POWERED BY

ISD Institute for Strategic Dialogue

CASM technology

Antisemitism on Twitter Before and After Elon Musk's Acquisition

Carl Miller
David Weir
Shaun Ring
Oliver Marsh
Chris Inskip
Nestor Prieto Chavana

With foreword & commentary by
Jacob Davey & Milo Comerford

New research from CASM Technology and ISD has found a major and sustained spike in antisemitic posts on Twitter since the company's takeover by Elon Musk on October 27, 2022. Powered by the award-winning digital analysis technology Beam – and based on a powerful hate speech detection methodology combining over twenty leading machine-learning models – researchers found that the volume of English-language antisemitic Tweets more than doubled in the period following Musk's takeover. In total, analysts detected 325,739 English-language antisemitic Tweets in the 9 months from June 2022 to February 2023, with the weekly average number of antisemitic Tweets increasing by 106% (from 6,204 to 12,762), when comparing the period before and after Musk's acquisition.

Whilst preliminary studies conducted immediately after the takeover pointed to spikes in specific hateful slurs, this research moves beyond keyword-based analysis to demonstrate the broader and longer-term impact that platforms de-prioritising content moderation can have on the spread of online hate. Our approach draws on a suite of natural language processing classifiers trained to identify antisemitic content in line with the IHRA definition, allowing us to identify messages at scale which can plausibly be categorised as hate speech.

Beam: defending information

Beam is a multi-lingual, multi-platform capability to expose, track and confront information threats online, from disinformation to hate, extremism, information operations, harassment and harmful conspiracy theories. It is co-developed by CASM Technology and the Institute for Strategic Dialogue (ISD). In 2021 Beam was the joint-winner of the US-Paris Tech Challenge for innovative approaches to counter disinformation, sponsored by the US State Department; the UK Department of Digital, Culture, Media and Sport; and NATO.

The methods used in this report draw on the ongoing work by CASM Technology and the ISD to measure online harms at scale, across hate, harassment, abuse and disinformation. This paper uses a pioneering approach to the automated detection of antisemitism by deploying an 'ensemble' of different classifiers within a single workflow.

Acknowledgements

The authors would like to thank Milo Comerford, Melanie Smith, Jacob Davey and Jakob Guhl for their feedback and remarks on drafts of this paper, and researchers across the Institute for Strategic Dialogue for their many years of cooperation and support in researching hate speech on social media.

Please note, that in order to preserve the meaning and realities of antisemitism on social media, we have maintained the use of language - epithets, slurs, phraseologies and monikers - that readers may find offensive.

Foreword & Commentary

Jacob Davey & Milo Comerford, ISD

Twitter's tumultuous takeover saw dramatic changes in the platform's approach to tackling online harms. Within days, fundamental changes were made to policies and enforcement, including the reinstatement of accounts previously permanently banned, the dissolution of Twitter's independent Trust and Safety Council responsible for advising on decisions around tackling harmful activity on the platform, and the laying off of over half of Twitter's staff, including many of those responsible for content moderation, online safety and conversational health.

The effect of these changes were reflected in the data analysis outlined in this report, which demonstrates a major increase in the number of antisemitic Tweets posted in the immediate aftermath of the takeover, which has crucially remained at an elevated level in subsequent months.

We also identified a surge in the creation of new accounts posting hate speech which correlated with Musk's takeover. In total 3,855 accounts which posted at least one antisemitic Tweet were created between October 27 and November 6. This represents more than triple the rate of potentially hateful account creation for the equivalent period prior to the takeover. Closer assessment of these accounts showed that many displayed characteristics of overt racism and ethnonationalism. This correlates with a rise in coordinated harassment and even pro-ISIS activity on the platform around Musk's takeover, suggesting that harmful online communities felt empowered by Musk's widely publicized shits to Twitter's management.

Despite Musk's claims that "hate Tweets will be max deboosted & demonetized" - indicating that they will not be algorithmically recommended to users on their news feeds (deboosted) and will not be able to be displayed as adverts or able to generate revenue (demonetized) - and that "New Twitter policy is freedom of speech, but not freedom of reach", the research showed no appreciable change in the average levels of engagement or interaction with antisemitic Tweets before and after the takeover. There is no clear evidence that 'de-boosting' had any impact, as the platform's algorithmic architecture seemingly continues to prioritize engagement over quality content. However, Twitter's lack of algorithmic transparency means it is not easy to test this hypothesis at scale, preventing Musk from being held accountable for his promises.

A new regulatory paradigm

Twitter's policy on hateful conduct claims to prohibit the incitement of harm against people based on race, ethnicity or religious affiliation; the harassment of individuals with reference to the Holocaust; and the use of slurs and racist epithets. However, our research surfaced a broad spectrum of antisemitic content on Twitter ranging from harmful conspiracy theories referring to Jewish control of finance, media and politics; to overt support for antisemitic comments made by public figures such as Kanye West; and the promotion of profoundly racist white supremacy.

Much of this falls in a grey area, where it doesn't contravene legal thresholds of hate speech, but nonetheless likely violates platform terms of service. Twitter purports to take a variety of actions on violating material, including removing content, and down-ranking and de-amplifying Tweets, but there is little clarity around how such platform interventions are enforced.

Significantly, our research did find that after Musk's takeover of the platform around 12% of the plausibly antisemitic messages we identified are now inaccessible on the platform, compared to roughly 6% versus pre-takeover. Whilst there are multiple possibilities for a Tweet not being retrievable, one cause would be the platform's own content moderation practices. However, crucially our research suggests that these moderation efforts are not keeping up with the increased volume of hateful content on the platform, and accordingly are having a limited impact on the increasingly hateful environment on Twitter under Musk, a finding affirmed by recent research from the ADL showing the low removal rate of antisemitic Tweets flagged to the platform.

Beyond a sustained increase in hate speech, and evidence suggesting that other counter-measures to de-boost harmful content are having limited impact, Twitter's commitment to transparency also appears to be moving in the opposite direction, with the platform revoking the free API access that makes a substantial amount of this research possible. This poses the significant risk of limiting the impact of third party efforts to assess the scale of harmful content on the platform, or the impact of their moderation efforts. New regulations incoming from the European Union (in particular the Digital Services Act) will mandate much greater transparency from social media platforms on the actions being undertaken to prevent the proliferation of harmful material online.

The rising threat of antisemitism

These findings come amidst wider concerns around the proliferation of online antisemitism, with weaponised hate manifesting in rising real world violence targeting Jewish communities. In 2021 the ADL tracked the highest number of antisemitic incidents including harassment, vandalism and assaults in the US since they started recording in 1979. This is not just a US phenomenon; in the UK the Community Security Trust recorded a similar spike in this concerning activity, whilst the Interior Ministry of Germany also recorded record highs in antisemitic crimes following the Covid-19 pandemic.

These offline hate incidents should be viewed in the context of surges in online hate, with digital platforms facilitating the radicalisation of individuals towards antisemitic world-views and the mass proliferation of narratives which seek to hold Jews responsible for the world's ills. If we are to limit the spread of antisemitism and other forms of hate it is essential that policy solutions are found to its proliferation online.

This includes emerging regulatory regimes such as the EU's newly introduced Digital Services Act, which seeks to enshrine a systemic approach to platform governance, addressing the platforms' business models and their underpinning algorithmic architectures which promoting hate. Our research suggests that Twitter is failing in their duties under this regime, amid calls from regulators for an increased commitment to meaningful transparency, sophisticated detection and proportionate enforcement by the platform.

Executive Summary

In October 2022, Twitter was acquired by Elon Musk at the head of a consortium of private investors. It precipitated one of the most dramatic shifts in social media's short and tumultuous history of grappling with online harms as, within weeks, previously banned accounts were reinstated, policies were upended, and a significant proportion of Twitter's staff were laid off, including reportedly many of those in the company responsible for online safety.

This paper looks to provide an initial snapshot of how these changes have impacted Twitter, through analyzing the scale of English-language antisemitism on the platform before and after the takeover. The study is neither exhaustive nor definitive, but we hope it is a useful early window into the dynamics of one form of online hate and the responses to it.

