Helena Caramelle
Shengyi Zhang
Georgios Saitis
Seng Moon Ja

# Evolution of Hate Speech on Twitter since Elon Musk ’s Acquisition

**Focus on hate speech towards the Black community**

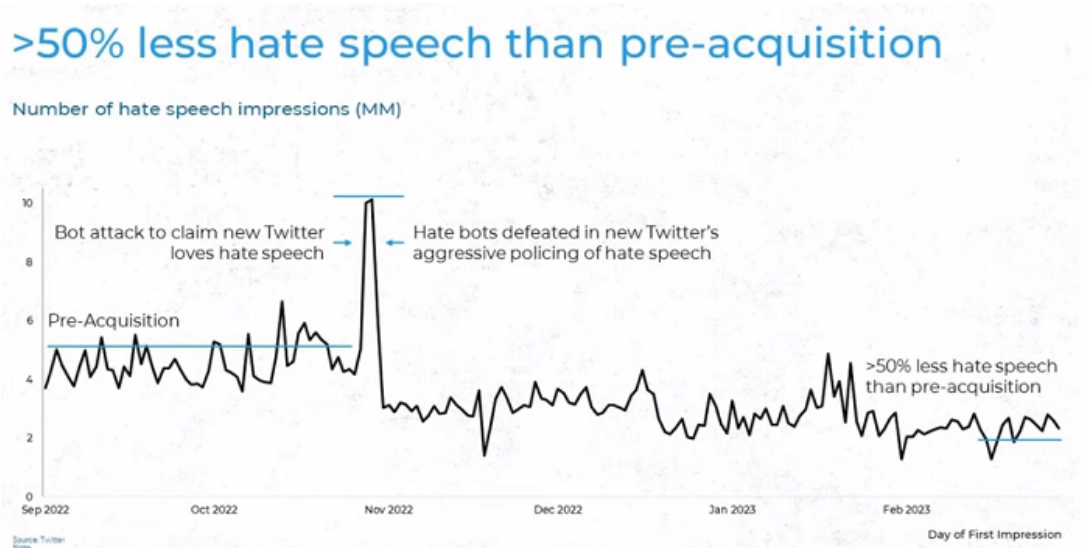
## Context and background

Elon Musk has been described as a free-speech absolutist. His chaotic takeover of Twitter has led to fear among users and politicians horrified by the prospect of Twitter amplifying the most radical online voices. Indeed, Elon Musk since his Twitter acquisition has reactivated a series of controversial accounts and has laid off a big percentage of the platform’s content moderation staff (CBS, 2022). According to a former content-moderation expert who worked for Twitter, 3,000 contract workers were fired during one night in November 2022. Apart from that, Elon Musk decided to unblock nearly 12,000 accounts that are related to racist, anti-semetic, misogynistic and transphobic comments. Some of them include Donald Trump, Andrew Anglin - a 38 year old nazi activist, and Carl Benjamin - an islamophobic theorist and well known gamer (Audureau and Leloup, 2022). At the same time, multiple studies have found that hate speech has augmented on Twitter since his acquisition with him claiming the exact opposite.

## Literature Reviews

Many researchers have identified an association between the increase of hateful content on Twitter and Elon Musk’s acquisition. For instance, a study conducted by Montclair State University collected Twitter data from Oct. 22nd to Oct. 28th, 2022 with the Tweetbinder analytics program and examined the frequency of use of multiple slurs targeting race, ethnicity and sexual orientation (Benton et al., 2022). The researchers concluded that there was an immediate spike in the number of tweets mentioning one or more of the analyzed hate terms immediately before the acquisition took place. Furthermore, using Tweetbinder’s sentiment analysis tool, the researchers found that 67.2 percent of the tweets mentioning these hate terms have a negative tone. Similarly, the Center for Countering Digital Hate suggested that the volume of tweets and retweets mentioning various slurs spiked during the first full week under Elon Musk’s ownership (Ortiz, 2022). Despite their findings, these researchers didn’t attempt to analyze the context in which the hateful terms were used. They primarily focused on the number of times that various slurs were mentioned, but tweets that contain slurs should not be classified as hate speech if they don’t express hatred towards a group or an individual based on race, sexual orientation etc. It is thus necessary to delve into hate speech detection which will

help investigate whether there is truly a rise of hate speech on Twitter associated with Elon Musk's acquisition. In addition, at the Morgan Stanley Tech, Media and Telecommunications Conference in March 2023, Elon Musk claimed that hate speech on Twitter has decreased by 50% from when he took over the platform as a result of defeating hate bots (see figure below). The contradiction of research findings and Elon Musk's statement makes it particularly interesting to examine whether hate speech on Twitter has really been on the rise since the acquisition.




