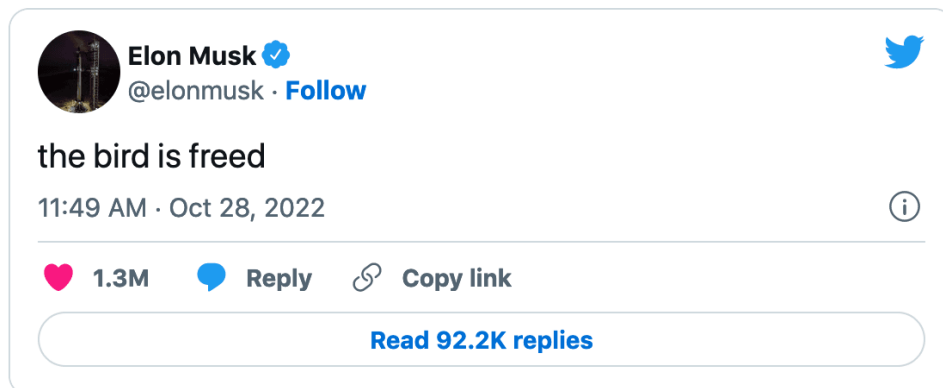


ONE YEAR IN: CONTENT MODERATION AND ANTISEMITISM ON X

Final Report



Prof. Sylvain Parasie
Sociology of the Digital Public Space
School of Public Affairs
Sciences Po Paris



Anthony Ammendolea, Joshua Bernstein, Christian Bissinger, Janine Ecker, Maria Chiara
Liviano D'Arcangelo

Table of Content

1. Introduction	2
2. Literature Review	3
Role of Content Moderation Policy	3
Content Moderation & Antisemitism on X	4
Methodological Contributions	5
3. Theoretical Argument	6
4. Quantitative Research Design	8
Choice of Sensitive Hashtags	8
Choice of Timeline	9
Scraping	9
Classification Antisemitic vs Non-Antisemitic using Natural Language Processing	9
Choice of Training Dataset & Pre-Trained NLP Model	10
Final Dataset & Event Study Using R	10
5. Results Quantitative Analysis	11
Hashtag Analysis	11
User Analysis	14
Results Event Study (Advanced Regression)	17
6. Qualitative Analysis	18
Cultural Transformation	20
Content Moderation Concerns	20
Antisemitism Landscape	20
Coping Mechanisms	21
Propaganda and Freedom of Speech	22
Financial Interests	22
Additional Themes	22
7. Discussion	23
8. Conclusion	24
9. References	26

1. Introduction

“I think it became a bit too much for me what Twitter was becoming. I felt as if there was this culture shift. Twitter was never really the ideal place to be in terms of hate speech. But it definitely changed.”

~ Nati Pressman, interviewed for the project

Elon Musk, the billionaire CEO of Tesla and SpaceX, officially bought Twitter for \$44 billion in October 2022. Shortly after the take-over, many users, just like Nati Pressman, voiced a feeling of sentiment change on the platform with increasing narratives of Twitter becoming a more toxic environment. In particular, vulnerable user groups such as women, LGBTQ+ community members and ethnic or religious minorities, have encountered higher levels of hate speech using the platform. This vague, seemingly personal feeling has been quickly corroborated by scholarly evidence which found a significant increase in hate speech on the platform (Hickey et al. 2023).

Why is that? It is important to understand that Elon Musk’s acquisition of Twitter was not simply a change in ownership, but marks a consequential turning point for one of the largest social media platforms globally. Twitter has been the primary online platform used for real-time information dissemination, facilitating public discourse on various topics, connecting individuals globally, and serving as a platform for news updates, trends, and discussions on a wide range of subjects. The rules for any interaction in this digital public space are set and enforced privately by the platform provider. Such measures are commonly referred to as content moderation rules. Elon Musk, a self-declared “free speech absolutist,” has made no secret of his intentions to decrease content moderation on the platform in his crusade to “protect free speech” (Dan Milmo 2022).

While content moderation is inherently an opaque and highly intransparent feature of digital platforms, the acquisition of Twitter by Elon Musk and the subsequent significant changes in content moderation policy present a unique opportunity to study the effects of this crucial tool. One year after the acquisition, we seek to provide an empirical analysis to answer the research question: *How did Elon Musk’s acquisition of Twitter in October 2022 influence antisemitic hate speech on the platform?*

In our research, we will focus on antisemitic hate as a specific form of hate speech on the platform. We employ a mixed-methods approach combining empirically quantitative and qualitative analysis. For quantitative analysis, we employ hashtag analysis and an advanced event study regression model. A major contribution of our study is also the extensive dataset of antisemitic hate speech Tweets that we have constructed, containing 158.439 antisemitic tweets in the period between 01.08.2022 and 30.09.2023.

The quantitative analysis is complemented by qualitative expert interviews to unravel further the underlying mechanisms and effects of content moderation and antisemitism on X.

Our report begins by giving an overview of the state-of the art of current research and consequently identifies the research gap that is tackled. We proceed with the theoretical part by outlining how the acquisition has influenced antisemitism on the platform through its change in content moderation policies. After, we outline the methodology of both the quantitative analysis using our newly constructed data set as well as the expert interviews conducted. Then, we discuss the results and elaborate on the implications of our findings in a substantive way. Lastly, we address potential limitations as well as future avenues of research.

2. Literature Review

The rise of digital public spaces and the inevitable emergence of online hate speech has been subject of profound scholarly interest in the last decade. There exists clear consensus about the rise of hate speech in digital public spaces. The growth of hateful movements like right wing extremism, white nationalism and nazi twitter follower growth have skyrocketed to rates of more than five times the rate from 2012 to 2016 (Berger 2016).

Role of Content Moderation Policy

These trends are indicative of a rising pattern of antisemitism that has continued to fester within different realms of the offline but notably also digital public space (Anti-Defamation League 2021). Tech giants like Meta, Google, TikTok and other platforms have aimed to counter antisemitism through content moderation policies. Content moderation is broadly defined as “the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse” (Grimmelmann 2015). The most common definition used is by Sarah T. Roberts who has introduced the term *commercial content moderation* which refers to “the organized practice of screening user-generated content posted on Internet sites, social media, and other outlets” (Roberts 2014). The user-generated content is analysed usually in a hybrid form, combining automated tools with human moderators for a comprehensive approach. The repertoire of content moderation tools to address hate speech ranges from flagging content and user warnings/educational prompts, removing hateful content, restricting comment functions to suspending user accounts or outright banning them. The practice of commercial content moderation received significant scholarly attention in recent years and has been deemed a crucial tool for providers to address and manage hate speech and disinformation on their platforms (Gillespie 2020, 2018; Langvardt 2017; Roberts 2017).

Content Moderation & Antisemitism on X

Since Elon Musk acquired Twitter in October of 2022, discourse on hate speech on the platform including antisemitism has gained significant public attention and sparked renewed scholarly interest. Firstly, it is to note that research confirms the overall increase of antisemitic hate speech on the platform since Musk's acquisition. However, researchers such as Hickey et al. (2023) call for further research on the nuances in this rise of antisemitic hate speech. In their recent-published study they investigated the engagement of real users versus bots and confirmed a presence of spammer bots which echo certain viewpoints and inflate their presence on a platform. They acknowledge the general increase in the level of hate speech on the platform, yet call for more in-depth research and case studies to understand these dynamics better (Hickey et al. 2023).

While the trend of increasing antisemitism and hate speech more broadly is unquestioned, the degree to which this is related to the acquisition by Elon Musk and the underlying mechanisms sparks division among scholars. Some find that this is directly related (Miller et al. 2023; Hickey et al. 2023; Benton et al. 2022). On the other hand Jikeli and Soemer (2022, 2023) conclude that they cannot rigorously establish that Elon Musk's acquisition of X caused the rapid spike in antisemitism on the platform. They point to potentially confounding factors including the presence of Kanye West's antisemitic tirade or professional basketball player Kyrie Irving's antisemitic posts which significantly drove up the level of antisemitism and thus constitute outliers rather than long term trends.

A relatively new but promising strand of research is emerging to further investigate through which channels content moderation influences the spread of hateful content and user behavior (Arttime et al. 2020, Miller et al. 2023). They focus on the dynamics through which content moderation policy and changes therein drives up hate speech. Miller et al. investigated the rise in antisemitic hate speech shortly after the take-over and observed a surge in the creation of new accounts with a total of 3.855 accounts posting at least one antisemitic tweet which have been created between October 27 and November 6. The authors suggest that users "feel empowered by Musk's widely publicized shifts to Twitter's management" and conclude that this dynamic contributes to an increase in antisemitism on the platform by 106% (Miller et al. 2023). The same logic is expressed by a study conducted by René D. Flores who studied the Arizona Immigration Law and analysed 250.000 tweets on anti-immigration sentiments and found that changes in public discourse were not caused by shifting attitudes toward immigrants but by the mobilization of anti-immigrant users and by motivating new users to begin tweeting (Flores 2017). This push-and-pull effect is confirmed by Ribiero et al. who focus on the migration effects of radical users towards less restrictive platforms in order to post antisemitic content.