- Our approach uses an innovative algorithmic architecture to classify Tweets which could be interpreted as 'plausibly antisemitic', where at least one reasonable interpretation of a message's meaning fell within the International Holocaust Remembrance Alliance's definition of antisemitism. There are inherent challenges in training language models on as nuanced a topic as antisemitism, but this architecture is evaluated to operate with an accuracy of 76%.

- Based on this criteria, between 1 June and 9 February 2023, we identified a total of **325,739 plausibly antisemitic Tweets** sent from **146,516 accounts**.

- Our analysis showed the volume of antisemitic Tweets **more than doubled after Musk's acquisition**. Between June and October 27th, the weekly average of plausibly antisemitic Tweets was 6,204. From October 27th until February 9, the average was 12,762, an **increase of 105%**.

- **We identified a significant surge of new accounts posting plausibly antisemitic content. 3,855** such accounts were created between Oct 27 and Nov 6**, an increase of 223% compared to the 11 days (the equivalent timespan) leading up to Oct 27**.

- Whilst Musk claimed that "hate Tweets will be max deboosted",[1] data showed only **a very small decrease** in the average levels of **engagement or interaction** with antisemitic Tweets before and after the takeover.

- The data used in this analysis was collected in two batches. The initial collection was performed on December 2 2022 and included Tweets posted between June 1 to November 30. A second collection was performed on February 9 2023 that extended this range from November 30 2022 to February 9 2023. All analysis presented in this report refers to the full collection range, with the exception of the enforcement analysis which focuses on the initial date range only.

[1] https://twitter.com/elonmusk/status/1593673339826212864

- Through a combination of topic modelling and manual appraisal, analysts drew the following key themes of antisemitism from the Tweets collected:
    - Conspiracist content, often referring to Jewish control of 'elites', media and politics.
    - Antisemitic attacks on 'Zionist states' - including Israel, but also Ukraine - often tied to the above idea of Jewish 'control' of Western elites.
    - Support for antisemitic comments made by Kanye West.
    - Racialised antisemitism, including white nationalism, nativism and ethno-supremacism.
    - Historical antisemitic tropes, including separating out historical 'races' of 'fake' and 'real' Jews and blaming Jews for the death of Jesus.

Analysis and Implications

Our data presents a clear picture: antisemitism spiked on Twitter during its acquisition by Musk, and has stayed at an elevated level in the months thereafter. Less clear is the enforcement response. We can see an increase in the proportion of antisemitic Tweets that are now unavailable. However it is unclear whether this is due to takedowns by Twitter, or other actions (such as deletion) by the users. Moreover, this increase in takedown rates has not kept up with the increases in absolute volume of antisemitic content. There are also a number of complex measurement effects likely present here, which we detail below.

A number of factors likely came together to produce these results. Musk's Twitter takeover brought questions of platform moderation to global media and public attention, with a change in Twitter's posture potentially encouraging antisemitic actors to join or rejoin Twitter. The acquisition also disrupted Twitter's workforce and operations, including its enforcement teams, with mass lay-offs and resignations. This may have contributed to the public impression that hate speech could be conducted with impunity, and had practical effects in terms of enforcement activity.

As with any study, the results we present must be caveated by the limitations of the methods we use. The study is not comprehensive, the algorithmic ensemble we use to detect antisemitism has a measurable error, applying definitions of antisemitism to the messiness of social media is challenging and the analysis is animated by our own interpretations and judgements. We discuss these more fully in the methods section below.

Part 1. Volume of plausible antisemitism on Twitter

Defining 'plausible antisemitism'

Measuring the amount of antisemitism on Twitter is a formidable task. It requires the sensitive and careful use of concepts and definitions that often collide with the uncertain and messy social realities found online.

The definition of antisemitism used throughout this project is from the International Holocaust Remembrance Alliance (IHRA): "a certain perception of Jews, which may be expressed as hatred towards Jews. Rhetorical and physical manifestations of antisemitism are directed towards Jewish or non-Jewish individuals and/or their property, towards Jewish community institutions and religious facilities." The IHRA working definition includes a number of practical examples, included in the annex.

Applying this definition can be challenging, particularly on a brief and discursive medium like Twitter. where the 'real' meaning of a message can be hard to establish from the individual Tweet alone.

Analysts observed many posts to fall within a 'grey' area where different coders might draw different, equally valid interpretations when trying to discern the real intention or meaning of the message. Sometimes this is due to ambiguous language; sometimes language the analyst felt was deliberately coded in an attempt to hide antisemitic intent; sometimes there was not sufficient context to understand full meaning; and sometimes the message was simply unclear or unintelligible in parts. There were also a number of edge cases, which we discuss more in Part 5 of this report.

Due to this context, analysts could often come to legitimately different views as to the 'real meaning' behind a piece of text. To respond to this challenge, we used a concept of 'plausible antisemitism' where at least one reasonable interpretation of the Tweet was that its meaning fell within the IHRA's definition. As we explain more in the methodology section below, this approach risks classifying some ambiguous texts as antisemitic when they are not. It also, however, limits the capacity of individual analysts' interpretation to skew or affect our findings.

Our technical approach was to combine 22 pre-existing classification models relevant to hate speech detection and five additional lexicons of hateful words into a so-called 'ensemble' of classifiers. These published models were developed with a variety of aims in mind, including the detection of toxicity, threats, and counter-speech, leading to different strengths and weaknesses when it comes to the detection of antisemitism.

Analysts then manually coded 400 Tweets for plausible antisemitism, according to the project's definition (with 213 not antisemitic and 187 antisemitic), which were then used to train a meta-classifier to learn the best combination and patterns of decisions made by the 22 component models that produced the most accurate overall classification.

Volume of antisemitism

We identified a total of 325,739 Tweets as 'plausibly antisemitic', posted between 1 June 2022 and 9 February 2023. Like many other online phenomena, antisemitism tends to be driven by both online and offline events. A peak occurred on August 6, which aligns with a series of rocket attacks and a subsequent ceasefire in Gaza. Multiple peaks occurred after October 9, the day Kanye West's account was suspended after saying he would go 'death con 3 On Jewish people'. The highest peaks culminate on October 25 - in line with reports that Musk would be imminently closing Twitter's acquisition - and November 4.

Towards the end of December, the very high peaks of antisemitic volume ceased. However, the weekly average from December, January and into February remained at 11,359, a stable increase of 97% over the average across July, August and September.

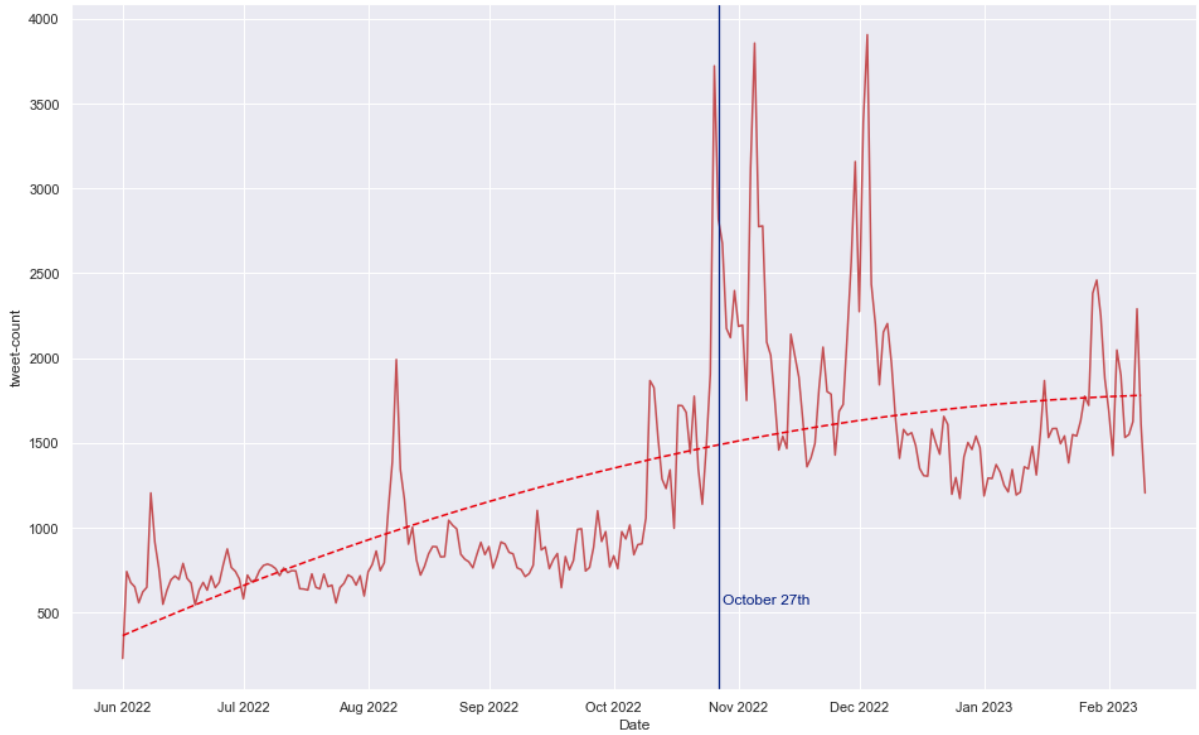

Figure 1: volume of potentially antisemitic Tweets over time, June 2022 – February 2023