## Research question

Our project aims to put the previous findings to the test. It is important to underline that we decided to revise our initial research question and make it more specific. Initially, we wanted to conduct a comprehensive investigation into hate speech towards black people on Twitter following Elon Musk's acquisition but then we decided to focus on a specific slur that best embodies racism towards black people, as we will discuss in the data section. Our new research question is: *Has the amount of hate speech that contains the racist slur "n*gger(s)" increased after Elon Musk's Twitter takeover?* In that way, we can delve into how the usage of this specific slur as hate speech has evolved in relation to Elon Musk's takeover.

## Data

Our data is based on tweets containing the n-word. We tried to collect tweets that contain this keyword written in more than 10 different ways and observe how the trend evolved. However, during the research, we found that some of them, especially the slur "n*gga'' is used positively and as an empowerment term in the black

community. This is why we decided to focus on the infamous racist slur “n*gger(s).” That helped us to quantify hate speech more effectively since it is known that it is predominately used negatively. We used the other variations of the n-word (i.e. niggar OR niggur OR nigor OR nigre OR nigar OR nigga OR niggah OR nigguh OR niglet OR nigglet OR nigra OR niggra OR nigrah OR nigruh OR niggars OR niggurs OR nigors OR nigres OR nigars OR niggas OR niggahs OR nigguhs OR niglets OR nigglets OR nigras OR niggras OR nigrahs OR nigruhs) as a proxy of the overall number of tweets. We wanted to make sure that any increase of hate speech is not due to an increase in the overall number of tweets. Apart from that, we decided to analyze a limited time span of 12 days when it comes to word count for the n-word (23/10/2022-11/04/2022). Also, our time span for the text clarification model was set to 8 days (10/23/2022-10/30/2022). The reason behind the decision for the limited time span was the difficulty of uploading a really big file to Github.

## Empirical Strategy

**Data Collection:** Data was collected using the Twitter Explorer Collector. The collected tweets were filtered “n*ggers” OR “n*gger” to include tweets related to hate speech against black people.
**Pre-processing:** Pre-processed the text data by removing stop words, punctuation, unnecessary symbols, and URLs to clean up the text data.
**Tokenization:** tokenized the text data to convert it into a format that can be processed by the machine learning models.
**Sentimental Analysis:** Sentimental analysis to get an overview of the tweets and identify whether they were positive, negative, or neutral. This step is performed just to see what was the general sentiment toward black people.
**BERTopic** using Non-negative Matrix Factorization:
We used BERTopic, a topic modeling algorithm that uses non-negative matrix factorization, to identify the topics within tweets that contain “n*gger” or “n*ggers.” BERTopic can identify the underlying topics of a text corpus based on the similarity of the text.

**Train the models to predict hate speech :**
**Annotation:** Before we trained the model, we annotated 467 Tweets manually and combined the annotated data with existing annotated data from Twitter Hatespeech from Kaggle. This Kaggle dataset has 29,530 hate speech labels, but it is labeling general hate speech Tweets. With our 467 hate speeches towards black people, we were able to fine-tune the model to specifically identify hate speech against black people. Tweets are labeled “0” if it is non-hate speech and “1” if it is hate speech.
**BERT:** BERT(Bidirectional Encoder Representations from Transformers) was applied to train a hate speech prediction model. BERT is a powerful pre-trained model that can be fine-tuned for specific NLP(Natural Language Processing) tasks. We fine-tuned BERT for hate speech prediction on our dataset.