They theorize that relatively higher content moderation on other platforms will motivate users to move to different platforms that moderate less heavily (Horta Ribeiro et al. 2021). In our case, a reduction of content moderation on X and the decreased threat of punishment would imply that radical users might be more encouraged to use X than before.

Methodological Contributions

The existing research provided meaningful contributions to the quantitative and qualitative analysis methods employed in this study. We want to highlight two studies which made notable contributions in regards to methodology. Firstly, Jikeli and Soemer (2023) used a combination of qualitative and quantitative approaches to examine the rise of antisemitism on X during Elon Musk's acquisition. They employed frequency and hashtag analysis, identifying keywords and messages indicating antisemitic sentiments. An important methodological contribution was the use of manual annotation, where two trained identifiers assessed tweets based on the International Holocaust Remembrance Alliance's definition of antisemitism. This method allowed for detailed analysis, revealing responses to antisemitism, such as counter-speech. This study influenced our research methodology, incorporating similar quantitative methods and the Anti-Defamation League's definition of antisemitism. Secondly, another noteworthy contribution to quantitative methodologies came from Hickey et al. (2023), who conducted timeline analysis on user accounts before and after Musk's acquisition of X. This approach helped measure the impact of the acquisition and detect automated accounts (bots). Lastly, Criss et al. (2021) provided a framework for interview methodology, developing questions to explore participants' physiological and psychological reactions to online discrimination. This framework guided our understanding of how participants reacted to antisemitic insults and their responses.

Conclusively, the literature about the intricacies between content moderation and hate speech, in particular antisemitic hate speech, is more relevant than ever. The groundwork has been laid pointing to a clear rise of antisemitism online. However, the literature review shows that this alarming trend is far from conclusive and especially lacks robust quantitative evidence that not only suggests a correlation but rigorously tests the causal relationship between Musk's acquisition of Twitter and the increase in antisemitic hate speech. One year after the acquisition, we find ourselves at a crucial point to evaluate the developments so far in order to make sound claims about what has happened so far, and more importantly, what lessons can be learnt both for the public as well as for policymakers to effectively address these challenges in the future. Furthermore, our research not only aims to provide valuable insights into the drivers of this phenomenon but also creates a comprehensive dataset of antisemitic tweets which can serve as the basis for future research.

3. Theoretical Argument

Clearly, the acquisition by Elon Musk per se did not influence hate speech. What exactly drives the pattern of increased hate speech? We theorize that the change in content moderation under Musk's leadership is moderating the causal relationship between the acquisition of X and the increased antisemitic hate speech.

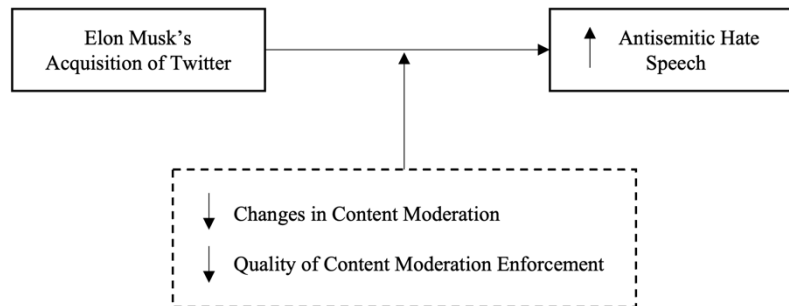


Figure 1 Visualization of Theory

We argue that Elon Musk's acquisition of Twitter has increased antisemitic hate speech through a change in content moderation. Our research thus seeks to test the following hypotheses:

H1: *The volume of antisemitic hate speech tweets has increased after the acquisition.*

H2: *Variations in content moderation policies have led to an increase in antisemitic hate speech on X.*

What has changed with regards to content moderation on X following Musk's acquisition? In its nature, content moderation is a highly opaque feature of digital platforms whose exact practices and consequently changes of it are only known to the platform itself. To study changes in content moderation it is therefore crucial to observe communication of the platform about policy changes, the launch of new features and study narrative evidence of users' experiences.

Changes arise from both changes in the substantial approach as well as from the quality of enforcement which we argue has significantly been weakened through the mass layoffs shortly after Musk's take-over. In the following we will briefly outline the changes in X's content moderation policies derived from their official communications.

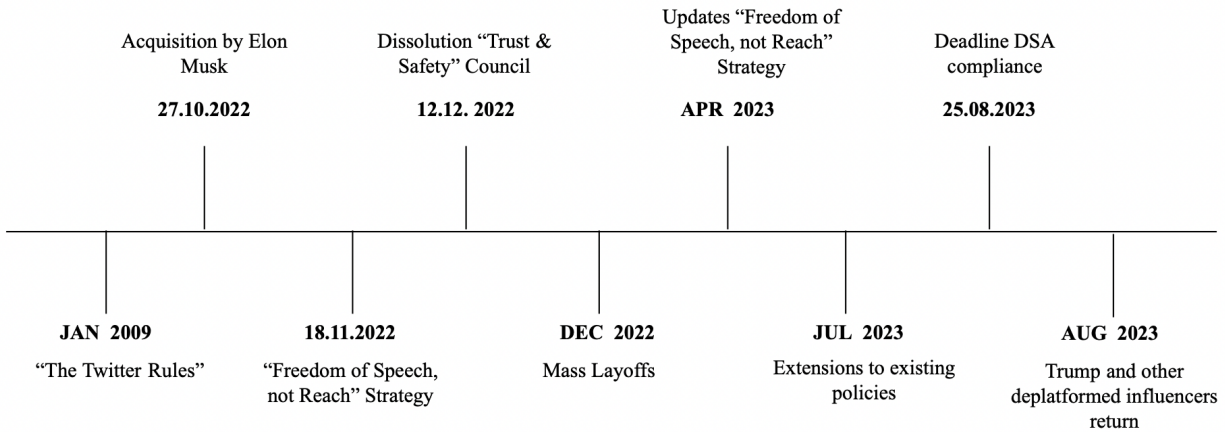


Figure 2 Timeline Key Events Content Moderation Policy Changes X

18.11.2022 – Introduction “Freedom of Speech, not Reach” Policy: The policy marks a new direction of content moderation, pursuing a model focused on limiting the visibility of hateful content instead of taking content down and blocking users. Musk announced that negative/hate tweets will be deboosted and demonetized. This constitutes a significant diversion from the original community guidelines.

12.12.2022 - Dissolution of the “Trust and Safety Council”: Musk dismissed the Trust and Safety Council, a volunteer group of around 100 independent civil, human rights, and other organizations, formed in 2016 advising the social media platform with the goal of improving the combat of hate speech and child exploitation. The same day, three members of the council resigned, denouncing publicly that “contrary to claims by Elon Musk, the safety and wellbeing of Twitter’s users are on the decline.”

12/2022: Mass Layoffs: Musk announces the layoff of more than 3.700 employees and more than 1.000 employees chose to leave the company such as Yoel Roth, Head of Trust & Safety. Among the many dismissed employees, a large part of the content moderation team has been suspended.

17.04.2023 - First Update “Freedom of Speech, not Reach” Policy / Choice of Twitter 2.0”: As part of the first update, X announced the visibility filter feature, which will add a label to tweets that likely violate the policy. Musk justified this as a tool that “allows us to move beyond the binary ‘leave up versus take down’ approach to content moderation”. X announced that illegal content will continue to be removed, and users suspended but this change in policy raises concerns about the extent of content moderation employed.

12.07.2023 - Last Update of the “Freedom of Speech, not Reach” Policy: X announced policy amendments to extend the Abusive Behavior and Violent Speech policies. Twitter claims that more than 99.99% of tweet impressions are from healthy content, or content that does not violate our rules.

4. Quantitative Research Design

Firstly, the quantitative part of our study aims to empirically investigate the relationship between Musk’s acquisition of Twitter and antisemitic hate speech by using the largest possible sample of antisemitic tweets through relevant hashtags and quantification of how much changes in Twitter’s policies have impacted the rise of antisemitism. A key contribution of our research is the creation of a novel data set containing 158.439 antisemitic tweets which we use to empirically test our hypotheses. The methodology is visualized in *Figure 3* with each step explained below.