From June until the week of October 27 the number of plausibly antisemitic Tweets posted per week stayed generally stable, with an overall average of 6,204. From that week onwards, this average rose to 12,762 per week, an increase of approximately 105% (until the end of the study period).

Month	Average Tweets per week classified as plausibly antisemitic
June	4,670
July	4,900
August	6,680
September	5,930
October	11,470
November	13,850
December	12,530
January	10,890
February	9,380

Table 1: average Tweets classified as plausibly antisemitic by month, June 2022 – February 2023

Part 2. Antisemitic Accounts on Twitter

We identified 146,516 accounts which had sent at least one Tweet classified as plausibly antisemitic. Analysing the creation dates of these accounts indicates a spike on October 28, the day after Musk's acquisition. 3,855 such accounts were created between Oct 27 and Nov 6, the total duration of the spike. This spike potentially correlates with reports of a coordinated trolling campaign designed to flood Twitter with hate speech at the point of Musk's takeover of the platform, although it should be noted that these new users only account for approximately 2.6% of the total number we observed to have sent at least one antisemitic message.[2]

[2] https://www.theguardian.com/technology/2022/oct/30/twitter-trolls-bombard-platform-after-elon-musk-takeover

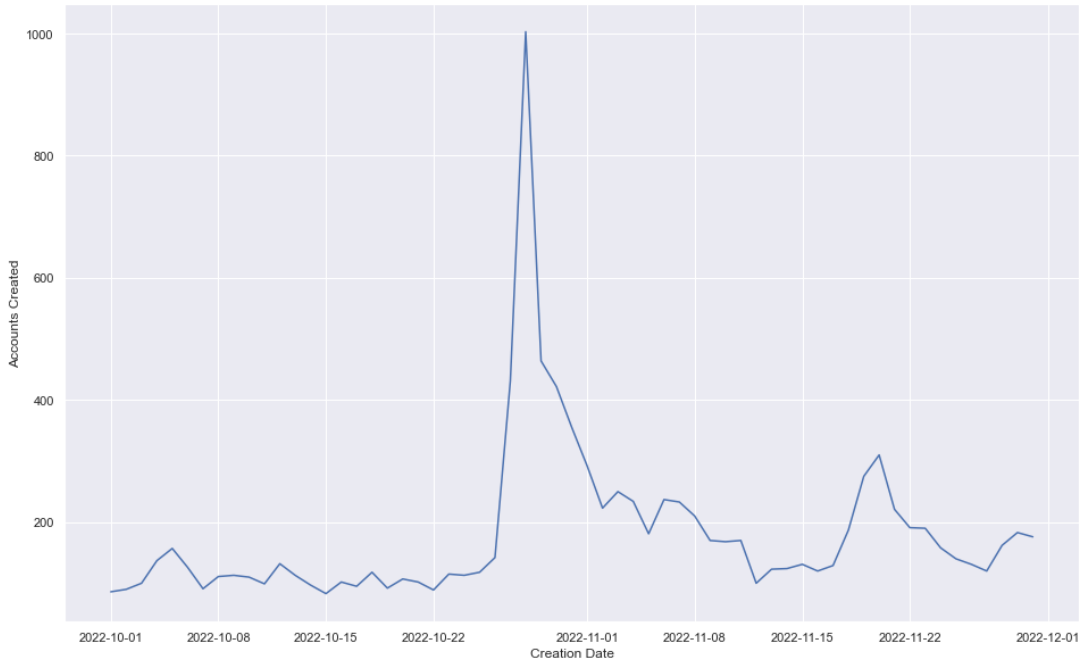

Figure 2: antisemitic account creation by date, October – December 2022

The most active account had sent 436 plausibly antisemitic Tweets, and the second 420, often multiple copies of the same Tweet tagging different people each time. Most accounts, however, sent fewer than 30 antisemitic Tweets. We performed a manual analysis of the 50 most prolific accounts, examining any explicit identifiers they mentioned in their biographies, and the wider behavior they conducted. We qualitatively identified a number of key themes exhibited by many accounts, often occurring in combination with each other:

- **Conspiracy theories.** Many accounts spoke about "noticing things" or "spotting patterns". For them, antisemitism often took the form of referencing Jewish figures (Soros, Rothschild) or claiming that Jews control the media, finance, and/or political parties (particularly the U.S. Democrats). For Soros in particular, a recurring theme was the idea that he has encouraged crime in the US by either sponsoring 'soft on crime' Democrats or directly 'sponsoring criminals'. Some referenced extreme conspiracy theories, such as 'elites' running satanic cults and/or microchipping citizens.

- **Global politics: Israel and Ukraine.** Some of the most active of these accounts were extensively critical of Israel and Zionism, often using the term 'Israhell'. Other geopolitical content also tied together Israel and Ukraine with wider conspiracy theories, suggesting that support for these countries was an identity marker of the 'elites' (sometimes related to perceived Zionist control of elites).

- **Criticism of online harms agenda.** Other accounts emphasised a desire to push freedom of speech, or criticised ideas of 'hate speech' and 'identity politics'. Some biographies included

references such as "Censored for criticizing the Group that uses its disproportionate Power to propagate Hatred against Me" or "I disagree with what you said. Therefore what you said is hate speech." Some of these accounts expressed support for Kanye West and/or praised Musk's takeover of Twitter (though it should be noted that other accounts suggested he may be part of the 'capture' of 'elite media'). A small number of accounts took an approach familiar from sites such as 4chan or 8chan, of using controversial and hateful speech in a semi-humorous or potentially parodic way. For example, one account was a Kanye West parody, which made antisemitic comments against "Jewish business"; others referenced Hitler in their usernames. It should be noted however that the use of 'parody' has long been identified as a tactic used by genuine hate actors to defend their accounts against platform enforcement. Many accounts we saw exhibiting this behaviour have now been suspended by Twitter.

- **Exclusionary and supremacist racial politics.** Some accounts included antisemitism as part of other race-related content; for instance promoting ideas of white people being superior to and/or needing to 'defend themselves' against all other races. There were also references to different historic races / genetic lineages of Jews, referencing narratives that some are 'fake jews'.

Part 3. Themes of Antisemitism

Unsupervised topic modelling is a form of Natural Language Processing that can be used to discover topics or themes in large bodies of text. From iterative exploration using this method, we established 10 topics that captured distinct recurring themes within the Tweets. We (i) explored the words and phrases that the algorithm found to be characteristic of each topic; and (ii) inspected randomly selected example Tweets for each topic, to provide an indicative, non-exhaustive picture of the different manifestations and forms that antisemitism actually took on Twitter. We also explored how the prevalence of these different topics varied over the period of data collection. Please note the exemplar texts presented below have been lightly bowdlerised.

Theme 1: 'Goy'/ 'Goyim'

Originally a Yiddish word for a non-Jewish person, 'Goy' or 'goyim', can be used in antisemitic speech to describe the non-Jewish victims of imagined Jewish plots and conspiracies. By connotation, it also often is used to reflect the contempt the speaker feels Jews hold (either openly or secretly) for non-Jews. There were also examples of counter hate speech, with numerous Tweets in this theme also criticising the antisemitic Goyim Defence League group.