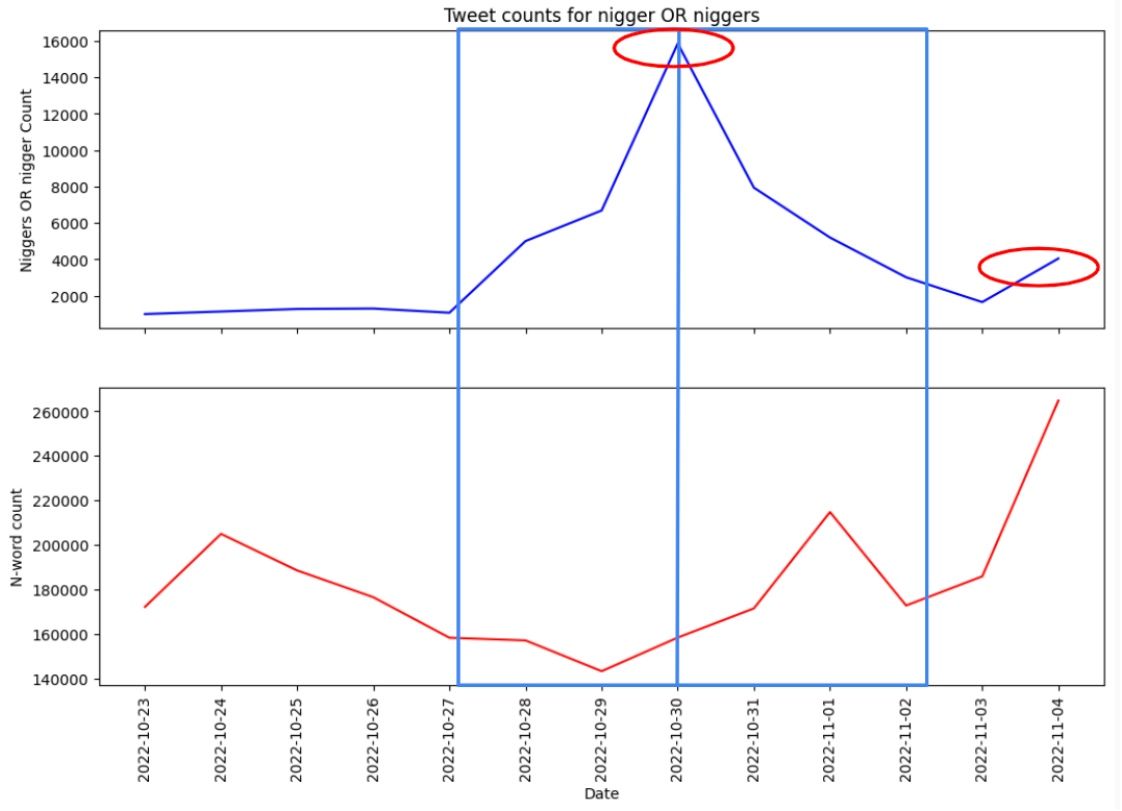
Helena Caramelle
Shengyi Zhang
Georgios Saitis
Seng Moon Ja

**Support Vector Classifier** (SVC) was used to detect hate speech against black people. SVC is a binary classification algorithm that can classify text into two categories: hate speech and non-hate speech.

## Main Findings

### Number of tweets using N-words

We first attempted to obtain the count of words “n*gger” and “n*ggers” appearing in Tweets during 2 weeks around the date of the official acquisition to see the magnitude of its use and evolution pre and post-take over of the platform by Elon Musk (c.f. blue graph).



