Toxicity Analysis of Elon Musk's Tweets Using the Perspective API

Simon Kruelle
Eberhard Karl University of Tuebingen
December 22, 2022

Abstract

This study analyzes the toxicity of tweets made by entrepreneur and business magnate Elon Musk using the Perspective API, a machine learning tool developed by Google that can identify patterns of toxic or abusive language. The API is used to classify Musk's tweets based on various toxic attributes, including toxicity, severe toxicity, identity attack, insult, profanity, and threat. The results are depicted in a visual representation to observe any changes over time. The resulting dataset, which includes Musk's tweets and their corresponding toxicity scores, can be used for further analysis, such as comparing the overall toxicity of Musk to that of other public figures. The study found that Musk's earliest tweets had the highest toxicity scores, with two notable peaks in 2012-2013 and 2019. There was also a slight decline in toxicity over time, although recent tweets from July-October 2022 suggested a potential increase in the future. Elon Musk's overall toxicity was relatively low, with an average of 5.3% across the entire dataset. The wordcloud analysis of the 48 most toxic tweets revealed that they often contain derogatory and offensive language and accusations of others being irrational or unstable. This study contributes to the understanding of toxicity patterns of social figures on social media platforms and the potential triggers or causes of toxicity in online behavior.

Keywords: Elon Musk, Perspective API, toxicity, Twitter

Introduction

Over the past few months, the public discourse around entrepreneur and business magnate Elon Musk and his companies has become increasingly heated and divisive. His recent acquisition of the social media platform Twitter has raised questions about his motivations and the potential impact on Twitter's future direction. While Elon Musk has been a vocal advocate for free speech, even calling himself a "free speech absolutist", his own use of social media has been criticized for inciting hateful and harmful discussions and for spreading misinformation [1]. According to The Atlantic [2], Elon Musk's commitment to free speech has been questionable, as he seems to prioritize his own ability to express himself over others' freedom of expression. Furthermore, he has been known to dismiss or even retaliate against those who disagree with him, both on Twitter and within his own companies. Upon acquiring Twitter, Elon Musk also emphasized his commitment to promoting free speech on the social media platform. For him the idea of free speech means not taking down anything and therefore allowing every opinion [1]. However this definition of free speech could lead to ostracizing minorities from expressing their opinion and increase the overall toxicity on the platform. Reactions to the buyout have therefore been divided, with some praising Musk's vision for the company and his support for free expression, but others expressing concern about the potential for an increase in misinformation, disinformation, harassment, and hate speech.

Consequently tools for moderating toxicity are gaining more and more significance because they help to create a safer and more inclusive online environment for users by identifying and removing harmful or abusive content. One such tool is the Perspective API [3], a machine learning tool developed by Google, that can analyze text and identify patterns of toxic or abusive language. The API uses a combination of natural language processing and machine learning techniques to understand the context and sentiment of words and phrases, and it can be customized to identify specific types of toxic language, such as hate speech, bullying, or harassment. By using the Perspective API, social media companies can proactively identify and remove harmful content, helping to create a safer and more inclusive online environment for their users.

The purpose of this study is to construct a dataset [1] of all tweets made by Elon Musk and use the Perspective API to classify them based on various toxic attributes, including toxicity, severe toxicity, identity attack, insult, profanity, and threat. The toxicity levels will be examined and depicted in a visual representation to observe any changes over time. The resulting dataset, which includes Musk's tweets and their corresponding toxicity scores, can be used for further analysis, such as comparing the overall toxicity of Musk to that of other public figures.

To conduct this analysis, I first describe the Perspective API and its capabilities as a tool for analyzing online text. I then outline the steps I took to collect and classify the data, including details of the data collection process and the specific methods used to classify the tweets using the Perspective API. Finally, I present my findings and discuss the implications of my results. I also acknowledge the limitations of my approach and suggest directions for further research.

[1] The resulting dataset "Elon_Musk_Twitter_Toxicity.csv" is available on the projects GitHub-repository at github.com/simonkruelle/Toxicity_Analysis_Elon_Musk

Method

The Perspective API is a machine learning tool developed by Google's Jigsaw team that is designed to analyze the toxicity of online text. It uses natural language processing (NLP) techniques to evaluate the likelihood that a given piece of text will be perceived as "toxic" by a human reader. Toxicity is defined by the API as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion" [4].