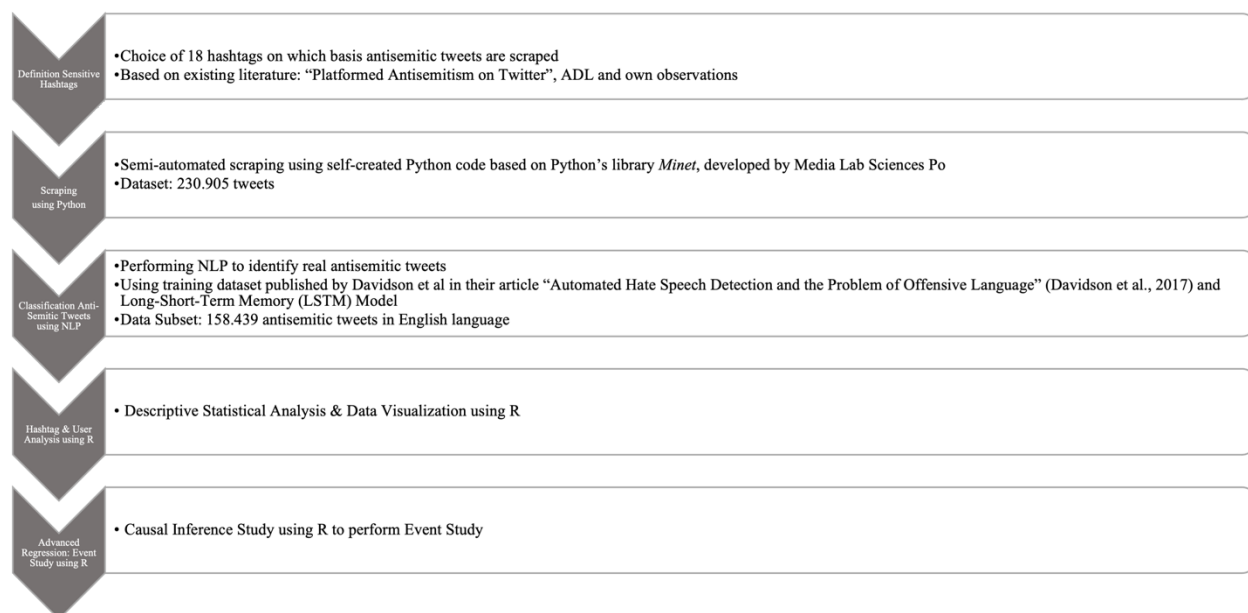


Figure 3 Quantitative Analysis Process

Choice of Sensitive Hashtags

To build a dataset of relevant tweets, we scraped tweets containing hashtags that were classified as potentially antisemitic. Those hashtags have been selected using two justification strategies: we mainly relied on existing literature and reports of agencies fighting antisemitism which have identified sensitive hashtags mainly building on the article “Platformed Antisemitism on Twitter” (Riedl et al. 2022) which identified several antisemitic hashtags based on quantitative content analysis. Moreover, we relied on hashtags classified as likely antisemitic by the Anti-Defamation League. The rest of the hashtags have been chosen based on our initial observations and based on similarities to the previously explained. Given the time intensity of scraping we limited the number of selection hashtags to 18 hashtags (see *Table 1*)

Table 1 Choice of 18 Antisemitic Hashtags

#rotschild	#jewishprivilege	#unbonjuif
#soros	#killthejews	#yeisright
#zionist	#hitlerwasright	#zionistjews
#thejew	#bantheadl	#zionismnazism
#jew	#exposethenose	#ye24
#jews	#jewishsupremacy	#jewishracialsupremacism

Choice of Timeline

In order to evaluate the impact of Elon Musk’s acquisition of Twitter and the implementation of new content moderation policies we have chosen the following period: 01.08.2022 T00:00:00 until 30.09.2023 T23:59:59. At this point, it is crucial to mention that we have purposely excluded the Hamas attack on Israel on 07.10.2023 which have dramatically increased antisemitism online. This choice was motivated by two reasons. From an operationalization perspective the scraping became impossible due to the influx of daily tweets which far exceeded the level prior to October 7th. More substantially, we limited our research as this political event can be seen as an outlier which presumably distorts the results.

Scraping

To build the dataset we had to gather a sufficient quantity of potentially antisemitic tweets. To maximize the number of tweets and given the limited choice in available datasets for this purpose, we chose to directly scrape tweets from X. This strategy required us tackle two main issues: Firstly, while the Twitter API could be previously used to scrape tweets free of charge, Elon Musk blocked all “simple solutions” for scraping from X’s using API in August 2023 and introduced the API as a paid feature. They also restricted access from well-known Python’s scraping libraries (Tweepy). Secondly, the scraping requires a precise query, that is to say the user has to specify what specific tweets he wants to scrape (part of the text, dates, user, etc.) instead of broad access to a large number of tweets. To collect tweets, we had to develop a novel coding strategy from scratch, inspired by Python’s library *Minet*, developed by the Media Lab of Sciences Po (see [notebook](#) for all technical details). This semi-automatic solution allowed us to scrape a substantial amount of tweets. It has to be mentioned that this process was immensely time-intensive and required a very deep level of technical knowledge and dedication.

Classification Antisemitic vs Non-Antisemitic using Natural Language Processing

In total we collected 230.905 tweets, based on the research of 18 different hashtags in the period between 01.08.2022 and 30.09.2023, with less than 0,02% dates missed. This dataset contains all tweets that include at least one of the targeted hashtags. However, the inclusion of the hashtag does not qualify by

itself as an antisemitic tweet. While they are more likely to include an antisemitic message than other tweets, they could also use the hashtags to denounce antisemitic behaviors. To have reliable data, we had to quantitatively distinguish between real antisemitic tweets and non-antisemitic tweets. We used pre-trained Natural Language Processing (NLP), which is a specific field of Deep Learning, including Large Language Models (LLMs) to analyze the content of tweets in a coherent way. The objective was to find a training dataset which labels a sufficient number of tweets as hate speech which can be used with a pre-trained model that statistically analyses our overall dataset of tweets to filter out the real antisemitic tweets. Constraints hereby were finding an efficient enough that is in accordance with our resources, both time-wise and regarding GPU.

Choice of Training Dataset & Pre-Trained NLP Model

We used the training dataset published by Davidson et al. in their article “Automated Hate Speech Detection and the Problem of Offensive Language” (Davidson et al., 2017) and which contains 24.783 labeled tweets classified as “hate speech” (1.430), “offensive language” (19.190), and “neither” (4.163). As the scope of our study is about antisemitism only, we considered only “hate speech” tweets. Secondly, we selected the pre-trained Long-Short-Term Memory (LSTM) Model. It builds on a deep learning architecture which is able to catch the “attention” of a tweet, that is to say, the sentiments behind a tweet. For time reasons we worked with 50 dimensions tokens out of the 300 available. We performed the training of the dataset on the previous training dataset which yielded the following metrics: 89,91% accuracy, 95,26% recall, and 91,88% precision (see final [dataset](#) and [notebook](#) for detailed code).

Final Dataset & Event Study Using R

This process resulted in a final dataset containing 158.439 tweets in English language (68,8% of all tweets). The prediction classified 26.934 tweets as antisemitic. These antisemitic tweets serve as the sample to test our hypotheses, we employed an event study (advanced regression). It is a difference-in-difference analysis measures the impact of a specific date on the increase in the number of antisemitic tweets. The regression requiring two sample groups which will be compared. The control group is the dataset of all English tweets that haven’t been flagged as antisemitic, amounting to 131.505 tweets. We have selected five dates (see *Table 2*) following the acquisition of Twitter by Elon Musk, based on their relevance and possible impact on content moderation on Twitter as explained in the theoretical section of this report.

Table 2 Selection of Key Events

Date	Event
27.10.2022	Acquisition of Twitter by Elon Musk
18.11.2022	Implementation of new policy on content moderation “Freedom of Speech, not Reach”
12.12.2022	Dissolution of the “Trust and Safety Council” of Twitter
17.04.2023	First update of the Freedom of Speech, not Reach” policy
12.07.2023	Last update of the Freedom of Speech, not Reach” policy

We have used R to perform the regression analysis (see [R code](#)).

5. Results Quantitative Analysis

Hashtag Analysis

The first two graphs give an overview of the trend of the total amount of antisemitic tweets over the studied period, where each vertical line corresponds to a event date, representing a milestone in X’s moderation policy change. *Figure 5* is a zoom of the first figure highlighting the period August 2022 until June 2023 to understand developments after the acquisition better. It becomes clear, that the number of antisemitic tweets has increased over time with peaks, most significantly after the introduction of the “Freedom of Speech, not Reach” policy.

