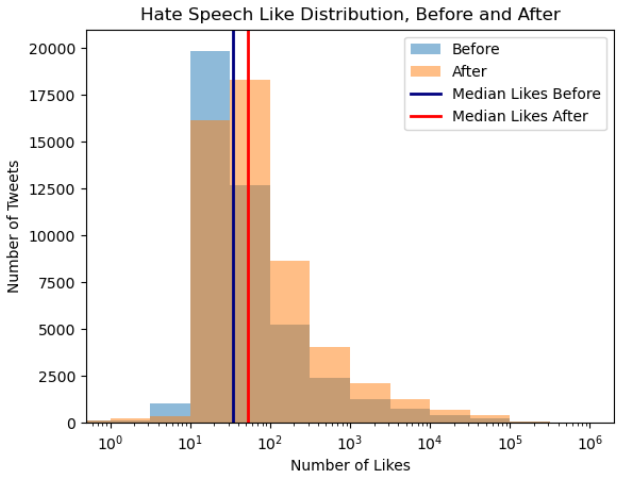

Figure 4. A histogram showing the number of likes received by tweets in the Overall Total dataset. Note that the horizontal axis uses a logarithmic scale to reflect the large variations in number of likes.

are simply not the primary target for hate speech on Twitter. The results also demonstrate a statistically insignificant change in xenophobia, implying that the change in ownership had little effect on this particular category of hate speech. This may be because much of the xenophobia-related speech on the platform contains slurs or other terms that are still considered bannable offenses under Musk's administration of Twitter. The results from Table 3, which depict a completely insignificant change in misogynistic and xenophobic speech as a relative proportion of each dataset, also suggest this. These findings would imply that the amount of xenophobia and misogyny remained relatively constant between the two time periods, but that anti-immigrant and anti-LGBTQ+ speech became more prominent.

A significant detail that further confirms the findings of this study is the p-value found from the data in Table 4 ($p = 0.059$). This p-value was found to be marginally insignificant, meaning that the change in average like count between the two time periods was likely not a result of the change in ownership. Because of this, it can be confidently concluded that the results found in Table 2 were the result of an increase in the volume of hate speech tweets and not an increase in the promotion of or engagements with hate speech tweets. This is important because the data was arranged based on its like count, so a lower p-value would indicate that there was a greater proportion of tweets from the second time period only because tweets received more likes. The number of daily Twitter users also appeared to decrease or remain the same every month since November 2022[35], suggesting that this change in the prominence of hate speech cannot be attributed to changes in the number of Twitter users.

In terms of generalizability, a medium-sized sample of Twitter posts was taken. Though it is unclear exactly how many tweets could be classified as hate speech, it is likely far greater than the sample provided in this study. Despite this, the most-liked hate speech posts were considered, meaning these were the posts that were most favored among far-right users on Twitter. The findings from this sample could thus likely be extrapolated to a larger sample of posts with less engagements. Though results could fairly be generalized to all hate speech on Twitter, the observed trends almost certainly will not be found on other platforms. Discussion around Musk's acquisition of the platform could have sparked small changes in trends on other platforms, however it is unlikely that it would have a notable nor long-lasting effect.

IMPLICATIONS

Considering the findings outlined above, Twitter users should exercise caution when engaging with the platform. Because of the aforementioned connection between hate speech and rates of far-right motivated domestic terrorism, it is possible that this increase in hate speech could act as a radicalizing agent for Twitter users. The allowance of hateful and radicalizing content on platforms other than Twitter (most notably 4chan but also including a variety of alternatives to mainstream social media sites such as Rumble) has been shown to influence several far-right domestic terrorists; it logically follows that this concerning increase in hate speech on Twitter could have a similar effect.

One notable difference between Twitter and platforms like 4chan is the size and reputation of the site. Though 4chan became a breeding ground for hate speech due to its lack of moderation, its tight-knit community allowed it to more easily radicalize specific actors to commit violence. Twitter's vastness, by comparison, could mean that it acts as a less direct influencer of domestic terrorism. Rather, Twitter is more likely to serve as an entry point to a pipeline that eventually leads to even more extreme spaces such as 4chan where complete radicalization occurs.

If the observed trend can be to any degree attributed to the increased promotion of hate speech through Twitter's algorithm, left-leaning or moderate users will have difficulty receiving the same publicity on Twitter as their far-right counterparts. The trend outlined in this study also concerns advertisers. In the view of the general public, advertising on a site that fosters an increasing amount of hate speech would suggest an endorsement of the platform's administration and general contents wary are thus expected to be wary when promoting their brand on Twitter. In fact, there exist various cases of advertisers leaving the platform throughout 2022 and 2023 in response to Musk's acquisition. Should the trend in hate speech continue, advertisers will most likely either lessen their presence on the platform or remove their advertisements entirely.