The language model is trained on a large dataset of comments in different languages from online forums such as Wikipedia and The New York Times. The training data is labeled by human raters for various toxicity attributes. The resulting score, between 0 and 1, gives the likelihood that a reader would perceive the comment as fitting the respective attribute. The architecture of the Perspective API is based on multilingual Bidirectional Encoder models (BERTs), which are distilled into single-language Convolutional Neural Networks (CNNs) for each supported language [4].

Developers can use the API by making HTTP requests to the server with the text they wish to analyze. The API will then return a JSON object containing the toxicity scores for the text, as well as scores for other attributes such as identity attack, insult, and threat. The API can be utilized to moderate online conversations by providing comments with toxicity scores to enable human moderators to sort comments more efficiently, as is already being done in the New York Times comment sections. Another use case is to provide feedback on the perceived toxicity of comments that are not yet posted to give commenters an oppertunity to change their wording. This tool named Coral by Vox Media uses the API to prophylacticly prevent toxic comments from beeing posted on more than 200 sites [4].

In the folowing analysis, the API is used to score the toxicity of tweets from Elon Musk. To collect a dataset of Twitter tweets, the Twitter API can be used to search and collect tweets from specific persons or about specific topics. Unfortunately, I was unable to work with this utility, because an "Academic Researcher" account is needed to access the full archive of tweets. With the free "Essential" or "Elevated" account I was limited to recent tweets, posted a maximum of seven days ago. Therefore I chose to combine a dataset of 17 thousand Elon Musk tweets [5] with a more recent dataset of 2542 tweets [6]. The resulting dataframe of Elon Musk tweets includes all of his tweets from June 4, 2010 to October 27, 2022. The raw tweet data was preprocessed and transformed into clean text by deleting emoticons, hashtags and other characters that cannot be parsed by the Perspective API. All tweets shorter than 10 characters were excluded from the analysis, in order to reduce the overall variance, since the API tends to be unreliable for single words or short sentences. The cleaned tweets were then classified, using the Perspective API's "analyze" method, by the attributes of toxicity, severe toxicity, identity attack, insult, profanity, and threat. The attribute scores obtained were added as new columns to the dataframe for further analysis [2].

[2] The source code is available on the projects GitHub-repository at github.com/simonkruelle/Toxicity_Analysis_Elon_Musk

Results

As previously suggested by Communalytic [7], the six toxicity attributes calculated by the Perspective API were analyzed for their correlation with one another. To determine the extent to which the toxicity scores are correlated with each other, a correlation matrix can be used. For each pair of toxicity scores, the matrix shows a correlation value ranging from 0 to 1, where 0 indicates no correlation and 1 indicates a high level of correlation. The values are represented in the matrix in shades of blue, with lighter shades indicating lower correlation values and darker shades indicating higher correlation values.

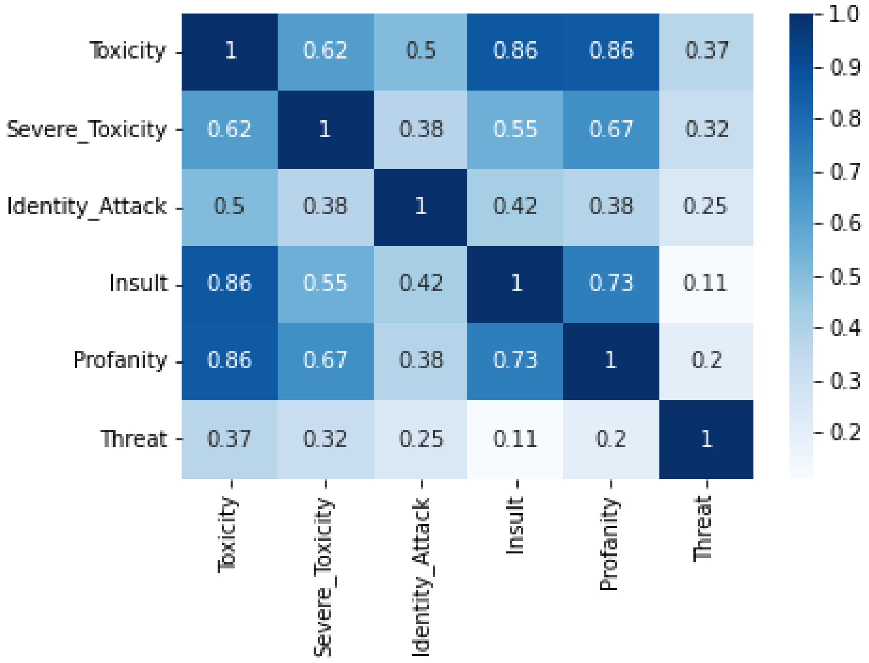

Figure 1: Correlation across toxicity attributes.