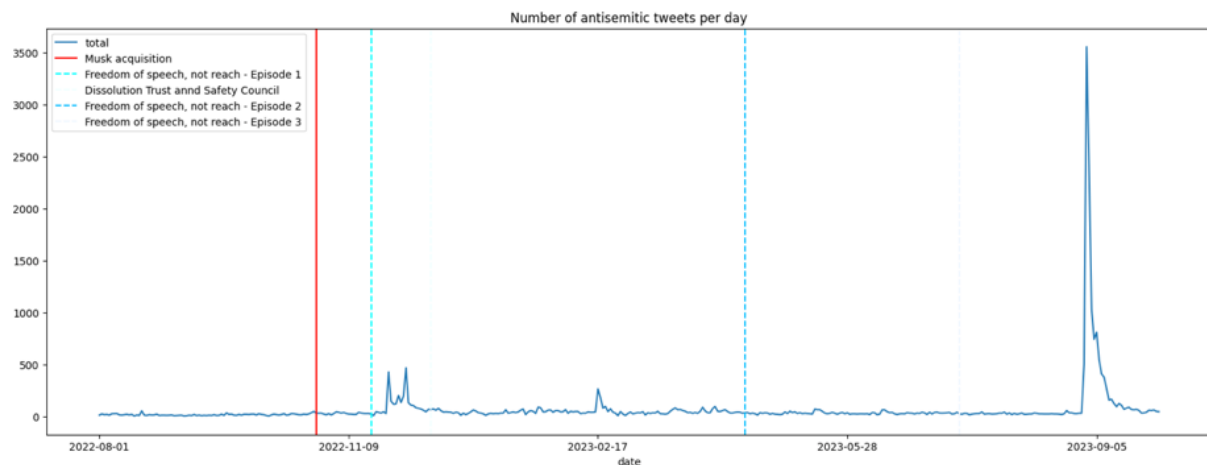


Figure 4 Number of Antisemitic Tweets by Day

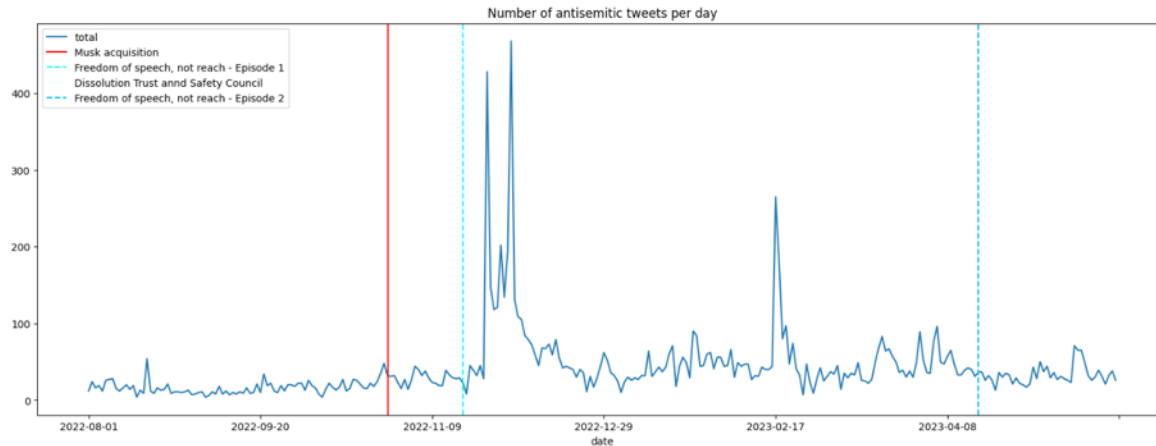


Figure 5 Zoom: Number of Antisemitic Tweets by Day

Figures 6 & 7 provide an overview of the tweet evolution per day including the specific antisemitic hashtags. It allows us to get a more fine-grained picture of the developments and what has driven the increases. Notably it provides insights into the two peaks in antisemitic hate speech. The first peak of December 2022 can be associated with the ban of Kanye West, called “ye” on Twitter, based on his repeated antisemitic incitement to violence, culminating in his adoration for Hitler (Rania Aniftos 2022). The second peak in September 2023 is correlated with the movement against the Anti-Defamation League (ADL), a civil rights organization combating antisemitism online. Following several claims by the ADL about the increase in antisemitism on X, Musk actively promoted the boycott of the ADL (Feuer 2023)

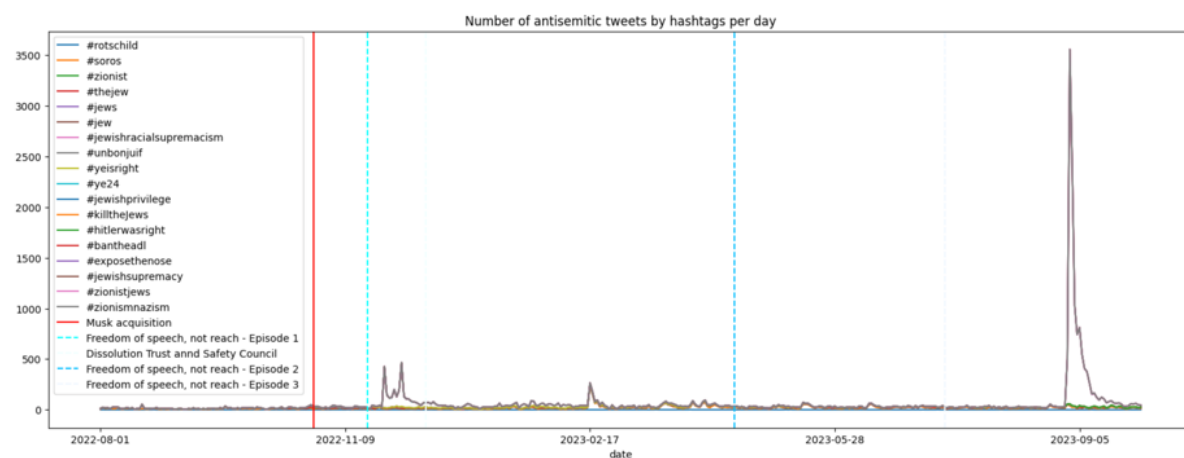


Figure 6 Number of Antisemitic Tweets by Hashtag by Day

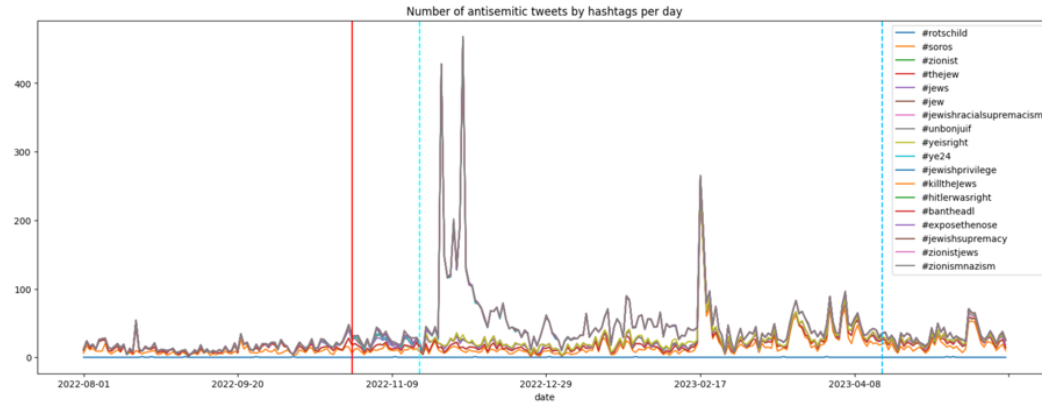


Figure 7 Zoom: Number of Antisemitic Tweets by Hashtag by Day

Figure 8 provides an insight into the number of antisemitic tweets per hashtag, visualized by month. It allows us to confirm that there has been a significant increase in antisemitic tweets following Musk's acquisition of Twitter.

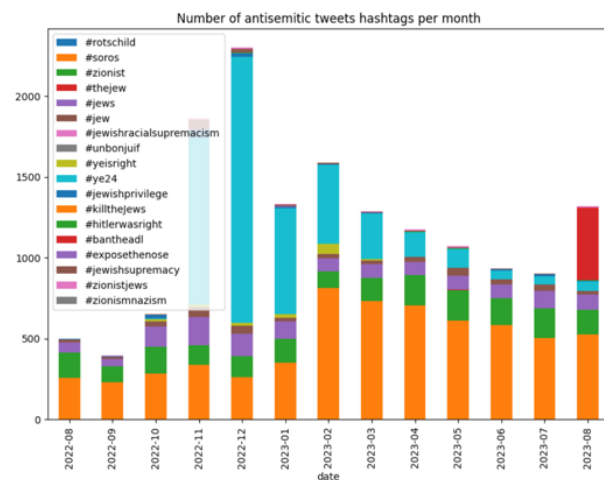


Figure 8 Number of Antisemitic Tweets by Hashtag by Month

Another visualisation to provide more substantial insights into the antisemitic hate speech is to look at the word clouds before and after Musk's acquisition. Figures 9 and 10 show respectively the Top 100 hashtags before and after Musk's acquisition, relative to their frequency. Comparing the word clouds we notice a higher diversity of antisemitic hashtags after Musk's acquisition which hints at a more diverse trend in antisemitism compared to spikes that have derived from common themes or debates before.