Hate speech is often associated with misinformation campaigns that aid in the promotion of pejorative stereotypes. With this in mind, the massive platforms offered to proponents of hate speech on Twitter will allow for misinformation campaigns like those during the COVID pandemic and 2020 election. Such trends were already popular prior to Musk's purchase of ownership, but the more relaxed attitude towards hate speech of the current moderation team suggests the trend could become even more pronounced during the 2024 election or other major political events.

LIMITATIONS

Although this study generally provided accurate results, some circumstantial limitations could have affected the data collection and results presented. The largest potential issue is the possibility of a non-negligible amount of false positives, i.e. posts that were flagged as hate speech during the data collection process but do not match the definition of hate speech provided in this study. Such posts may have been included if they contained quotations of others' hate speech; if they sarcastically criticized hate speech in a way that uses such speech within their post; or if the post contained a misspelled word that is identical to

a hate speech term. Furthermore, members of minority groups will often reclaim slurs and hateful language that is typically used against them for comedic or political purposes [36]. Some false positives likely emerged from this phenomenon.

Similarly, false negatives could have been a contributing factor to the collected results. A false negative is any post that meets the established definition of hate speech but was not detected by the method implemented in this study. Words omitted during the process of shortening search queries (described in the methodology section) were ignored by the algorithm but could have yielded posts containing hate speech. The use of euphemisms and dog whistling by far-right posters likely also resulted in a large quantity of false negatives.

Also important to note is that the hate speech detection methods used within this study were based on one definition of hate speech and were not all-encompassing of the types of hate speech on Twitter. Similarly to how there may have been some false positives in the dataset, it is just as likely that posts utilizing dog whistles or other forms of communication that do not contain overtly hateful language were not detected despite being hate speech.

Besides like counts, another significant indication of engagement with a tweet is its retweet count. Unfortunately, TWScrape did not provide a means to systematically access the number of retweets of each tweet collected. Thus, though retweets would have been a useful measure of engagement to combine with investigation into like counts, it was unfeasible to consider them with the use of TWScrape.

Additionally, the results for this study were limited to 50,000 posts across a two-year time period. This number was chosen because it was a reasonable amount of posts to collect during the given time frame and is a large enough sample to represent the entirety of Twitter's posts. Nevertheless, it is possible that the results collected during this two-year time period would not show the same trend as a set of tweets collected over a time period stretching further before or after Musk's purchase of Twitter. The sample size was also limited by the number of accounts used in this study; with access to a tool that creates a large number of bot accounts (which could not be purchased for use in this study due to monetary concerns) to collect tweets, a larger sample size likely could have been achieved.

FUTURE RESEARCH

Going forward, researchers could replicate the method used in this study to assess if the trend discovered continues throughout 2024 and beyond. Future analyses could also include elements of a trend analysis in addition to a comparative analysis to investigate how other spikes in hate speech may be explained by other events that occurred throughout the period of analysis.

International researchers could also consider analyzing hate speech in languages besides English. This could provide useful insight into whether or not Elon Musk was able to influence members of the far-right despite speaking a different language. To address the issue of false positives, a more sophisticated hate speech detection method that uses machine learning or a non-lexicon-based model of detection could potentially detect such posts more accurately. Creating a more complex algorithm like this could also allow the delivery method of the hate speech to be investigated; this could include classifying certain posts as using irony or humor to communicate a hateful message. Furthermore, the four categories of analysis in this study could be subdivided into more exact types of hate speech. Xenophobic speech, for instance, could be distinguished based on the particular race or ethnicity being attacked by the tweet.

Another aspect that could be considered in future research is how hateful accounts have changed in prominence and volume before and after Musk's purchase of Twitter. This could provide a window into whether or not the change in ownership attracted new accounts that promote hate speech or, instead, encouraged existing accounts to propagate such content. This could further extend into investigating the accounts liking and otherwise engaging with hateful tweets, perhaps comparing their activity before and after the change in ownership to assess if they have grown more radicalized as a result. Another element to consider is the reinstatement of many previously banned accounts after Musk purchased Twitter. Future research could examine if a significant proportion of hate speech on the platform could be attributed to these previously banned accounts.

Finally, more thorough qualitative investigation into what particular aspects of Twitter have caused the trend discovered in this study are necessary to confirm that causality can be observed. A comparison could be conducted between the various changes made by Musk, such as when the Twitter Blue subscription was introduced or when . The implications of this study must also be investigated to determine if they have manifested in any form throughout 2023 and 2024.

References

1 B. Lewandowska-Tomaszczyk, A. Baczkowska, C. Liebeskind, G. V. Oleskeviciene and S. Žitnik, *Lodz Papers in Pragmatics*, 2023, **19**, 7–48.

2 S. Jackson, *International Centre for Counter-Terrorism*, 2019.

3 M. O'Brien and B. Ortutay, *Musk's Twitter disbands its Trust and Safety advisory group*, 2022, `https://apnews.com/article/elon-musk-twitter-inc-technology-business-a9b795e8050de12319b82b5dd7118cd7`, AP News.