As shown in Figure 1, the "toxicity," "insult," and "profanity" attributes are highly correlated with each other (correlation values of 0.7 or higher). The "severe toxicity," "identity attack," and "threat" attributes exhibit moderate correlation with all other attributes, with correlation values ranging from 0.35 to 0.7. The "threat" attribute has only weak to moderate correlation with the other toxicity attributes, with correlation values below 0.35. This means that, in general, high values in any of these toxicity attributes will correspond to high values in the other toxicity attributes, with the exception of the "threat" attribute, which captures a somewhat different aspect of Elon Musk tweets. For the purposes of our exploratory analysis, I will only consider the "toxicity" attribute, as it can serve as a proxy for the majority of the toxicity attributes provided by the Perspective API.

To visualize changes of toxicity over time, the toxicity scores from all tweets for each day were averaged in order to obtain one score value for each day.

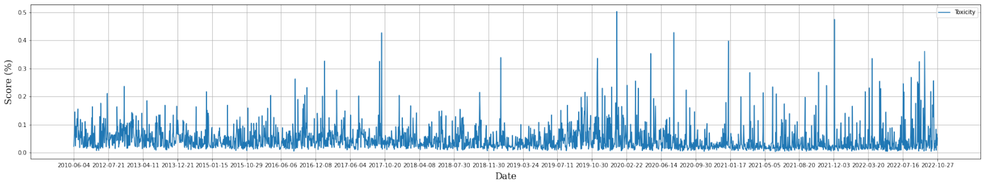

Figure 2: Timeline of the average toxicity scores.

The resulting curve has a lot of noise, so a localized regression was used to smooth the curve to better visualize the underlying trend. It works by fitting multiple regressions in the local neighborhood of each point specified by the percentage of data points nearby that should be considered to fit the regression model. A value of 5% was used.

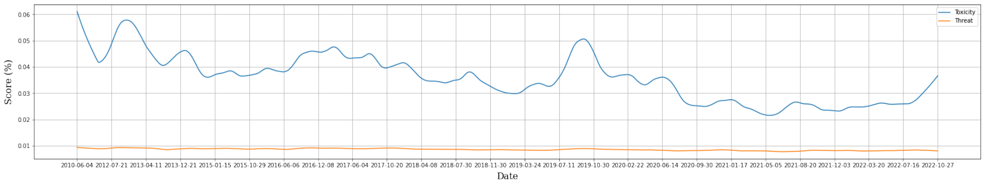

Figure 3: Smoothed timeline of the average toxicity scores.

In Figure 3 the smoothed curve for the "threat" category was also plotted, but it approached zero and was therefore not significant and was not considered in further analysis. Contrary to what may have been anticipated, Elon Musk's earliest tweets had the highest toxicity scores, with a maximum of approximately 6% on June 4, 2010. Two notable peaks in toxicity were observed, one between July 21, 2012 and March 11, 2013 and one between July 11, 2019 and October 30, 2019. While there has been a slight decrease in overall toxicity over time, tweets from July 16, 2022 to October 27, 2022 suggest a potential increase in the future. Future research could investigate these peaks further, for example through wordcloud analysis, to identify any patterns or trends within this time.
By plotting the maximum toxicity score for each day along with a smoothed curve, the resulting curve appears nearly identical to the curve obtained by averaging toxicity scores. The main difference is that the scores are more extreme, with a maximum of 18%. This representation also more clearly demonstrates the decrease in toxicity over time.

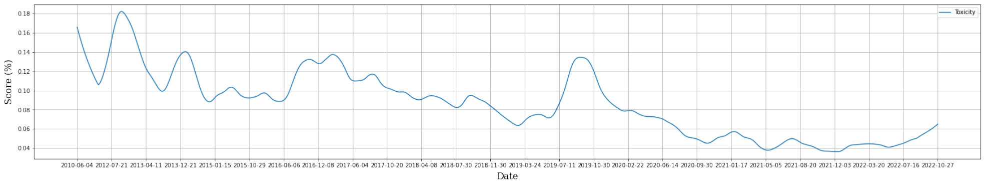

Figure 4: Smoothed timeline of the maximum toxicity scores.