Figure 9 Word Analysis Pre-Acquisition

Figure 10 Word Analysis Post-Acquisition

User Analysis

The following graphs visualizes the user analysis we have conducted to unravel the nature of the rise in antisemitism and provide insights into the users responsible for the antisemitic hate speech.

First, we will focus on the user's activity, to understand whether there is a large community or active users or rather a small community responsible for the antisemitic tweets. We find that the 26.934 antisemitic tweets have been created by 12.410 different users. *Figure 11* depicts the number of tweets created by users per quartile during the entire studied period. The graph highlights that most users had a very moderate contribution to antisemitic and 90% of users have published less than three tweets. On the contrary, the more active users represent a comparatively small minority which is much more active, underscored by the last quartile which argues that only the top 0,1% of most active users has generated more than 70 tweets. Thus, it seems that the increases have been driven by a small fraction of users. A closer look at the top 10% (quartile [90%-100%]) of most active users in *Figure 12* highlights large disparities with the average tweets per user at 10 tweets while the most active user responsible for 300 antisemitic tweets, followed by six other users with more than 100 tweets considered as antisemitic. Thus, we conclude to have a significant heterogeneity of users with the increase driven by a minority of very active users.

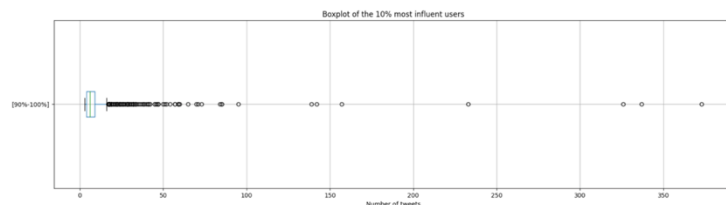
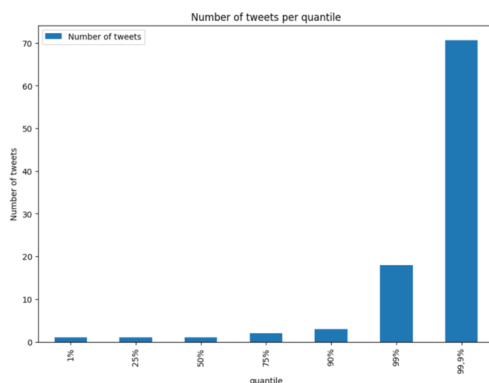


Figure 11 Number of Tweets by User Quartile

Figure 12 Antisemitic Activity of the Upper 10% of Active Users

Looking at the regularity of the most active users, *Figure 13* shows the activity of the ten most active users during the entire period, which confirms a regular involvement over several months. To confirm the preponderance of the activity of the most influential users, that seem to stay month after month, *Figure 14* shows the dominance of these few users whose activity fluctuates between 7% and 33% for the top ten only and 59% of the total activity for the top 100.

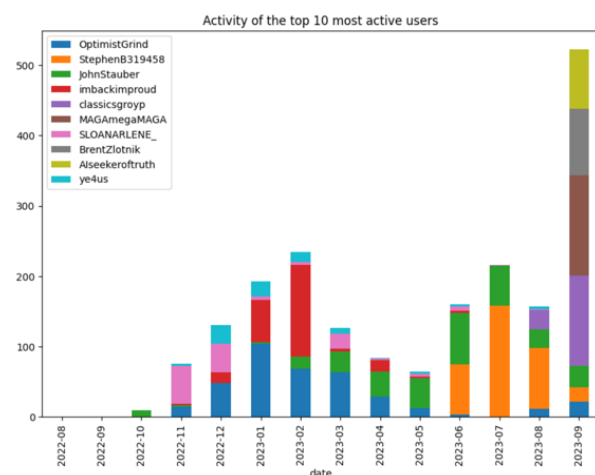


Figure 13 Activity of the Top 10 Most Active Users

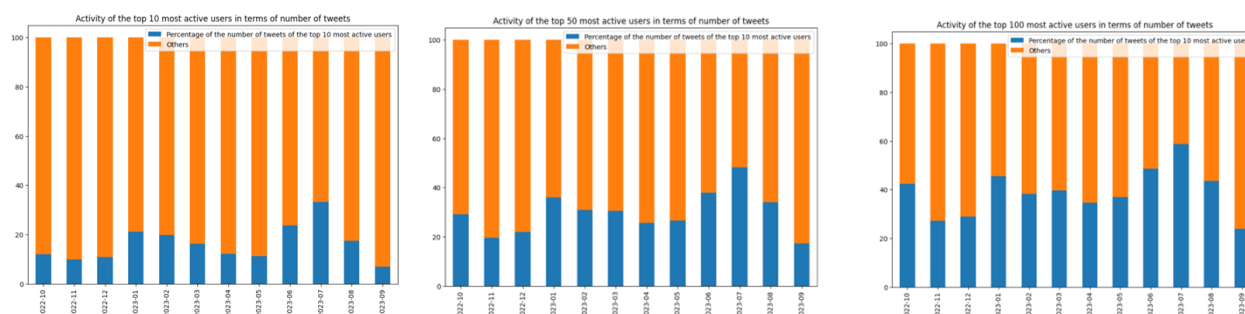


Figure 14 Comparison Activity of the Most Active Users by Number of Tweets

Furthermore, we focus on the evolution of different users during the entire period and the entrance of newcomers. We can obtain two interesting findings from the data: Firstly, there is a fluctuation in the number of different users, depending on the month with no significant rise in new entrants, except in September 2023 as depicted in *Figure 15*. However, *Figure 16* underscores the significant turnover of users, both month-on-month and compared to September 2022. Indeed, these graphs show that for most of months, more than 80% of users have not published antisemitic tweets in the following month. Moreover, only 8% of users that have tweeted something in September, have also tweeted something in September 2023. This finding is interesting as it hints to a “casual antisemitism” and not a pursuit of an antisemitic agenda. We have to note, that this trend deserves further research as it could be driven by e.g. users being blocked from posting.