@ADL @kanyewest ADL cries out as they strike Kanye. 'Oy vey goyim don't listen to Kayne!'

[@-tags] @elonmusk The damn Goyim keep noticing.

[@-tag] That's the goyim cant touch me smirk

Theme 2: 'Kike'

Kike is a derogatory slur for a Jewish person, and most of the antisemitic Tweets in this category either used it as an insult against another Twitter user or Jews in general. This theme also picked up more 'noise' than other themes, particularly misspellings of 'like' and references to footballer Kike García.

"[@-tag] THE KIKES AND HILTER MUST BE RELATED!!! THEY HAVE EVERYTHING IN COMMON WHEN IT COMES TO HURTING INNOCENT PEOPLE!!!!!

"if you could call your boss a 'Jew kike' and still keep your job- Then EVERYONE would do it"

A kike spotted [accompanied by an antisemitic cartoon]

Theme 3: 'Soros'

These Tweets were about George Soros and his participation in a secret plot to variously destroy America, Christians, Western morality, or the entire world. Frequently he was accused of either backing Democrats in the US and/or 'funding criminals'. In these accounts, Soros is seen to work in concert with hidden forces, sometimes explicitly referenced as Jews, sometimes in the more coded language of 'globalists', 'puppeteers' and 'financiers', which we have interpreted as plausibly antisemitic.

Josh Shapiro is backed, funded, & endorsed by George Soros the globalist Nazi. Soros installed Shapiro as PA AG & now he's trying to move him into the Governor role. Don't let it happen PA! Don't vote for the Nazi's candidate!!

[@-tags] You fools, all you do is point fingers, learn to do RESEARCH...OPEN your eyes and ears. Beto is a SNAKE bought and paid for by George Soros....Beto is apart of the Elite Cabal....RESEARCH IT,,LEARN

[@-tags] Yeah because it's impossible to overthrow a criminal WEF/Soros-puppet

Theme 4 - 'Zionism'

Antisemitic Tweets that fell within this topic were critical of Zionism to the extent that they denied the legitimate existence of Israel, alleged that the Holocaust was either fabricated or exaggerated, equated Israel with Nazism or referred to global Zionist plots or conspiracies - often, it was claimed, supported by 'liberals' and/or 'The West'.

[@-tag] Yeah cuz Palestinians control the media and the Jews must submit. What an absurd statement. Keep peddling your Zionist propaganda, but fewer people buying it. Jews in the media and holly wood brainwash the gentile goy every day

[@-tag] THE ZIONIST JEWS in Israel(Palestine land) & zionist jews in America and zionist jews in Europe want to take 95 persen all the land of Palestinian and give only 5 persen to Palestine. this typical tricky, greedy, dirty brain zionist jews politics trouble makers in this world

[@-tag] Every country has an army but in Israel it is the opposite: it is an army (criminal and cowardly) that has a country. All Zionist money goes to support this army and to colonize Palestine but this also shows how much the Holocaust is used as a propaganda weapon by the Zionists.

Theme 5 - 'Synagogue of Satan'

Formally a reference to a Biblical line in Revelations, the idea of a 'synagogue of Satan' was used in Tweets to describe Jews, and often to either imply or explicitly argue that they were engaged in the persecution of Christians and Christian ethics. We generally interpreted the use of this descriptor to be inherently antisemitic.

[@-tags] Black people are God's chosen (Deuteronomy 28) and their identity was stolen. The so-called Jews are the synagogue of satan (Revelation 3:9)

[@-tag] The Synagogue of Satan money changers are at the heart of everything that is wrong in the West today! They were not kicked out over 100 countries for no reason.. Research it.

Conservatives will say 'well it's in God's hands now' and do absolutely nothing. Your children & grandchildren are going to suffer due to spineless cowards who allow the synagogue of Satan to flourish in this country

Theme 6 - 'Jewish control of the world'

This theme focussed on Jewish control of politics, the media, business, and finance. We have seen these sentiments suggested in previous themes, but sometimes they were expressed independently of the other themes. There was some counter-speech in this theme, in which people used 'Jewish control' to satirise antisemitic conspiracy theorists.

Having money's not everything, not having it is. Jared Kushner is an example of how the Jewish people have their hand on every single business that controls the world.

The Jews using their control over the central banks to keep society divided, money makes the world go round, and if people are fighting each other their busy to look up and realize why they're even fighting, politicians keep people divided to keep power

[@-tag] ok you obviously a tankie who believe jews have space lazers and control the moon [Counter-speech example]

Theme 7 - Historical, Religious and Racial References

This broad theme contained a range of references to alleged historical events and trends, including Jews killing Jesus, Jews attempting to 'divide Christians', and Biblical prophecies.

The Synagogue of Satan invented racism! Jews use it to divide & conquer Christ's flock! All are equal in His salvation for He loves all, not a chosen few!

[@-tags]You're still waiting on the 12th of Mom which will be the antichrist which is what the Jews will also unify under a New World order I've read the Scriptures I've done my research I know more than you know there's nothing that you can tell me that I don't already understand Jesus.

[@-tag] @kanyewest The Jewish lie that the Romans killed Christ is just another attempt by Jews to distort the Christian Religion.

Theme 8 – Kanye West

Beginning in early October, Kanye West made a number of appearances and interviews. He made a number of antisemitic remarks, including references to the 'Jewish underground media mafia' and 'Jewish business people', and praise for Adolf Hitler. His Twitter account was locked on 9 October for threatening to go "death con 3 On Jewish people". His account was later unlocked and he Tweeted sporadically from 3 November, until a more high-profile return 20 November. He was then banned again by Musk on 2 December for posting a swastika blended with a Star of David. We identified Tweets defending West's antisemitic remarks, or sending additional antisemitic commentary to him, or about him.

[@-tags] I agree with @kanyewest. Insallah @kanyewest can end facist zionist control of America . @elonmusk these Facist zionists will come after you next. We love you @kanyewest

Fake Jews and #ZionistScams are attacking #KanyeWest aka #Ye because he has something superb and magnificent to offer for humanity in general and the #BlackNation in particular. Their attack on Ye is an attack on everyone of you and your future. #StandWithYe #YE24 #KanyeisRight

Now the Jewish Congress asking for Keynes music to be removed. Hmmmmâ€¦â€¦ now you are just proving that Kanye is right and now the Jewish cabal is exposing themselves.

Theme 9 - 'Israhell'

Often more activist in nature, these Tweets almost exclusively criticised Israel. Judgements separating antisemitic and legitimate criticism of Israel were often fine-grained and difficult, and it is likely that this topic over-includes Tweets that are not antisemitic. We tried to make distinctions between legitimate criticism and those Tweets that denied Israel's right of existence, compared Israel to the Nazi regime, or conflated Israel with wider Jewish influence or characteristics.

[@-tag] Israhell & its loyal dog US are the Terrorists!!

Hey #BDS fake Jews @IfNotNowOrg Is your #BDS hatred sponsored by #BDS billionaires from #Gaza

Theme 10 - Russian invasion of Ukraine

This included claims that the Russian invasion of Ukraine was caused by Jews, that Jews have secretly caused the USA to support Ukraine, and criticism of Volodymyr Zelensky as a Jew. Some Tweets also drew links between Ukraine and Israel as alleged 'Zionist Projects'.

Rabbi Larry Fink of BlackRock is buying Ukraine like Rothschild did with Israel since the Orthodox Jews want the Zionist scum out and the zionist jews need to start pandering to the purse strings of Russia and China now

[@-tag] Amerikan's have no knowledge. Zelensky is a Zionist puppet for the globalist. He's not a real president. He was hand picked.

"[@-tags] The Nazi Jew Dwarf who bombs his people and demands $100 billion a year VS Putin who offered free citizenship to all Ukraine, who rebukes the trans monsters and asks for nothing. Hmmn. Which shall I viscerally dislike?

Topics over Time

The graph below shows the volume Tweets were labelled from different topics over time. We observe:

- Peak in 'Israhell' topic in early August, in line with the rocket attacks in Gaza followed by a ceasefire.
- Discussion of Kanye West increased dramatically in late October, in line with him being dropped by multiple brands for his antisemitic comments; in early November, when he was reinstated to Twitter; and in early December, when he was banned again.
- Increased commentary around West was accompanied by increases in the 'Jews controlling the world' topic, as supporters of West claimed his 'silencing' was evidence for antisemitic conspiracy theories.