It should be noted, that Elon Musk's overall toxicity was relatively low, with an average of 5.3% across the entire dataset. The Perspective developer portal [4] recommends different thresholds for determining whether a text should be considered toxic for different use cases. For example, social science researchers studying harassment should use thresholds between 0.7 and 0.9, while machine learning researchers filtering out toxic content from their data should use even higher thresholds like 0.9 or 0.95. Filtering the dataset using a threshold of 0.9 resulted in only three tweets exceeding this threshold and being considered toxic. A threshold of 0.7 resulted in 25 tweets being considered toxic. For the sake of effectively visualizing the data, a lower threshold of 0.6 was selected in order to include a sufficient number of tweets (48).

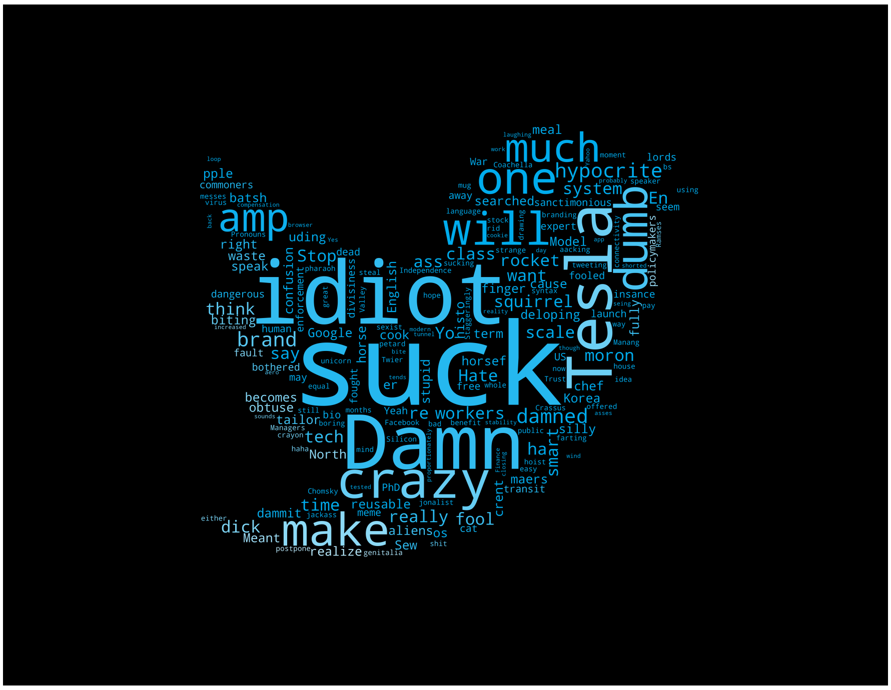

Figure 5: Wordcloud of the 48 most toxic tweets.

The resulting wordcloud visualizes the most common words used by Elon Musk in his 48 most toxic tweets, with the size of each word proportional to its frequency in the tweets. Upon analyzing the wordcloud, it is clear that certain themes emerge. The most frequent words used in the most toxic tweets include derogatory and offensive language, such as 'idiot', 'suck', and 'damn'. These words suggest that Musk's toxic tweets often involve insults or derogatory language. The presence of the word 'Tesla' suggests that Musk's toxic tweets may often involve discussions of his company. The word 'crazy' could indicate that Musk's toxic tweets may often include accusations or portrayals of others as irrational or unstable. Overall, the wordcloud provides insight into the language and themes present in Musk's most toxic tweets. By understanding the common words and themes in these tweets, it may be possible to identify potential triggers or causes of toxicity in Musk's online behavior. Further analysis of the language and themes present in the rest of Musk's tweets could provide a more comprehensive understanding of his overall Twitter use and help to identify any patterns or trends in his toxic language.

Limitations and future directions

While the Perspective API is a useful tool for analyzing the toxicity of online text, it has limitations that should be taken into account when interpreting the results. One key limitation is the potential for bias in the API's algorithms, which may affect the accuracy of the toxicity scores generated. The API is trained on a dataset of text that may not be representative of all forms of online discourse, and may therefore be prone to bias or error when applied to other contexts. Hosseini et al. (2017) demonstrates that the API can be deceived by strategically manipulating the wording and structure of a comment, leading to incorrect toxicity scores. This finding highlights the potential for bias in the API's algorithms, and suggests that the results of its analysis may not always be accurate or reliable.