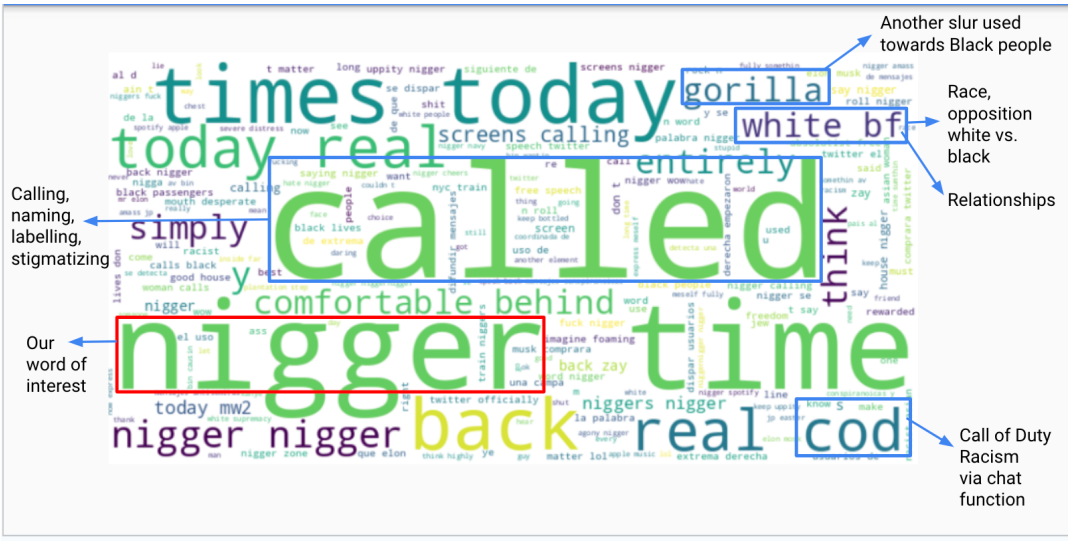

We can observe a notable increase in tweets containing the racial slur “n*gger(s)” right after the official acquisition, on October 27th. More specifically, the average number per day, which was c.1000 pre-acquisition, gradually increased to c.5000 on Oct. 28th, a little less than 7000 on Oct. 29th, before spiking at c.16,000 on Oct. 30th. The daily number of use of the racial slur then gradually decreased, until coming back on Nov. 3rd to a level only slightly higher than the average pre-acquisition. A factor explaining this increase could be that considering that Oct. 28th, 29th, 30th correspond to Friday-Saturday-Sunday, Twitter users had more free

time and were more inclined to tweet compared to weekdays. However, we can rule out this hypothesis: we argue that this notable increase in the use of the word “n*gger(s)” is not due to an increase in the overall number of tweets on the platform. Indeed, we have plotted in the red graph the count of other derivations of the n-word, such as “n*gga(s)”, “n*gro(s), to create a proxy of the overall number of tweets. We observe that this number did not increase between Oct. 27th and 30th. Thus, the notable increase in the use of the word “n*gger(s)” might be a lagged reaction to the announced looser content moderation related to the change in ownership.
We then observe a slight increase in the use of the slur “n*gger(s)” again on Nov. 4th, reaching c.4000 and thus almost doubling compared to the day before. This phenomena could be explained by a reaction to NBC News reports’ announcement on November 4th that many employees in charge of fighting against misinformation were fired by the new CEO, Elon Musk. This announcement could have caused hope for less regulation on content and thus giving more space to hate speech. However, this slight rise on Nov. 4th could also merely be proportional to the increase in the overall number of tweets seen in the bottom graph. If it is the case, the observed rise in use of the “n*gger(s)” slur might not be related to looser content moderation and thus to an increase in use of the slur in a hateful way.

**Word Cloud**

We generated a word cloud for the tweets that we collected (those containing the slur “n*gger(s)”) to identify the most common terms associated with the slur. Notable ones were: “called”, “gorilla”, “cod”, and “white bf”. The verb “called” can be interpreted as naming, labeling and thus categorizing or stigmatizing Black people in a negative or even hateful way. The word “gorilla” is another racial slur used against the Black community. “Cod” is an abbreviation for the video game Call of Duty, where the live voice chat function is used by racist players to insult Black players. Several tweets we analyzed were posted by Black players complaining that there had been called a “n*gger” while playing. Lastly, “white boyfriend” is interesting as it refers both to race and relationships: we see the opposition between white and black ethnicity, as well as an association to couples and sometimes a power relationship within.