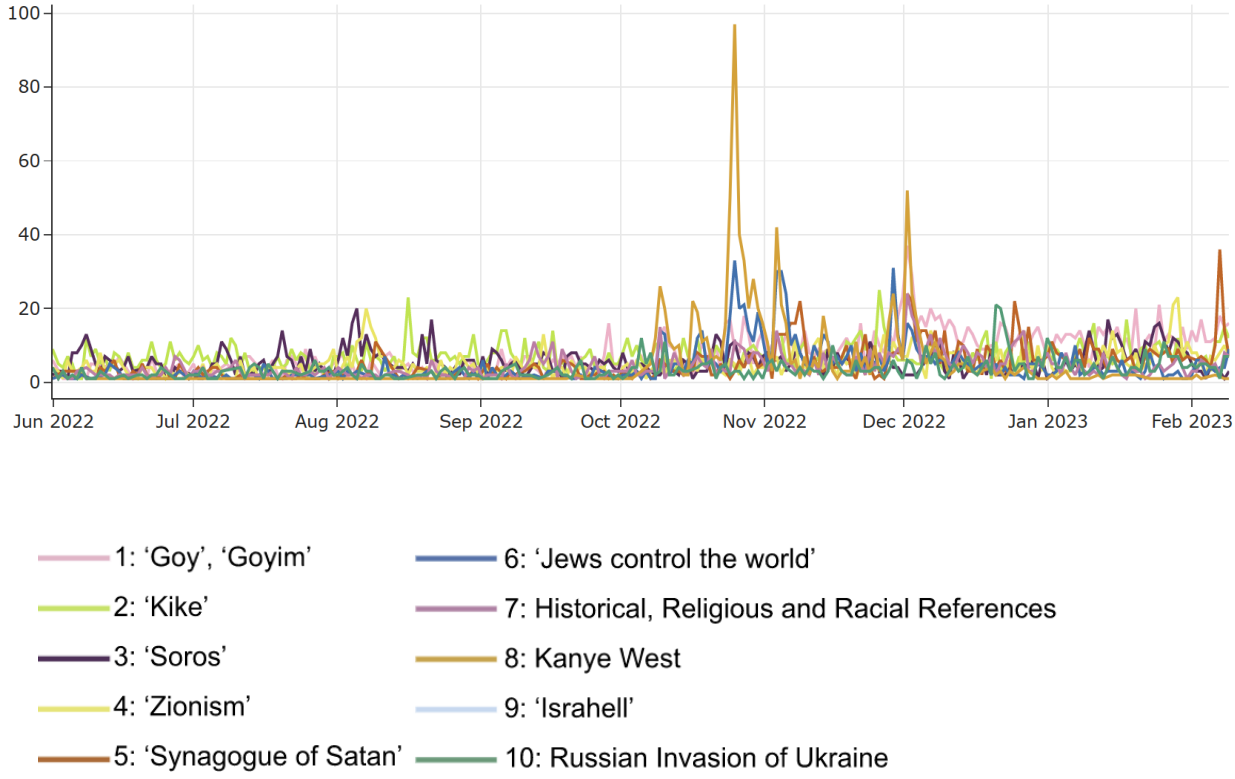

Figure 3: antisemitic Tweets by topic over time, June 2022 – February 2023

Mentions of People & Institutions

To provide an additional view of the type of content being shared, we extracted person and organisational entities identified within each antisemitic Tweet by the Stanford Named Entity Recognition (NER) tool.[3] We ranked these entities by the number of unique antisemitic Tweets that include them.

George Soros is - by far - the most mentioned person within antisemitic Tweets, followed by Jesus, Hitler and Trump. The names reflect the array of different themes we identified, ranging from conspiracy theory, race and politics, to Nazism, Ukraine and Kanye West.

[3] https://nlp.stanford.edu/software/CRF-NER.shtml

Person Entity	Occurrence in antisemitic Tweets
soros	19069
georgesoros	9185
jesus	5929
hitler	5737
biden	3928
trump	3566
obama	2553
mizrahi	2260
ye	2181
kanye	1887
putin	1697
christ	1552
zelensky	1310
billgates	1077
schwab	983

Table 2: person entities mentioned in antisemitic Tweets, by volume

Entities classified as 'Organisations' by Stanford's Entity Recognition tool also reflected the preponderance of conspiracy theories. 'NWO' – an abbreviation of 'new world order' – is a common trope of conspiracy theories describing a secret totalitarian globalist government. Likewise the United Nations, European Union, American political parties, the World Economic Forum, the CIA, FBI and American Israel Public Affairs Committee are all regular targets of 'globalist' conspiracy theories.

Org. Entity	Occurrence in antisemitic Tweets
nwo	3553
un	2787
ppl	1716
nazi	1361

judaism	1286
eu	1167
wef	1165
nazis	1095
gop	1038
congress	1017
cia	1005
nato	939
cdc	899
hamas	872
fbi	846
persen	781
dec	763
rothschild	655

Table 3: organisation entities mentioned in antisemitic Tweets, by volume

Part 4. Possible Platform Enforcement

Twitter has a policy on hateful conduct that states "you may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories."[4] This includes violent threats, wishing for harm on a group of people, and harassing individuals by referring to (amongst other mass murder events) the Holocaust. It also includes the repeated and non-consensual use of slurs and racist epithets.

When behaviour that contravenes this policy is identified by Twitter, they can:

- Down-rank and de-amplify the Tweet in various ways to make it less visible. This includes in replies, in search results, and recommendations.
- Remove the offending Tweets.

[4] This was updated in February 2023 to be: "You may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease."

- Suspend the accounts that sent the Tweets.

It is likely that this policy partly but not completely overlaps with the IHRA's definition of antisemitism which we used for this paper. In this section, we try to measure the possible effect of any enforcement activity.

We initially collected the Tweets studied in this report on December 2, covering a date range from June 1 to November 30. This collection was later extended to include Tweets up to February 9. We then made an attempt re-collect the Tweets we identified as plausibly antisemitic on February 15 2023, a process we call 'recollection'. We can use this process to see how many plausibly antisemitic Tweets are no longer available on the platform, and to look at any changes in the amounts of engagement they've received. This analysis was only conducted for Tweets posted until December 2, because analysing very recent Tweets would have left too little time for any enforcement to have happened, which could have made takedowns appear misleadingly low.

This is certainly an imperfect way of studying platform enforcement, and the results presented here should be read as extremely tentative for the following reasons:

- First, there are some forms of Twitter enforcement that we cannot measure, such as read-only suspensions.
- Second, what we can measure is not necessarily due to platform enforcement. Tweets may have been deleted by their sender rather than Twitter, for instance, and many factors can affect engagement alongside de-ranking.
- Third, we would not have been able to collect Tweets already removed by Twitter before December 2. This is likely to reduce the overall volume of measurable antisemitism and measurable removals of antisemitism for older Tweets.
- Fourth, Tweets sent very close to the collection date would have had less chance to have been taken down than older Tweets.