Since the API is unable to automatically detect the written language, it was not possible to filter out non-english tweets in my analysis. All Elon Musk tweets were therefore processed by the API as "english," even though the API is capable of handling other languages. This also leads to unreliable toxicity scores for tweets in different languages, altough the percentage of tweets in different languages was marginal.

Another limitation is the difficulty of accurately analyzing the toxicity of entire Twitter discourses, which often include replies and mentions that can not be captured in the analysis. These interactions can significantly impact the overall tone and content of a conversation, and can not be accounted for by the API's algorithms, when only the raw text of a tweet is provided. In an effort to address this issue, the Perspective API will soon be adding the feature of using context to more accurately classify toxicity, as stated on the developer website.

One final limitation of the Perspective APi is the quota limit of one "analyze" request per second. As a result, the toxicity classification process is relatively slow for large datasets, taking approximately four hours for all of the tweets by Elon Musk in this study. This hindered the ability to perform a comparison between different individuals within the time constraints of the study. In order to facilitate more efficient and comprehensive analysis in future studies, it is recommended to apply for a higher quota.

It should also be noted that this analysis of the results is not necessarily scientifically rigorous and should be mainly considered as a tool for visualization and inspiration for future research. To more accurately determine whether Elon Musk's overall Twitter use could be considered toxic, it would be beneficial to compare his toxicity values to those of other social figures on Twitter. This approach would provide a more comprehensive and fair evaluation of his online presence, as it would take into account the broader context of Twitter discourse. By comparing his toxicity values to those of other individuals with a similar level of visibility or influence, it would be possible to better understand how his use of the platform compares to that of others, and to determine whether his behavior is consistent with societal norms and standards. This type of comparison would provide a more scientifically rigorous and robust analysis of the data, as it would consider multiple variables and factors that may impact the results.

The recent acquisition of Twitter by Elon Musk presents an opportunity for future research to examine the impact of this acquisition on the toxicity of his tweets or the overall toxicity on the platform. In order to accurately assess any changes resulting from the acquisition, it is necessary to allow sufficient time for any implemented changes to be fully integrated. While the recent nature of the acquisition prevented the examination of this question in the current analysis, it would be a valuable direction for future studies to explore.

Conclusion

In conclusion, this study demonstrated that the "toxicity," "insult," and "profanity" attributes of Elon Musk's tweets are highly correlated with each other, while the "severe toxicity," "identity attack," and "threat" attributes exhibit moderate correlation. By analyzing the toxicity scores over time, I found that Elon Musk's earliest tweets had the highest toxicity scores, with two notable peaks in 2012-2013 and 2019. There was also a slight decline in overall toxicity over time, although recent tweets from July-October 2022 suggest a potential increase in the future. Overall, Elon Musk's average toxicity score was relatively low at 5.3%, with only a small number of tweets exceeding a threshold of 0.7 or 0.9 for toxicity. The wordcloud analysis of the 48 most toxic tweets revealed that they often contained derogatory and offensive language and accusations of others being irrational or unstable. It could help identify potential triggers or causes of toxicity in his online behavior. The resulting dataset of this study can aid other researchers to further investigate patterns of toxicity in the online behavior of social figures on social media platforms, by comparison of various individuals or by analysis of trends over time. This information can inform the development of strategies aimed at promoting a safer and more inclusive online environment.

GitHub

The source code for this study, as well as the resulting dataframe, can be found in the project's Github repository at:
https://github.com/simonkruelle/Toxicity_Analysis_Elon_Musk

References

1. Dana Hull, Sarah Frier, and Maxwell Adler. Musk wants free speech on twitter after spending years silencing critics. *Bloomberg*, 04 2022.

2. Marina Koren. Of course Elon Musk wanted Twitter. *The Atlantic*, 04 2022.

3. Perspective api website. https://www.perspectiveapi.com/how-it-works/. Accessed: 2022-12-22.

4. Perspective api developer website. https://developers.perspectiveapi.com. Accessed: 2022-12-22.

5. Yasir Raza. Elon Musk tweets dataset (17k). https://www.kaggle.com/datasets/yasirabdaali/elon-musk-tweets-dataset-17k, 2022.

6. Marta Castrillo. Elon Musk's tweets dataset 2022. https://www.kaggle.com/datasets/marta99/elon-musks-tweets-dataset-2022?resource=download, 2022.

7. Communalytic. Toxicity analysis of a Twitter thread: Analyzing replies to president trump's tweet about contracting the novel coronavirus. https://communalytic.com/2021/01/13/toxicity-analysis-of-a-twitter-thread/, 2021.