Helena Caramelle
Shengyi Zhang
Georgios Saitis
Seng Moon Ja




## BERTopic

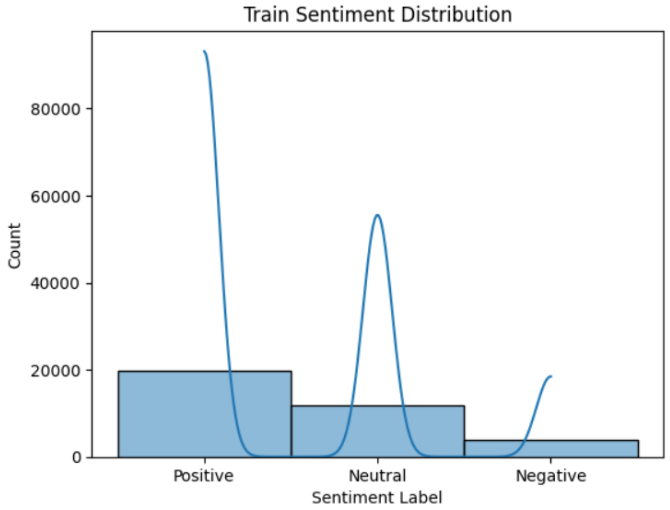
We also used topic modeling, and more specifically BERTopic, to identify the topics within tweets that contain “n*gger” or “n*ggers.” However, most of the topics that we found had random keywords or words that corresponded to Twitter accounts and they didn't give us any tangible information about the relevant topics.
However, some of the topics were very interesting. For example, topic 15 shows that the usage of the n-word is related to the decision of Elon Musk to be more “democratic’’ and allow Twitter users to express themselves freely. Similarly, topic 71 shows that people are using the n-word in tweets where they also talk about Elon Musk’s controversial decision, and topic 46 is directly related to him and his Twitter acquisition. Another topic that keeps coming up in the tweets containing the n-word is hate-related. In topic 25 the word ‘’die’’ has the largest bar and it’s closely followed by the slur “n*ggers.” Similarly, in topic 26 the word hate is being used significantly. Based on the data, it can be inferred that the most prominent subjects of tweets containing the n-word are related to Elon Musk's implementation of a free speech policy on Twitter and the expression of hate speech.

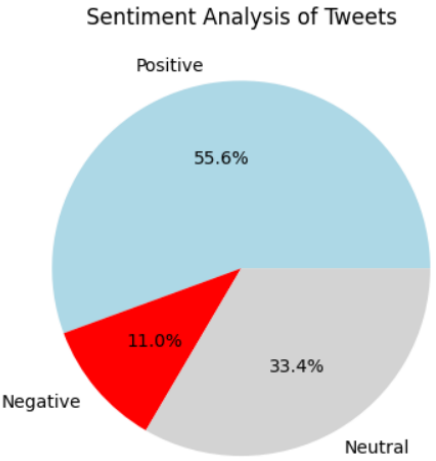
Helena Caramelle
Shengyi Zhang
Georgios Saitis
Seng Moon Ja


## Sentiment analysis

We ran a sentiment analysis to get a sense of how the slur “n*gger(s)” was used in tweets in the period of interest. Surprisingly, the majority (55.6%) of tweets had a positive sentiment, around one third (33.4%) were neutral, and only 11% were associated with negative sentiment. Contrary to our intuition, the sentiment analysis shows that tweets containing the racial slur were rarely used in an offensive way, as an insult. It seems to have been mostly employed in a friendly or neutral way, for example as a way for black people to call each other. However, it is important to note that this analysis is purely descriptive and does not take into account the context (user profile like ethnicity, place of residence, age, history of posts, retweets, when the tweet was posted and to what external event it might be correlated, etc.). In addition, this sentiment analysis is not specific to racial hate speech. Thus, we must take some distance from these results, and analyze the tweets in their context in our next steps.

Helena Caramelle
Shengyi Zhang
Georgios Saitis
Seng Moon Ja




## Train Hate Speech Models

As we understand that slurs are not always used as hate speech, we delve into hate speech detection by using the two models below.