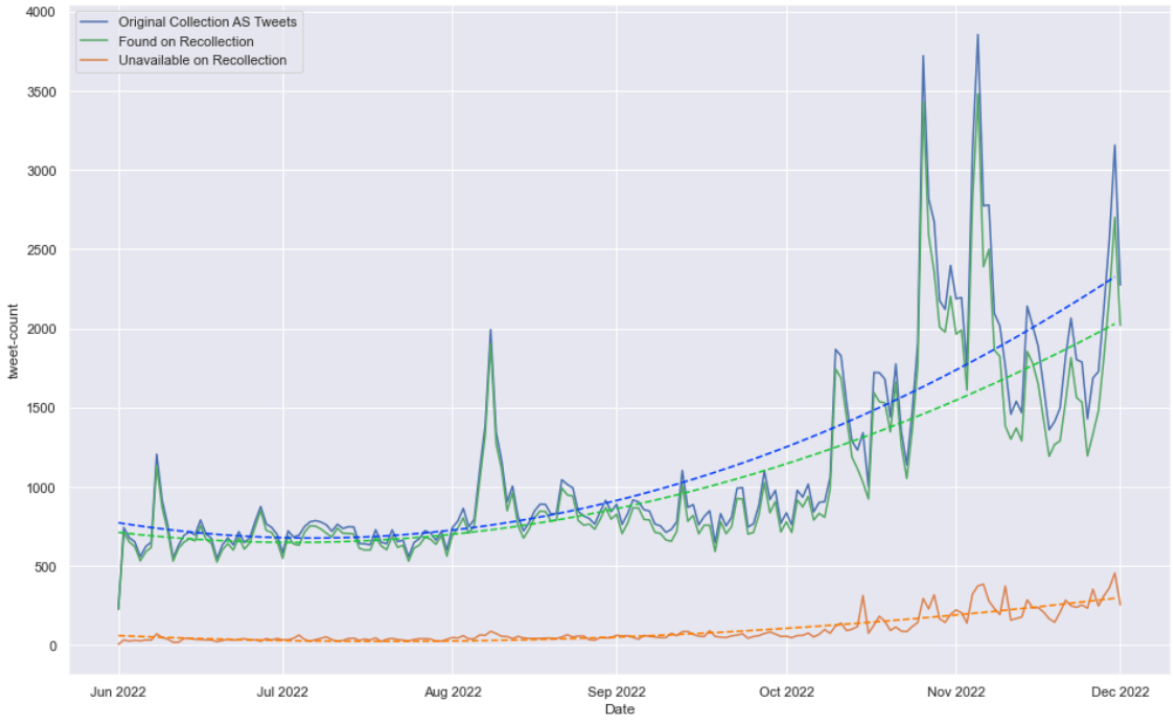

Figure 4: rehydration statistics for antisemitic Tweets over time, June – December 2022

17,589 Tweets we classified as antisemitic in the original collection range were no longer available on the platform, 8.5% of the total.

Overall results	
Antisemitic Tweets in Original Date Range	208,141
Recollected	190,552
Recollected %	91.5%
Unavailable	17,589
Unavailable %	8.5%

Table 4: rehydration statistics for antisemitic Tweets gathered between December 2 and February 15

The further back in time we look, the more time there has been for the Tweet to have already been removed (whether through enforcement or not). This caveat duly noted, we can observe an increase in unavailable Tweets beginning in late October, broadly mirroring the overall increase in the volume of plausibly antisemitic Tweets. However, the rate of increase in potential takedowns is much lower than the increase in antisemitic posts.

The larger absolute quantity of unavailable Tweets in the later months is partly due to simply more Tweets being collected over those months. However, a greater proportion of Tweets were also unavailable, too. As noted above, this can be due to a number of factors.

Recollection results per month					
Month	**Antisemitic Tweets**	**Recollected**	**Recollected (%)**	**Unavailable**	**Unavailable (%)**
June	21,466	20,410	95.1%	1,056	4.9%
July	21,683	20,528	94.7%	1,155	5.3%
August	29,228	27,622	94.5%	1606	5.5%
September	25,410	23,564	92.7%	1846	7.3%
October	49,794	45,611	91.6%	4183	8.4%
November	60,560	52,817	87.2%	7743	12.8%

Table 5: rehydration statistics by month, June – November 2022

As Musk has claimed that part of Twitter's enforcement strategy is "max deboosting hate Tweets" - i.e. reducing how much they are seen and engaged with - we also investigated levels of engagement to test this claim. There was a small decrease in the average engagement metrics for antisemitic Tweets posted before and after October 27. As the table below shows, antisemitic Tweets posted before October 27 received an average of 6.4 'favourites', while those posted after received an average of 6. In the case of ReTweets, Tweets posted before October 27 received an average of 1.2 ReTweets, while those after received 1. Therefore despite Musk's claims of 'max deboosting', our data showed only **a very small decrease** in the average levels of **engagement or interaction** with antisemitic Tweets before and after the takeover..

Engagement metrics before and after October 27 for antisemitic Tweets		
	Before October 27	After October 27
Total Records	136,024	189,715
Favourites Per Tweet (Mean)	6.4	6

Favourites Sum	864,661	1,128,827
ReTweets Per Tweet (Mean)	1.2	1
ReTweets Sum	157,538	188,569

Table 6: engagement metrics for antisemitic Tweets before and after October 27

Part 5. Methodology and Caveats

Measuring the amount of antisemitism on Twitter is a formidable task. It requires the sensitive and careful use of concepts and definitions that often collide with the uncertain and messy social realities found online. It also poses a huge technical challenge to collect and reliably distinguish antisemitic messages from everything else at great scale. Many other forms of speech use language very similar to the antisemitism we were trying to identify, from counter-speech to the the appropriated use of language by targeted communities, and people talking about hate and racism online.

This is a research challenge that forces us to go far beyond simply counting the frequency of certain words or phrases, and instead use machine learning and natural language processing to train models and workflows capable of handling complex and multi-faceted forms of language, meaning and expression. This project falls within a much broader effort by both CASM and ISD to measure hate speech online, and below we describe the method used here, drawn from that broader endeavour.

Step 1. Defining 'plausible antisemitism'

The definition of antisemitism used throughout this project is from the International Holocaust Remembrance Alliance (IHRA): "a certain perception of Jews, which may be expressed as hatred towards Jews. Rhetorical and physical manifestations of antisemitism are directed towards Jewish or non-Jewish individuals and/or their property, towards Jewish community institutions and religious facilities." Referring to the IHRA's definition, this can include:

- Calling for, aiding, or justifying the killing or harming of Jews in the name of a radical ideology or an extremist view of religion.
- Making mendacious, de-humanising, demonising, or stereotypical allegations about Jews as such or the power of Jews as collective — such as, especially but not exclusively, the myth about a world Jewish conspiracy or of Jews controlling the media, economy, government or other societal institutions.

- Accusing Jews as a people of being responsible for real or imagined wrongdoing committed by a single Jewish person or group, or even for acts committed by non-Jews.

- Denying the fact, scope, mechanisms (e.g. gas chambers) or intentionality of the genocide of the Jewish people at the hands of National Socialist Germany and its supporters and accomplices during World War II (the Holocaust).

- Accusing the Jews as a people, or Israel as a state, of inventing or exaggerating the Holocaust.

- Accusing Jewish citizens of being more loyal to Israel, or to the alleged priorities of Jews worldwide, than to the interests of their own nations.

- Denying the Jewish people their right to self-determination, e.g., by claiming that the existence of a State of Israel is a racist endeavor.

- Applying double standards by requiring of it a behavior not expected or demanded of any other democratic nation.

- Using the symbols and images associated with classic antisemitism (e.g., claims of Jews killing Jesus or blood libel) to characterise Israel or Israelis.

- Drawing comparisons of contemporary Israeli policy to that of the Nazis.

- Holding Jews collectively responsible for actions of the state of Israel.

Applying this definition to Twitter was challenging, particularly given the brief and discursive nature of the medium where the true intent of a user can be hard to establish. It can be hard to come to a conclusive, objective decision on *one* piece of content, let alone the thousands used in this analysis.

We observed many posts to fall within a 'grey' area where different coders might draw different, equally valid interpretations when trying to discern the true intention or meaning of the message. This was due sometimes to ambiguous language, sometimes language the analyst felt was deliberately coded, sometimes when there was not sufficient context to understand full meaning, and sometimes to the message being unclear or unintelligible in parts. There were also a number of edge cases where analysts consistently found the definitional boundaries of antisemitism to be the most difficult to apply. These included:

- References to racist tropes that may be genuinely racist or may be sarcastic or parodic attacks as counter-speech. For instance, we saw numerous references to 'jews controlling the world' which, in context, were more likely to be using tropes to insult conspiracy theorists (as one example Congresswoman Marjorie Taylor Greene, known for supporting conspiracy theories, was frequently accused of believing in 'Jewish Space Lasers').

- Criticism of elites, globalists, billionaires and so on sometimes seemed to use coded language for Jews, and on other occasions did not. For instance, some attacks on George Soros are based on antagonism to his support for liberal causes; but others bring in tropes of Jewish elites 'controlling' politicians.