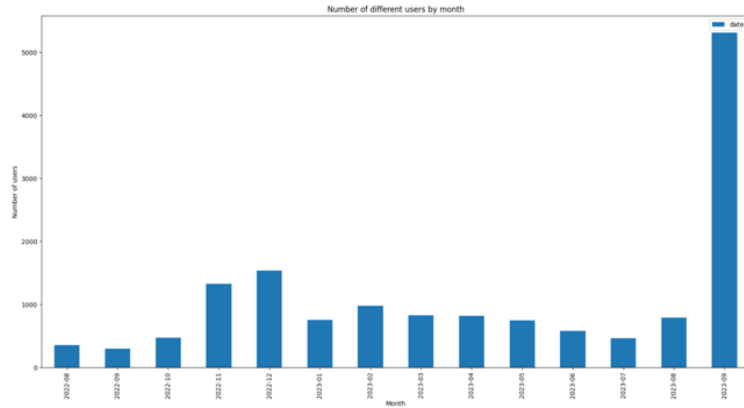


Figure 15 Number of User Variation by Month

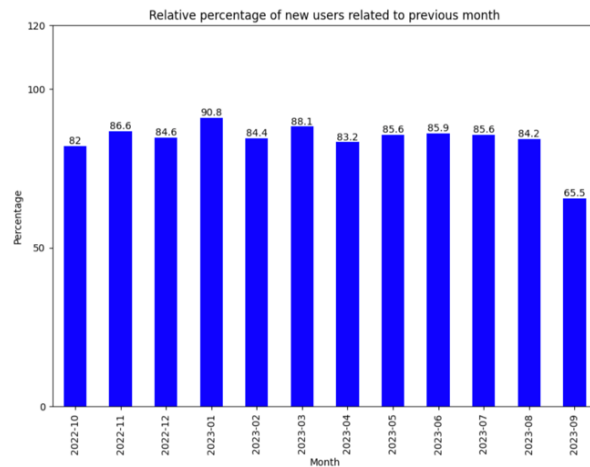


Figure 16 Change of New Users Compared Previous Month

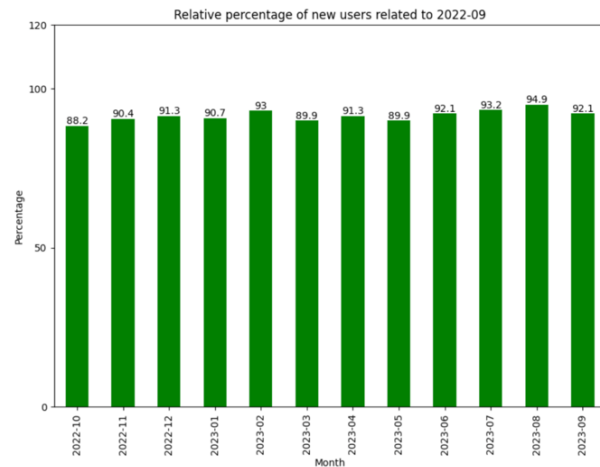


Figure 17 Change of New Users Compared to 2022-09

Overall, concerning the users that drive the increase we can conclude that the main traffic of antisemitic tweets is due to a small minority of users (less than 100 individuals), very active (more than half of the traffic over several months), whose activity is very steady over the period studied. Moreover, we notice a large month-on-month turnover, outlining a typical type of user, who tweets once or twice and will never send another antisemitic tweet. We notice a lack of a steady increase in the number of “antisemitic users”, mainly due to the huge turnover of users. Following the acquisition we can observe an increase in the cumulative number of users with a “one-off” antisemitic behavior. However, we have to note that more in-depth research would be required to study the antisemitic minority responsible for the majority of antisemitic tweets as it could be expected that the policy changes might have attracted more of these hard-core, structured antisemitic users.

Results Event Study (Advanced Regression)

The main goal of this part is to quantify the impact of each flagged date on the rise in the number of antisemitic tweets. All the process has already been described in the methodology part and this part will present the results with an interpretation and the limits of the study. The different dates of interest described below have been selected for their potential huge impact on content moderation of X, regarding the changes implemented.

Table 3 provides an overview of the statistical analysis. According to the results and the interpretation of the R^2 metric, the three most significative dates, that is to say the dates and the measures that had the main impact on the rise in the number of antisemitic tweets, are 27.10.2022 with the acquisition of Twitter by Elon Musk with an 84% explanation of the model, 17.04.2023 with the introduction of the “Freedom of Speech, not Reach” policy with a complete (100%) explanation of the model and 12.07.2023 which introduced the latest update of the “Freedom of Speech, not Reach” policy with a 82% explanation of the model. While all of the dates are statistically significant, the two remaining dates could explain 30% of the variation.

An explanation of the the major effects of the dates 17.04.2023 and 12.07.2023 could be that the real “revolution” in X content moderation policy happened through the implementation of masks of filtering, allowing complete freedom of speech for those who decide to watch sensitive content and which seems to feed the spread of antisemitic behaviors on X. The 12.07.2023 date refers to the same policy and counter-intuitively, the expansion regarding Abusive Behavior and Violent Speech policies did not appear to have prevented or slowed down the rise of antisemitism on the platform.

time_to_treat_5:-12:antisemitic		-4367.00 ***
time_to_treat_5:-11:antisemitic	(0.00)	-3864.00 ***
time_to_treat_5:-10:antisemitic	(0.00)	-5735.00 ***
time_to_treat_5:-9:antisemitic	(0.00)	-8195.00 ***
time_to_treat_5:-8:antisemitic	(0.00)	-8835.00 ***
time_to_treat_5:-7:antisemitic	(0.00)	-6024.00 ***
time_to_treat_5:-5:antisemitic	(141.28)	-6134.50 ***
time_to_treat_5:-4:antisemitic	(0.00)	-6144.00 ***
time_to_treat_5:-3:antisemitic	(0.00)	-6147.00 ***
time_to_treat_5:0:antisemitic	(0.00)	-5866.00 ***
time_to_treat_5:1:antisemitic	(0.00)	-6048.00 ***
time_to_treat_5:2:antisemitic	(0.00)	-24148.00 ***
Num. obs.	28	28
Num. groups: date	14	14
R ² (full model)	0.99	0.97
R ² (proj model)	0.99	0.94
Adj. R ² (full model)	0.84	0.30
Adj. R ² (proj model)	0.81	0.13
	28	28
	14	14
	1.00	0.99
	1.00	0.97
	1.00	0.82
	1.00	0.78

*** p < 0.001; ** p < 0.01; * p < 0.05

Table 3 Overview Regression Results

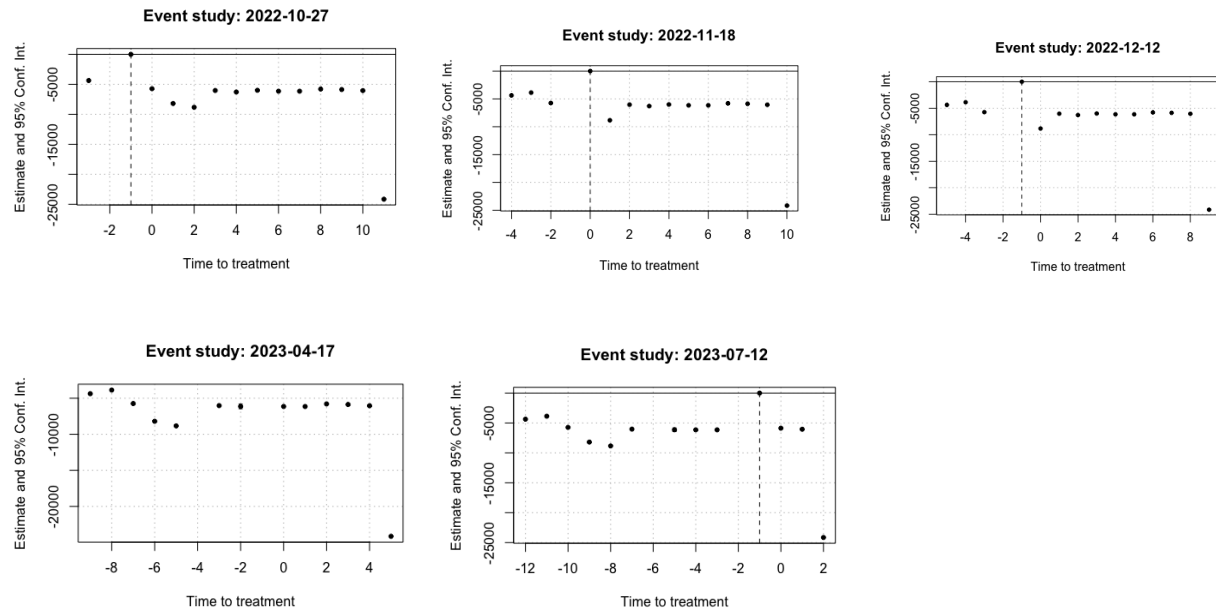






Figure 18 Results of Event Study for Key Events

Overall, the quantitative analysis allowed us to investigate the increase in the rise of antisemitic behavior on X using descriptive analysis, gaining rich insights of the impact of each hashtag on the number of antisemitic tweets as well as the behavior of users. We can thus confirm our first hypothesis (H1) and conclude that there has been a statistically significant increase in the volume of antisemitic hate speech tweets after the acquisition of Twitter by Elon Musk. Moreover, through the advanced linear regression we were able to quantify the impact of the key events in the timeline of X's content moderation policy which corroborates our second hypothesis (H2) that variations in content moderation lead to an increase in antisemitic hate speech on Twitter/X.

6. Qualitative Analysis

To enrich the quantitative analysis method of our paper we incorporate qualitative research using expert interviews. We interviewed four individuals deemed highly pertinent to the subject matter, each representing diverse perspectives. Our participant cohort comprised three university professors, with expertise spanning Jewish studies and law, alongside a dedicated activist. From a demographic standpoint, our cohort was gender-balanced, featuring two women and two men. Moreover, three participants were aged over 50, while one belonged to the younger demographic, being under the age of 20. Crucially, all interviewees self-identified as Jewish and were active users of Twitter/X. Comprehensive details about our esteemed interviewees can be found in *Table 4* (see [here](#) details and full interviews).