**Support Vector Classifier**

**Model accuracy and precision**

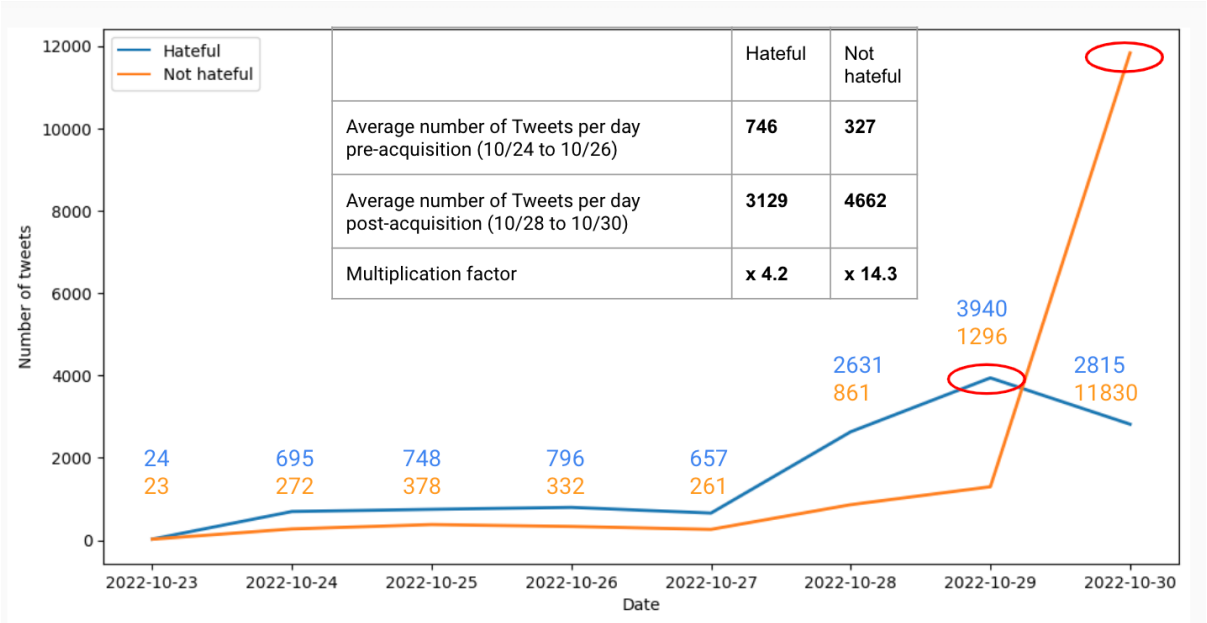
We first used the Support Vector Classifier (SVC) to detect hate speech in our data. The final model has a 84% F1 macro score and 96% accuracy.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.96 | 0.99 | 0.98 | 5945 |
| 1 | 0.85 | 0.60 | 0.70 | 543 |
| accuracy | | | 0.96 | 6488 |
| macro avg | 0.91 | 0.79 | 0.84 | 6488 |
| weighted avg | 0.95 | 0.96 | 0.95 | 6488 |

**Results**

In terms of the evolution of the number of tweets containing the slur of interest, it is very similar to the word count analysis we did in the beginning: the frequency increases after the acquisition and spikes 3 days after.

Helena Caramelle
Shengyi Zhang
Georgios Saitis
Seng Moon Ja




| | Hateful | Not hateful |
|---|---|---|
| Average number of Tweets per day pre-acquisition (10/24 to 10/26) | **746** | **327** |
| Average number of Tweets per day post-acquisition (10/28 to 10/30) | **3129** | **4662** |
| Multiplication factor | **x 4.2** | **x 14.3** |

This analysis however allows us to determine if the slur is used in a hateful way or not. During the 3 days pre-acquisition, the slur was used more frequently in a hateful way: average of 746 tweets a day compared to 327, corresponding to a factor of 2.3 During the 2 days following the acquisition, the slur continued to be more frequently used in a hateful way, in a much greater magnitude compared to the non-hateful use, and peaks on the 29th. On Oct. 30th, the trend reversed, and the majority of tweets containing the slur were used in a non-hateful way: 11830 compared to 2815, thus a factor of 4.2. Contrary to our intuition, the use of the slur in a non-hateful way increased by a greater factor (x14.3) after the acquisition on average compared to it being used in a hateful way (x4.2).

## BERT

### Model training - annotation

We then started training a BERT model by annotating some of the tweets that contain the word “n*gger” or “n*ggers” and carefully examining how they are used. We are interested in whether a tweet counts as hate speech. Accordingly, we created two categories: “1” denotes “hate speech” and “0” denotes “non-hate speech.” For the tweets that are too vague for us to determine whether they are hate speech, we didn’t include them in the training model.

### Model accuracy and precision

The model was trained for four epochs to achieve higher quality. The final model has a 90% F1 macro score and 97% accuracy which are higher than the scores of the SVC model.