- Criticism of Israel was one of the most difficult judgements analysts had to make, as the distinctions between a criticism we would regard as antisemitic, and one we would regard as not, could be subtle and fine-grained. As discussed in our definition, for criticism to be antisemitic, it needed to go beyond the criticism of a nation state, whether by directly comparing Israel to Nazism, by denying the right of the Jewish state to exist, or by suggesting Western support for Israel emerged from 'Jewish control of elites'.

- Finally, in automated text-based collection and classification, distinctive words which appear in antisemitic contexts - such as 'Nazi' or 'shabbos goyim' - can appear in other contexts in the dataset in a non-antisemitic manner. For instance, 'Nato Nazis' was used a lot in anti-Ukrainian rhetoric, without necessarily clear antisemitic intent.

Step 2 - Collecting plausibly antisemitic Tweets

We collected all acquirable Tweets that contained either (a) any one of 119 slurs, racist epithets, derogatory references for Jews or generally language that highly correlates with antisemitic speech, or (b) that contained a combination of more general language drawn from two further lists of words and phrases. This allowed us to capture a range of linguistic combinations such as 'jews' and 'control'. From this collection criteria we obtained 1,544,142 Tweets from the Full-archive Search API. Tweets we classified as non-English were removed, resulting in a collection of 764,983 Tweets to be classified.

Step 3 - Classification

Our principal technical task was to create a workflow that could automatically classify any Tweet as either plausibly antisemitic or not according to the IHRA's definition, and also measure the accuracy of that classification. Our strategy was to combine a number of pre-existing classification models, creating a so-called 'ensemble' of classifiers. The ensemble is comprised of 22 pre-trained machine learning models and 5 lexicons, which we outline in the annex. These models were developed with a variety of aims in mind, including the detection of hateful speech towards a single target group, detection of hate targeting any one of multiple target groups, as well as those that aim to detect toxicity, threats, and counter-speech. This approach exploits the fact that different models typically have different strengths and weaknesses.

We manually coded 400 Tweets (213 not antisemitic and 187 antisemitic) as to whether they were plausibly antisemitic, according to the project's definition. These were used to train a meta-classifier (called an XGBoost classifier) that learns how best to combine the decisions of each of the

component models (classifiers) in order to produce a more accurate overall classification of whether any given Tweet is plausibly antisemitic.

Step 4 - Evaluation

To evaluate the overall classification process for the data we collected, we took a further sample of randomly selected Tweets and manually annotated them as either antisemitic or not until we had roughly 100 in each category. We then compared these human decisions to those of the algorithm. When they were the same, we regarded the algorithm to have got the decision correct, and when they were different we regarded the algorithm to have made an error. On the basis of this comparison, we established the following:

- The algorithm's precision. This is the proportion of those Tweets that the model classified as either antisemitic or not antisemitic that were judged to be the same by a human.
- The algorithm's recall. This is the proportion of those Tweets considered to be either antisemitic or not by a human that the model classified as the same.
- F1 score. The geometric average of the precision and recall scores.
- The accuracy of the classifier (the percentage of the Tweets that were correctly classified).

Classifier outcome	**Precision**	**Recall**	**F1 score**
'Not antisemitic'	0.79	0.75	0.77
'Antisemitic'	0.72	0.76	0.74
F1 Score	0.76	0.76	0.76
Accuracy			75.5%

Table 7: classifier outcomes for the ensemble hate speech classifier

Step 5 - Topic Clustering

Unsupervised topic modelling is a form of Natural Language Processing that can be used to discover topics or themes in large bodies of text. We used the BERTopic Python module created by Maarten Grootendorst, which is one of the most advanced open-source techniques available. BERTopic makes use of Bidirectional Encoder Representations from Transformers (BERT) which is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google, and which can be used to create a vector representation of a document that captures aspects of the meaning of the document. BERTopic clusters documents together that have similar BERT encodings, and identifies clusters of words that characterise each of these clusters (topics)

1. Our full dataset (from June to February) was filtered to include just Tweets classified as plausibly antisemitic by the classifier model.

2. Since BERTopic is resource intensive, we applied it to a 20% sample that was created by ordering the Tweets by date/time, and selecting every 5th Tweet in this sequence. We sampled across time to ensure we captured a spread of topics as they grew and waned in prominence.

3. We ran BERTopic, and experimented with a number of output topics to best capture relevant, distinctive themes. We found that 10 topics optimally reduced the number of overlapping topics, while still capturing a diversity of themes.

4. We inspected the word lists of 10 characteristic words/bigrams/trigrams for each topic – listed below - and manually inspected a random sample of 30 Tweets classified with each topic, to interpret and apply a label to each topic; and plotted volume over time using BERTopic's visualisation commands (which are built on Plotly).

A limitation of this form of topic modelling is that a small number of words and phrases are used to represent topics spanning a wide variety of situations, and for topics which appear very frequently it may only be feasible to manually inspect a small proportion of the Tweets in that topic. Nonetheless, the word combinations plus our inspections revealed recognisable repeated patterns which allowed us to develop broad and descriptive labels for topics.

Step 6 - Recollection

The idea of recollection was to measure how many Tweets were still on the platform, by trying to re-collect all the Tweets we classified as antisemitic. The original data collection was performed on December 2nd 2022, searching for Tweets posted between June 1st and November 30th that matched the set of antisemitic keywords described above. The recollection itself was performed on February 15th 2023. Tweets were aggregated by date, according to their posting time in PST. The inability to recollect some Tweets can be attributed to multiple reasons. The Tweet may have been flagged for inappropriate content and taken down. It's also possible that the posting account itself

may have been flagged for posting inappropriate content and suspended, making all their Tweets unavailable. Finally, the account may have been set to private or deactivated by the owner themselves rather than by Twitter.

It is also possible that moderation may have already been applied before the date of the original collection in December. This could explain the lower rate of unavailable Tweets (potential takedowns) on earlier dates, identified in the recollection process. Thus, the increased proportion of unavailable Tweets in more recent dates could be a reflection of the usual attrition rate of Tweets rather than an increase in moderation effort. Additionally, Tweets being unavailable for recollection does not always indicate a moderation effort, as Tweets removed by the poster would suffer from the same problem.

Limitations/Caveats

As with any methodology, the approach used here carries with it a series of strengths and weaknesses. When interpreting the data, the following caveats should be regarded. We discuss many of the most important challenges faced by this research process throughout the methodology section. The most important are collected here, and the results must be interpreted with these challenges in the background.

- **Applying definitions of antisemitism to Twitter data was challenging.** We observed many posts to fall within a 'grey' area where different coders might draw different, equally valid interpretations. This causes an issue when making a binary classification of antisemitic or not, as both training and evaluation data can represent a high degree of analyst bias. Our response was to continue to discuss edge cases and examples between the analysts in order to evolve how we applied the definitions to real world data.

- **Tweets are taken out of context for the purposes of analysis, making some difficult to interpret.** When analysing hate speech, context matters. Given the way the models were trained, however, analysts needed to make decisions on the basis of the Tweet alone, and to ignore the identity of its sender and the broader conversation it may have occurred within. Equally importantly, automated detection approaches such as the one used for the analysis in this report only make their decisions on the linguistic content of the message itself rather than the entire conversational thread from which the message was possibly drawn.

- **The machine learning ensemble does not perform with perfect accuracy.** The classification technology we used is inherently probabilistic. According to our evaluation, it is likely to make the correct classification decision around 75.5% of the time. Analysts observed a number of mistakes that the classification made:
    - The classification of counter-speech to antisemitism as antisemitism
    - The classification of non-antisemitic criticism of Israel as antisemitism
    - The classification of people quoting antisemitic speech as antisemitism.

- **This study is not comprehensive.** We did our best to identify as many plausibly antisemitic Tweets as we could. However it is possible that the inclusion of additional collection keywords may have increased the total volumes we identified. When doing a study like this, there is a trade-off to be made when deciding how to collect the dataset being studied. On the one hand there is good reason to aim for high coverage of potentially antisemitic posts, but this typically leads to the collection of a dataset that contains a very high proportion of posts that are not antisemitic, resulting in a dataset that is very challenging for machine learning algorithms. On the other hand, there are significant benefits in being more selective in terms of the data that is collected, as this results in a dataset with a higher proportion of Tweets that are antisemitic: a dataset, therefore, that can be more accurately classified by a machine learning model.