4 A. Yang, *Twitter quietly changes its hateful conduct policy to remove standing protections for its transgender users*, 2023, `https://www.nbcnews.com/tech/twitter-changes-hateful-conduct-policy-rcna80338`, NBC News.

5 S. F. Aziz and K. A. Beydoun, *Boston University Law Review*, 2020, **100**, 1151–1191.

6 V. A. Auger, *Perspectives on Terrorism*, 2020, **14**, 87–97.

7 J. Tinnes, *Perspectives on Terrorism*, 2020, **14**, 168–189.

8 M. Conway, R. Scrivens and L. Macnair, *International Centre for Counter-Terrorism*, 2019.

9 J. M. Berger, *Nazis vs. ISIS on Twitter: a comparative study of white nationalist and ISIS online social media networks*, 2016, https://extremism.gwu.edu/sites/g/files/zaxdzs5746/files/downloads/Nazis%20v.%20ISIS.pdf, The George Washington University Program on Extremism.

10 V. S. Greene, *Studies in American Humor*, 2019, **5**, 31–69.

11 S. Hagen and D. de Zeeuw, *Big Data Society*, 2023, **10**, year.

12 R. Labadie Tamayo, B. Chulvi and P. Rosso, *Procesamiento Del Lenguaje Natural*, 2023, **71**, 383–395.

13 A. Nagle, *The Baffler*, 2016, **30**, 64–76.

14 M. Akram, A. Nasar and A. Arshad-Ayaz, *Online Journal of Communication and Media Technologies*, 2022, **12**, 1–17.

15 D. Fallis, *Library Trends*, 2015, **63**, 401–426.

16 A. Jungherr and R. Schroeder, *Social Media + Society*, 2021, **7**, year.

17 S. Wang, F. Su, Y. Lu and J. Yuan, *International Journal of Environmental Research and Public Health*, 2022, **19**, 16849.

18 E. Díez-Gutiérrez, M. Verdeja, J. Sarrión-Andaluz, L. Buendía and J. Macías-Tovar, *Comunicar*, 2022, **30**, 97–109.

19 C. Arcila-Calderón, P. Sánchez-Holgado, C. Quintana-Moreno, J. Amores and D. Blanco-Herrero, *Comunicar*, 2022, **30**, 21–34.

20 Center for Countering Digital Hate, *Toxic Twitter: How Twitter makes millions from anti-LGBTQ+ rhetoric*, 2023, https://counterhate.com/wp-content/uploads/2023/03/Toxic-Twitter-II-Final-Report.pdf, Final Report.

21 B. Ganesh, *Journal of International Affairs*, 2018, **71**, 30–49.

22 S. Singh and M. Blase, *New America*, 2020, 39–43.

23 J. Noguera-Vivo, M. del Mar Grandío-Pérez, G. Villar-Rodríguez, A. Martín and D. Camacho, *Revista Latina De Comunicación Social*, 2023, **81**, 44–62.

24 I. Bonifacic, *Twitter's community notes feature starts rolling out globally*, 2022, https://www.engadget.com/twitter-community-notes-rolling-out-globally-195650660.html.

25 L. Hirsch, *Elon Musk completes $44 billion deal to own Twitter*, The New York Times, 2022, https://www.nytimes.com/2022/10/27/technology/elon-musk-twitter-deal-complete.html.

26 E. Heinze, *The Modern Law Review*, 2006, **69**, 543–582.

27 A. Picchi, *Elon Musk faces growing backlash over his endorsement of antisemitic X post*, CBS News, 2023, https://www.cbsnews.com/news/elon-musk-actual-truth-antisemitic-post-backlash-advertisers/.

28 K. Tangalakis-Lippert and H. Getahun, *Elon Musk Is Empowering Right-Wing Extremists on Twitter: Researcher*, Business Insider, 2022, https://www.businessinsider.com/elon-musk-right-wing-extremism-twitter-mythology-of-the-center-2022-12.

29 K. Bell, *Engadget*, 2023.

30 J. Papcunová, M. Martončik, D. Fedáková, M. Kentoš, M. Bozogáňová, I. Srba, R. Moro, M. Pikuliak, M. Šimko and M. Adamkovič, *Complex & Intelligent Systems*, 2023, **9**, 2827–2842.

31 F. Alkomah and X. Ma, *Information*, 2022, **13**, 273.

32 Z. U. Rehman, S. Abbas, M. A. Khan, G. Mustafa, H. Fayyaz, M. Hanif and A. S. Muhammad, *Computers, Materials, & Continua*, 2021, **66**, 1075–1090.

33 E. Bassignana, V. Basile and V. Patti, Proceedings of the 5th Italian Conference on Computational Linguistics, 2018, pp. 1–6.

34 O. U. Press, *Oxford English Dictionary*, n.d., https://www.oed.com/.

35 D. Ingram, *NBCNews.com*, 2024.

36 N. Ghanea, *Human Rights Quarterly*, 2013, **35**, 935–954.