Interviewee	Short Description & Justification
<p>Prof. Marc Caplan</p> 	<p>Prof. Dr. Caplan is a professor for Jewish studies and currently a senior lecturer at the university Heinrich Heine in Düsseldorf, Germany. He has obtained his Bachelor at Yale University and holds a Master in Comparative Literature from New York University. He has taught in the United States, Poland, Israel and Germany. He specializes in Yiddish Literature, Comparative Literature, Jewish Literature and Jewish Culture in Eastern Europe. He was one of the signatories of the “X Out Hate” coalition which published an open letter calling out antisemitism on X. We have reached out to Prof. Dr. Caplan because of his extensive academic background in Jewish studies, his activism in the X Out Hate coalition and because he worked for more than a decade for the Anti-Defamation League, a Jewish NGO that combats antisemitism. His full CV can be found here.</p>
<p>Prof. Eric Fink</p> 	<p>Prof. Fink is a professor for Law and currently a professor at the Elon University School of Law in North Carolina. He has obtained his education in both sociology and law from NYU School of Law, University of Chicago, London School of Economics and John Hopkins University. From 1997 to 2007 he practiced law in different states in the US. Since 1992 he has taught law in the United States and Czech Republic. He specializes in employment and labor law, administrative law, civil procedure, and has a research interest in socio-legal perspective. He was one of the signatories of the <u>“X Out Hate” coalition</u> which published an open letter calling out antisemitism on X. We have reached out to Prof. Dr. Fink because of his academic background in law, his activism in the X Out Hate coalition and his research interest to combine sociology and law. His full CV can be found here.</p>
<p>Prof. Alana Vincent</p> 	<p>Prof. Vincent is an Associate professor at the Department of Historical, Philosophical and Religious Studies of Umeå University. She has previously held posts at the University of Chester (UK), Glasgow University (UK), and Swedish Theological Institute (Jerusalem). Her background is in Jewish Studies (modern Judaism, holocaust and genocide, Jewish-Christian dialogue) and religion and literature. She holds a PhD from the Centre for Literature, Theology, and the Arts of the University of Glasgow. She was one of the signatories of the “X Out Hate” coalition which published an open letter calling out antisemitism on X. We have reached out to Prof. Vincent because of her extensive academic background in Jewish studies and her activism in the X Out Hate coalition. Her most recent CV available online can be found here.</p>
<p>Nati Pressman</p> 	<p>Nati Pressman is a student at Queen's University in Kingston, Canada. She is currently completing a Bachelor of Arts (Honours) Majoring in History with a minor in Jewish Studies. Nati is a staunch advocate for combating antisemitism, as a published columnist with Toronto’s popular newspapers, a Ronald S. Lauder Fellow and currently a Co-Chair Canadian Jewish Political Action Committee Fellowship program. In addition to her work, Nati also constantly advocates for Jewish voices on her university campus and conducts holocaust education. We reached out to Nati because of her extensive advocacy background and activism. In addition to the data we have collected from twitter, Nati’s experience under Elon Musk’s new content moderation policies are useful to gain insights into if and how present antisemitism is on Twitter/X.</p>

The insights gleaned from the conducted interviews with members of the Jewish community shed light on several discernible patterns within the context of antisemitism on the platform X.

Cultural Transformation

All participants attested to a notable cultural shift on Twitter, where the overall sentiment appeared to worsen after the acquisition. Nati Pressman stated: “I felt as if there was this culture shift. Twitter was never really the ideal place to be in terms of combating hate speech, but it definitely changed and I noticed it much more”. Similarly, Professor Vincent declared: “it is quite clear the direction of the platform shifted significantly.” This transformation was characterised by an increase in toxicity and a heightened interaction with accounts and content espousing hate, as professor Caplan highlighted: “I do think that when Elon Musk took over Twitter, it became much more aggressively toxic.” Notably, the platform transitioned from merely hosting hate speech to actively encouraging it, as emphasised by Professor Vincent “There's a big difference between a platform where there is hate speech, there is violent rhetoric that we can't completely eliminate, that's the nature of social media being social. But there's a difference between that and platforms that seem to actively encourage. And post 2022 Twitter has been the latter”. Professor Fink even noticed that “most well known accounts that had been banned before or suspended before were just brought back, many of them as verified. Accounts that are now the paid accounts”.

Content Moderation Concerns

The cultural shift is directly related to the reduction in content moderation efforts on Twitter perceived by our Interviewees. According to Nati Pressman, Elon Musk “literally cancelled the moderator system.” User-driven moderation tools that our interviewees used to regularly use in the past, notably the report and block functions, were deemed less effective. Professor Fink recollected that “when I did report things, the responses seemed to be much less frequent, there was no action taken at all and much more frequently they would say, you know, this just doesn't violate our policy.” He also underscored a sense of inadequacy and a perceived lack of good faith efforts in Twitter's content moderation practices: “I think even in the past, it wasn't, it was inadequate, but there were at least at least I had the impression that they were trying [...] but now, there seems to be no good faith effort.” Other similar platforms like Instagram (IG) and Facebook (FB) were considered to have more robust moderation mechanisms.

Antisemitism Landscape

A confirmed increase in antisemitic incidents on X was a prevalent theme across interviews. In particular, our two female interviewees, had experienced direct antisemitism, while Professor Caplan reported witnessing similar events to two close female friends of his. Nati Pressman recollected that after responding to a post she deemed inappropriate, multiple users started retweeting her and “saying that they want to kill me and run me over.” Moreover, Professor Vincent told us that “there was one account that old Twitter banned repeatedly and kept coming up repeatedly, something like “fellow white people”. And that account existed entirely to screen capture and circulate pictures of Jewish people so that they and their followers could comment on me, and, oh, these people claim that they're white. It was vile. And my picture circulated fairly regularly along with personal details”. An interesting conclusion from our interview was that the extent of exposure to antisemitic content was portrayed as contingent upon individual engagement and characteristics. For example, professor Fink noted that “there's no antisemitism in Twitch and fishing Twitter, that I've seen, you know, people just talked about fish”, while professor Caplan argued that he had not been a direct victim of antisemitism because “I don't have a lot of followers. Somebody could target me, but where are they going to go with it? They're not going to get any exposure for doing it.” Additionally, the interviews highlighted the intersectionality of hate on the platform, with mentions of other forms such as misogyny and Islamophobia.

Coping Mechanisms

Diverse coping mechanisms were employed by interviewees to navigate the challenging online environment throughout the years. Before the acquisition, most interviewees were adopting playful problem-solving strategies, such as reporting and blocking hateful accounts. For example, professor Vincent stated: “I do not owe every single random person on the internet a fair hearing and an exhaustive dissection of what they're saying. Over the years, I have become much more aggressive “block and move on” because life is short, I am not going to spend more time trying to interpret someone's random internet posting than they spent writing.” After the changes in content moderation policies, users were forced to change their coping mechanisms as the report and block functions were not working anymore. As a consequence, three of them opted for escape avoidance, choosing to leave X altogether. For example, professor Caplan stated: “I made a vow to myself [...] that if Donald Trump is left back on the platform I'm leaving.” Professor Fink even reported the psychological consequences of constant exposure to hateful content: “after a while, when you just read this stuff all day, it kind of echoes around your brain. And it really takes a weight, and it was really causing me stress. So, I actually took it off my phone, and I stopped because it's, you know, it's not useful, and I don't want to subject myself to it.”

Our youngest interviewee reported instead that she “used to be very active on Twitter”, but now her “notifications are off.” The three professors interviewed also embraced confrontive coping, participating in initiatives such as the “X Out Hate” coalition, a group of almost two hundred Jewish leaders raising awareness for the problem of the widespread presence of antisemitism on X. In this regard, professor Caplan stated that “the ability of X to fail, can serve a very valuable cautionary lesson, to the oligarchs, and to the consumers of these platforms. Anything that causes Elon Musk discomfort, or loss of wealth is going to bring me some degree of pleasure, so I’m going to pursue it on those merits.”

Propaganda and Freedom of Speech

The concept of freedom of speech on X is a subject of considerable complexity and divergence of perspectives, as articulated by Professor Fink and Professor Vincent. Professor Fink believes that “there’s a lot of misunderstanding or misuse of this concept of freedom of speech [...] it’s become a term that people use ideologically to defend something that’s not really about freedom of speech at all”. He aptly pointed out that the legal principles of the First Amendment, which protect speech from governmental restrictions, do not directly apply to private communication on platforms like Twitter. Additionally, he expressed skepticism about the efficacy of algorithms, critiquing the notion of “freedom of speech, not reach” (the new content moderation motto of X) as a catchy but ultimately inconsequential slogan. Professor Vincent adds a nuanced dimension to the discourse, highlighting the European perspective that embraces constraints on certain types of speech to safeguard broader participation. She challenges the efficacy of the Anglo-American free marketplace of ideas model, suggesting that, in practice, it may fall short in the protection of diverse voices and perspectives: “the Anglo-American free marketplace of ideas model of free speech just doesn’t work.”

Financial Interests

The notion of the financial interests of X was a recurring theme throughout the interviews as well, scrutinised through the lenses of the diverse perspectives of Nati Pressman, Professor Caplan, and Professor Vincent. Nati Pressman shed light on the economic underpinnings of the platform, emphasising the ubiquitous presence of ads on Twitter and the revenue generated thereby. Her conscientious attempt to limit exposure to these ads exemplifies an awareness of the financial dynamics at play. Professor Caplan, in characterising X as a “toxic platform”, places responsibility on Elon Musk as a “powerful private citizen”, calling for collective efforts to curtail potential risks. Finally, Professor Vincent delves into the intricacies of financial interests, elucidating that the appeal to advertisers constitutes a significant leverage point, given that advertisers play a pivotal role in the platform’s revenue generation.