- **We will not have collected Tweets removed between June 1 and December 2.** This is because we conducted a data collection on December 2, collecting available Tweets sent any time from June 1 onwards. However, if Tweets had been removed before December 2, they would not have been available for our collection. This effect will likely skew the volumes of antisemitism to appear smaller during the older parts of the study, because the older the Tweet is, the more likely it will have been to have been subject to either user action (deletion) or platform moderation (take-down) before we had a chance to collect it.

- **The descriptions of the themes are impressionistic.** Other researchers may have drawn different contrasts or similarities from an appraisal of accounts in this network, or may have placed emphasis on different places.

- **The recollection exercise is certainly an imperfect proxy for platform enforcement.** There are two elements to this measurement where bias can be introduced.
    - For Tweets posted before December 2: the longer the period of time between when a Tweet was posted and our collection point (2 December), the longer time it had to be removed, and so not be present in the study in any form (as stated above). This will likely produce the effect where takedowns appear to be higher the more recent the time window. In other words, as the original collection was done in December, some Tweets posted in June that could have matched the keyword search would have already been taken down. This could explain the lower rate of unavailable Tweets (potential takedowns) on earlier dates, identified in the recollection process.
    - For Tweets posted after December 2: the older the Tweet is, the longer it will have to have been initially collected and then removed. This may mean that, for Tweets posted after December 2nd, this effect may decrease the relative volume of the most recent moderation efforts vis a vis older moderation efforts.
    - Additionally, Tweets being unavailable for recollection does not always indicate a moderation effort, as Tweets removed by the poster would suffer from the same effects.

Part 6. Technical Annex

The Antisemitism classification Ensemble

Layer 1
The ensemble consists of 22 pre-trained models and 5 lexicons (see below) which every Tweet is classified with. Each provides one or more signals such as the presence of toxic language or a particular slur term in a sequence of text. The text of each Tweet is processed with the ensemble which annotates the Tweet with a total of 69 ensemble signals.

Layer 2
The signals extracted for each Tweet are passed through a XGBoost classifier which has been trained to determine whether the overall combination of ensemble signals is likely to indicate that the content is antisemitic according to the IHRA's definition. The classifier was trained on 400 manually annotated Tweets, 213 not antisemitic and 187 antisemitic. The classifier was evaluated on 200 separate, manually annotated Tweets (109 not antisemitic and 91 antisemitic).

Models/ Lexicons used in the ensemble

Hatebert. This is a model trained using a transformer-based machine learning technique called Bidirectional Encoder Representations from Transformers or BERT. It is trained on a large dataset from Reddit (called RAL-E) of comments banned for being offensive, abusive or hateful[5]. It determines whether a post is hateful or not. Subset models include **Hateabuse** based on the Hatebert approach above, but instead is trained to identify abusive posts. **Hateoffence** is also based on the Hatebert approach above, but instead is trained to identify offensive posts. **Hateval** is based on the Hatebert approach above, but instead is trained to identify hateful posts.

Dehatebert. This was an attempt to detect hateful speech in 9 languages across 16 different sources. It was a comparison of different approaches in different languages[6]. **Mono** is a version of Dehatebert to identify hateful posts.

HateXplain was an attempt at automated hate speech detection, also to identify the target community and identify what study calls the 'rationales'; the portion of the post on which the labelling decision most depended. This is intended to increase the interpretability of the model[7]. **Rational2** determines if a post is abusive or not, whilst hate-explain-bert-base-uncased determines if a post is hate, offensive or neither.

---

[5] https://arxiv.org/abs/2010.12472

[6] https://arxiv.org/pdf/2004.06465.pdf

[7] https://arxiv.org/abs/2012.10289

Detoxify. These are a set of models that provide a score on how likely a post is to contain certain 'toxic' traits[8]. The **Original[9], Unbiased[10]** and **Multilingual[11]** models each give each post a score on the following attributes:

- Toxicity
- Severe toxicity
- Obscene language
- Threatening language
- Insults
- Identity attack
- Sexually explicit language (in the case of the latter two).

Hate alert-counter. These models focus on counter-speech,[12] language that is calling out or undermining, opposing or mocking hateful speech in some way. The models usually classify these as hateful speech, so these models are useful to increase the precision of the hybrid ensemble but removing counter-speech as examples of false positives. **Binary** identifies if a post is counter speech or not. Multi-label identifies what kind of counter-speech is being used, including:

- Presenting facts
- Hypocrisy or contradiction
- Warning of consequences
- Showing affiliation with the group
- Denouncing the hate speech
- Humour
- Posts that have a positive tone
- Posts that are hostile to the hate speech poster

A series of additional models also identify counter-speech specific to posts targeting Black, Jewish and LGBT communities.

Perspective. These are a series of models that can be accessed via an API on the Google Cloud Platform.[13] Originally created to help moderators moderate online conversations, they use finely tuned multi-lingual BERT-based models distilled into single-language Convolutional Neural Networks. These models are then used to evaluate the probability of a comment having an attribute of toxicity. Perspective evaluates the following attributes:

- Toxicity
- Severe toxicity

---

[8] https://github.com/unitaryai/detoxify

[9] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[10] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

[11] https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification

[12] https://github.com/hate-alert/Countering_Hate_Speech_ICWSM2019

[13] https://perspectiveapi.com/

- Identity attack
- Insult
- Profanity
- Threat

With more experimental models also classifying for:
- Sexually explicit
- Flirtation

It is important to note the probability scores from these models do not correlate to the severity of the toxicity, just the likelihood of the comment being toxic.

HateALERT-EVALITA. These are a series of models trained for 'Automatic Misogyny Identification' (AMI), which won a prize at EVALITA2018, a period campaign to assess the performance of NLP tools.[14] This includes an overall decision about whether a post is misogynistic, whether the post targets an individual or a more general group, and the type of misogyny being expressed, covering:
- Discrediting
- Derailing
- Dominance
- Sexual harassment
- Stereotype

Hatesonar. An approach that used crowd-sourcing to train models to distinguish between hateful and other instances of offensive language.[15]

Also included were 4 models that have been trained by CASM technology;

roberta-hate. A fine tuned roberta model that was trained on data collected from 4chan, Facebook, Instagram, Twitter, Reddit and YouTube (from known hateful actors or message boards) then annotated in line of whether the post was hateful or not.

roberta-offense. A fine tuned roberta model that was trained on data collected from 4chan, Facebook, Instagram, Twitter, Reddit and YouTube (from known hateful actors or message boards) then annotated whether a post was offensive or not offensive.

Antisemitism-1. A fine tuned roberta model trained on separate data annotated from this dataset consisting of 198/122 not antisemitic/antisemitic Tweets and tested on 49/31 not antisemitic/ antisemitic Tweets. This model had a precision of 0.7 and recall of 0.61 on this dataset.

Antisemitism-2. A fine tuned roberta model trained on separate data annotated from this dataset consisting of 198/122 not antisemitic/antisemitic Tweets and tested on 49/31 not antisemitic/ antisemitic Tweets. This model had a precision of 0.66 and recall of 0.68 on this dataset.

---

[14] https://arxiv.org/pdf/1812.06700.pdf

[15] https://arxiv.org/pdf/1703.04009.pdf

Lexicons

In addition to the models described above, messages can also be analysed more simply by whether or not they contain a given word. First, several externally compiled corpora have been identified.

T-davidson. 178 words that are commonly used in hate speech- manually curated. Each has a score of how likely the post is to be hate speech when the phrase is included[16].

Hatebegets-hate. A list of 187 offensive terms that are used against different groups of people commonly in hate speech posts[17].

Spread_Hate_Speech_WebSci19. A list of 81 offensive terms commonly present in hate speech[18].

A lexicon of generic terms relating to Judaism, and slurs towards Jewish people was also created. They were incorporated within the ensemble in the same way as the machine learning models, by annotating Tweets which contained any one of the words contained within any of the lexicons. This was another way of generating signal that the meta-classifier could use.

[16] https://github.com/t-davidson/hate-speech-and-offensive-language

[17] https://arxiv.org/abs/1909.10966

[18] https://arxiv.org/abs/1812.01693