Helena Caramelle
Shengyi Zhang
Georgios Saitis
Seng Moon Ja

Average test loss: 0.10
Validation took: 0:00:43

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.97 | 0.99 | 0.98 | 5963 |
| 1 | 0.90 | 0.65 | 0.76 | 525 |
| accuracy | | | 0.97 | 6488 |
| macro avg | 0.94 | 0.82 | 0.87 | 6488 |
| weighted avg | 0.96 | 0.97 | 0.96 | 6488 |

Epoch 1

Average test loss: 0.10
Validation took: 0:00:43

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 0.98 | 0.98 | 5963 |
| 1 | 0.79 | 0.81 | 0.80 | 525 |
| accuracy | | | 0.97 | 6488 |
| macro avg | 0.89 | 0.90 | 0.89 | 6488 |
| weighted avg | 0.97 | 0.97 | 0.97 | 6488 |

Epoch 2

Average test loss: 0.15
Validation took: 0:00:43

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 0.99 | 0.99 | 5963 |
| 1 | 0.90 | 0.75 | 0.82 | 525 |
| accuracy | | | 0.97 | 6488 |
| macro avg | 0.94 | 0.87 | 0.90 | 6488 |
| weighted avg | 0.97 | 0.97 | 0.97 | 6488 |

Epoch 3

Average test loss: 0.16
Validation took: 0:00:43

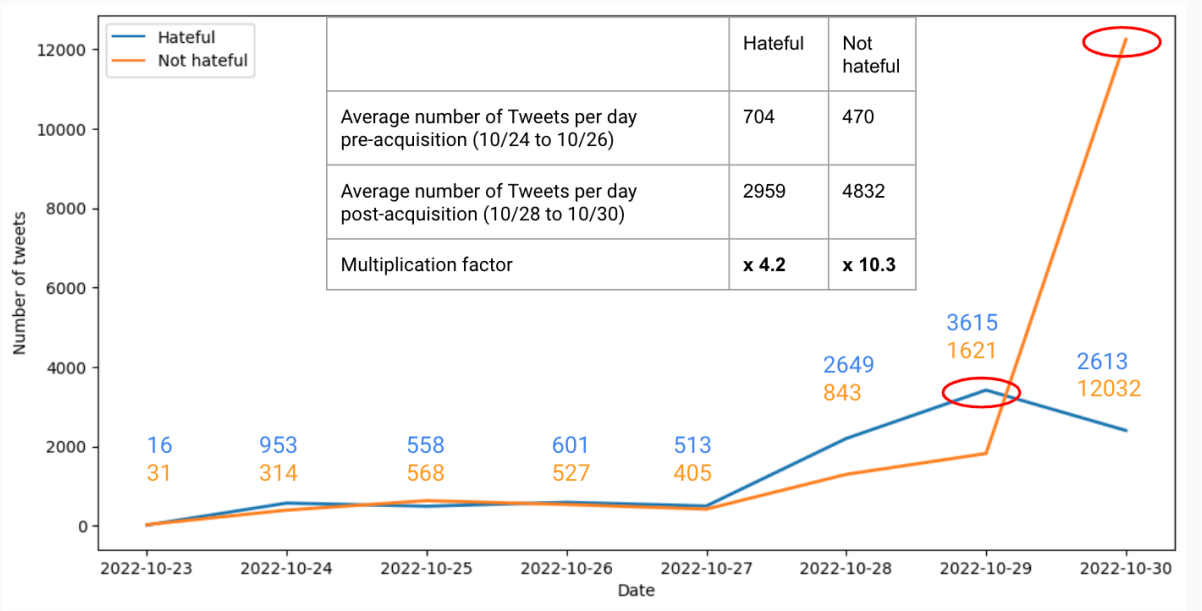
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 0.99 | 0.99 | 5963 |
| 1 | 0.89 | 0.76 | 0.82 | 525 |
| accuracy | | | 0.97 | 6488 |
| macro avg | 0.93 | 0.88 | 0.90 | 6488 |
| weighted avg | 0.97 | 0.97 | 0.97 | 6488 |

Epoch 4

*Classification table*

**Results**

The overall trend is similar to our findings with SVC. The slur is also slightly used more frequently in a hateful way pre-acquisition: average of 704 tweets a day compared to 470. The trend post-acquisition is also similar: a rise in the overall use, and considerable rise in use in a hateful way until the peak on Oct. 29th. The trend inverses the day after, with an explosion of use in a non-hateful way.




| | Hateful | Not hateful |
|---|---|---|
| Average number of Tweets per day pre-acquisition (10/24 to 10/26) | 704 | 470 |
| Average number of Tweets per day post-acquisition (10/28 to 10/30) | 2959 | 4832 |
| Multiplication factor | **x 4.2** | **x 10.3** |

Compared to the SVC model, our trained model shows a more moderate increase in non-hateful use post-acquisition, despite being significant: we found an increase by a factor of 10.3, compared to 14.3 previously. The change in factor for hateful use pre and post-acquisition is the same for both models: a factor of 4.2.