Additional Themes

Beyond the central themes, interviewees touched upon various other concepts integral to the discussion. These included social media bubbles, the notion of shadow banning, considerations of anonymity, concerns about misinformation, and the pervasive influence of algorithms, adding layers of complexity to the broader discourse on platform dynamics. For example, Nati Pressman believes she might have been shadow banned, because the numbers she gets on her tweets “aren't consistent with the followers” that she has, while Professor Caplan reflected on the topic of anonymity: “I think that the thing that I always found kind of just strange and maybe intimidating was the anonymity of Twitter. You know, I didn't like the idea that you're engaging with people that you have absolutely no idea who they are, and that they might not even be the people that they're claiming to be, you know, that aspect of Twitter, I found really long before Elon Musk, I found that to be a very off putting, I would say.”

7. Discussion

In conclusion, both the quantitative as well as the qualitative analysis corroborate our hypotheses and we can confirm that the volume of antisemitic hate speech has increased after Elon Musk's acquisition, as well as make robust statements that variations in content moderation policies have led to an increase in antisemitic hate speech. Notably, the introduction of the “Freedom of Speech, not Freedom of Reach ” policy has had a direct effect on levels of antisemitic hate speech.

The theoretical findings of our study not only add to the scientific literature in this field, but more importantly, they revisit the dilemma of content moderation with its delicate trade-off between protection of free speech versus limiting adverse social consequences, notably the observed rise in hate speech, as established in our research. The acquisition and significant content moderation shifts initiated by Elon Musk give rise to concerns regarding the accountability of private entities and even individuals within the digital public sphere. It prompts reflection on how society should respond to curtail the prevalence of hate speech and other harmful content. Our interviewees have resorted to endorsing online activism, such as the “X Out Hate Coalition”, which seeks to draw public attention to this issue and explicitly urges large advertisers to cease funding X through their advertising expenditures. While private combatting is definitely needed, the fundamental question emerges regarding whether the regulation of such vulnerable social spaces should be entrusted to private tech companies or to what extent policymakers should intervene to safeguard citizens.

In this context it is especially interesting to note that the European Commission has just announced on December 18th to launch "formal infringement proceedings" under the Digital Services Act (DSA) against Elon Musk's X platform to investigate whether X has breached obligations concerning countering the dissemination and amplification of illegal content and disinformation in the EU, transparency of the platforms and design of the user interface. EU Commissioner Thierry Breton remarks that "today's opening of formal proceedings against X makes it clear that, with the DSA, the time of big online platforms behaving like they are 'too big to care' has come to an end" (Supantha Mukherjee 2023). The discussion on regulation involves a nuanced exploration of the multifaceted challenges surrounding hate speech in the digital public space. Striking a balance between preserving free expression and safeguarding citizens requires a thoughtful and collaborative approach involving platforms, policymakers, and the wider society. Continuous dialogue, research, and ethical considerations are crucial to developing effective and fair regulatory frameworks. Scholarly literature can help disentangle these challenges and shed further light on the dynamics in which content moderation influences hate speech. Our research has meaningfully contributed to this through quantitative and qualitative analysis. However, we do want to acknowledge the limitations of our research.

First of all we need to note that from a quantitative perspective the detection and meaningful detection of antisemitic content of tweets poses a reliability challenge in the data. While we employed advanced Natural Language Processing to maximize the level of accuracy, there remains potential for over-or underrepresentation in the dataset, especially considering the cryptic nature of such tweets. Furthermore, the hashtag analysis can in its essence only include tweets that use antisemitic hashtags, omitting potentially antisemitic tweets that did not use hashtags.

Regarding the qualitative analysis it has to be mentioned that the interview selection process favored individuals active in the space and all participants hold very high levels of education. Additionally, due to personal motivations an above-average altercation with the issue can be expected. It would be very interesting to include perspectives of platform users engaging in behavior that we labeled antisemitic, however for obvious reasons such interviews are very complicated to obtain.

While these limitations are acknowledged, they also highlight areas for future research and refinement of study design. By addressing these constraints in subsequent studies, a more nuanced understanding of the relationships between content moderation and hate speech online can be accomplished.

8. Conclusion

In this report we have addressed the research question how Elon Musk's acquisition of Twitter in October 2022 influenced antisemitic hate speech on the platform hypothesizing that it led to an increase in the volume of antisemitic tweets, as well as that variations in content moderation following the acquisition have led to an increase in antisemitic hate speech on X. Using event study regression analysis and qualitative expert interview we can corroborate both hypotheses and confirm that the acquisition of Twitter by Elon Musk caused an increase in antisemitic hate speech on the platform.

These theoretical findings have a real-world impact and raise important questions about the governance of digital public spaces and inform current debates on the role of platforms in providing appropriate content moderation practices to limit hateful content and disinformation for citizens.

We furthermore contributed to the literature by providing a comprehensive dataset containing 158.439 antisemitic tweets which we hope can facilitate further research into this alarming trend of rising antisemitism online. In particular, analysing the dynamics of how changes in content moderation influence user behaviour, and thus hate speech, would greatly advance the understanding of this field.

9. References

- Anti-Defamation League. 2021. "Online Hate and Harassment.: The American Experience 2021." *Center for Technology and Society: New York, NY, USA*, 10–23.
- Arttime, Oriol, Valeria d'Andrea, Riccardo Gallotti, Pier Luigi Sacco, and Manlio De Domenico. 2020. "Effectiveness of Dismantling Strategies on Moderated vs. Unmoderated Online Social Platforms." *Scientific Reports* 10 (1): 14392.
- Benton, Bond, Jin-A Choi, Yi Luo, and Keith Green. 2022. "Hate Speech Spikes on Twitter after Elon Musk Acquires the Platform." *School of Communication and Media, Montclair State University*.
- Berger, John M. 2016. "Nazis vs. ISIS on Twitter: A Comparative Study of White Nationalist and ISIS Online Social Media Networks."
- Dan Milmo. 2022. "How 'Free Speech Absolutist' Elon Musk Would Transform Twitter." *Guardian*, April 2022.
- Feuer, Menachem. 2023. "#BanTheADL Hashtag Only Benefits Antisemites and Jews Who Want to Be Liked by Them." *Jewish Journal*, September 13, 2023.
- Flores, René D. 2017. "Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 Using Twitter Data." *American Journal of Sociology* 123 (2): 333–84.
- Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.
- . 2020. "Content Moderation, AI, and the Question of Scale." *Big Data & Society* 7 (2): 2053951720943234.
- Hickey, Daniel, Matheus Schmitz, Daniel Fessler, Paul E Smaldino, Goran Muric, and Keith Burghardt. 2023. "Auditing Elon Musk's Impact on Hate Speech and Bots." In *Proceedings of the International AAAI Conference on Web and Social Media*, 17:1133–37.
- Horta Ribeiro, Manoel, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. "Do Platform Migrations Compromise Content Moderation? Evidence from r/The_donald and r/Incels." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): 1–24.
- Jikeli, Gunther, and Katharina Soemer. 2022. "Conversations About Jews on Twitter: Recent Developments Since Elon Musk's Takeover." *Computational and Mathematical Organization Theory*.
- . 2023. "The Value of Manual Annotation in Assessing Trends of Hate Speech on Social Media: Was Antisemitism on the Rise during the Tumultuous Weeks of Elon Musk's Twitter Takeover?" *Journal of Computational Social Science*, 1–29.
- Langvardt, Kyle. 2017. "Regulating Online Content Moderation." *Geo. LJ* 106: 1353.
- Miller, Carl, David Weir, Shaun Ring, Oliver Marsh, Chris Inskip, and N P Chavana. 2023. "Antisemitism on Twitter before and after Elon Musk's Acquisition." *Institute for Strategic Dialogue*.
- Rania Aniftos. 2022. "Kanye West Praises Hitler in Alex Jones Interview: 'I Also Love Nazis.'" *Billboard*, January 2022.
- Riedl, Martin J, Katie Joseff, Stu Soorholtz, and Samuel Woolley. 2022. "Platformed Antisemitism on Twitter: Anti-Jewish Rhetoric in Political Discourse Surrounding the 2018 US Midterm Election." *New Media & Society*, 14614448221082122.
- Roberts, Sarah T. 2017. *Content Moderation*.
- Supantha Mukherjee. 2023. "EU Targets Musk's X in First Illegal Content Probe." *Reuters*, December 18, 2023.