Helena Caramelle
Shengyi Zhang
Georgios Saitis
Seng Moon Ja

## Conclusion

Our research question was: *Has the amount of hate speech that contains the racist slur "n*gger(s)" increased after Elon Musk's Twitter takeover* (10/27/2022)? Our hypothesis was that yes, it has increased after the acquisition.

Our results show that based on the word count, the frequency of use of "n*gger(s)" has increased right after the acquisition and spiked 3 days later, and gradually decreased but remained higher than pre-acquisition. Based on our trained model, use in a hateful way increased moderately right after the acquisition, spiked 2 days later, and then slightly decreased. On the other hand, use in a non-hateful way increased at a lesser extent during those 2 days, but experienced a surge on October 30th. Thus, we can validate that overall there was an increase in the frequency of use of "n*gger(s)". However, we cannot conclude that there is a massive increase in the n-word being used as hate speech after the acquisition. Additionally, it's important to note that the relationships we observed here are only correlational. We cannot conclude that Elon Musk's reinstatement of banned accounts and layoff of content moderation staff led to the increased frequency of "n*gger(s)" and more hate speech.

Lastly, we realized through our analysis that merely counting how many times the slurs are used, as done in several recent scientific and academic research on this topic, is not sufficient for hate speech detection. The results of our trained model show that it is crucial to understand in which context the n-word is being used, as it is not always used in a hateful manner. The method of quantifying hate speech employed by the Twitter team is unclear, but this methodological difference in hate speech detection may be a cause of the contradiction between the academic researchers and Elon Musk.

## Limitations

Due to the time constraints, we only investigated the use of n-word on Twitter. Future studies should broaden the scope of research by examining other slurs that target the Black community as well as other minority groups. Also, we initially hoped to retrieve the total number of tweets posted each day to make sure that the increase of hate speech is not due to an increase in the overall number of tweets. However, Twitter doesn't allow us to retrieve these numbers. Moreover, we only annotated tweets written in English, as we wanted to be sure that we understood the content of the tweets. However, the final collection of all the tweets that mentions "n*gger(s)" contains multiple languages, as Twitter explorer collector doesn't have a language filter. It was also difficult to differentiate between hate speech and non-hate speech in

the annotation process without enough context and account information, which led to many blanks and possible interpretation errors. Initially, we intended to examine the amount of hate speech one month prior to and after the acquisition. However, because of the enormous amount of tweets and the high level of difficulty of analyzing them, we had to narrow the timespan to 8 days. The free GPU usage limit also reduced our capacity to train the BERT model.

## Bibliography

Audureau, W., Leloup D. (2022) *Conspiracy theorists, homophobes, neo-Nazis: Ten accounts that embody Twitter's change under Musk,* Le Monde, Available at: https://www.lemonde.fr/en/les-decodeurs/article/2022/12/19/conspiracy-theorists-homophobes-neo-nazis-ten-accounts-that-embody-twitter-s-change-under-elon-musk_6008352_8.html

Benton, B., Choi, J., Luo, Y., and Green, K. (2022). *Hate Speech Spikes on Twitter After Elon Musk Acquires the Platform*. School of Communication and Media Scholarship and Creative Works. 33.
Available at:
https://digitalcommons.montclair.edu/scom-facpubs/33?utm_source=digitalcommons.montclair.edu%2Fscom-facpubs%2F33&utm_medium=PDF&utm_campaign=PDFCoverPages

CBSnews (2022). *Musk fires outsourced content moderators who track abuse on Twitter,* Available at:
https://www.cbsnews.com/news/elon-musk-twitter-layoffs-outsourced-content-moderators/

Hutchinson, A. (2023). *Elon Musk Outlines His Vision for Twitter at Morgan Stanley Tech Conference*. [online] Social Media Today. Available at: https://www.socialmediatoday.com/news/Musk-Outlines-Vision-for-Twitter-at-Morgan-Stanley-Conference/644417/.

Ortiz, A.G. (2022). *Musk's claim about a fall in hate speech doesn't stand up to scrutiny*. [online] Center for Countering Digital Hate | CCDH. Available at: https://counterhate.com/blog/fact-check-musks-claim-about-a-fall-in-hate-speech-doesnt-stand-up-to-scrutiny/