CyberWell

DATA INSIGHTS: The State of Antisemitism on Twitter

CyberWell is the first ever data platform of online antisemitic content, collected, vetted and curated with the goal of driving enforcement and improvement of community standards and hate speech policy across the digital space. CyberWell was created to leverage data to hold social media platforms accountable for the antisemitism they host and fight for a safer digital future for all. By showcasing antisemitic content on an open platform, we are democratizing the raw data of online antisemitism so that non-profits, digital rights researchers, lawmakers, educators and journalists can propel their own activism, policy-making and research forward.

Recently the Adopt IHRA Coalition - a group of over 180 organizations that previously penned an open letter to Facebook urging the adoption of the International Holocaust Remembrance Alliance's IHRA working definition on antisemitism as a cornerstone of their community standards – utilized CyberWell's database to **highlight the state of antisemitism on Twitter** and similarly **call for the platform's full adoption of the working definition** as part of their hate speech policy. CyberWell uses the widely supported working definition to categorize offending content into specific types of antisemitism, and we too support the call for Twitter to adopt the definition as part of its hate speech policy.

But, what does that mean practically? How do you translate the IHRA working definition into hate speech compliance on a major social media platform? As an addendum to the Adopt IHRA Coalition's efforts, we put together this brief methodology and data insights report to shed more light on the details of what CyberWell's initial monitoring of Twitter revealed about the state of antisemitism on the platform and the overarching hate speech policy gaps.

METHODOLOGY

As a regular part of our working methodology, CyberWell uses a combination of broad search keywords (comprised of general terms, slang, and code words for Jew-hatred), relevant images, and videos to identify a pool of content that is likely to include antisemitic material. We then apply our own exclusive specialized dictionary of antisemitic terms and expressions, based on the IHRA working definition, to flag highly likely antisemitic content. Each piece of content is reviewed by at least two professional analysts who are trained in the fields of antisemitism, linguistics and digital policy. All confirmed antisemitic content is reported to the relevant social media platform, in this case Twitter, and monitored to track how long it takes the platform to remove the content.

The specific dataset cited in the call for Twitter to adopt the IHRA working definition examined antisemitic content published on Twitter dating back to January 2020 through September 2022 in English and Arabic. For the relevant test period, the specialized dictionary **flagged nearly 40,000 Tweets that had a high likelihood of being consistent with one of the eleven criteria of Jew-hatred** laid out in the working definition. CyberWell's analysts reviewed a sample of 2,810 Tweets over the test period and confirmed 1,079 antisemitic Tweets. Flagged antisemitic Tweets were also reviewed to determine whether they complied with Twitter's Rules & Policies, and if not, which specific policy section was violated. This innovative process is unique to CyberWell, combining antisemitism monitoring with compliance to bridge the gap between digital policy and platform-implementation rules by generating data on specific policy failures.

CyberWell's technology further analyzed the confirmed antisemitic Tweets. Documenting the geo-location of the Tweet,[1] the repeating themes that go against general digital hate speech policy best practice ("Policy Violation Themes"),[2] and giving each vetted Tweet an engagement and potential reach score. The engagement score is an estimated interaction grade given to each antisemitic Tweet, based on retweets, likes and comments. All antisemitic Tweets also received a potential reach score which is calculated based on the number of followers of the publishing account.

As of the writing of this report, since CyberWell began monitoring for online antisemitism in May 2022, there a total of 1,898 antisemitic Tweets that were collected, vetted and reported to Twitter. While this sample may be small, it is indicative of our capacity as a start-up non-profit supported by philanthropic donations – CyberWell's technology and bundle of open-source intelligence tools, coupled with our IHRA specialized lexicon have the capacity to detect tens of thousands of pieces of antisemitic content. The full Twitter dataset from the relevant testing period is available here.

[1] Most social media content and accounts are not geo-located. While the CyberWell provides the open-source data on the geo-location of vetted antisemitic content, it is important to acknowledge the limitations of the reliability of geo-located data on social media platforms in general.

[2] For more information, please refer to CyberWell's Policy Guidelines available at: cyberwell.org/how-it-works/policy-guidelines/

DATA INSIGHTS VISUALIZED

Platform: Twitter **Test period:** January 2020 – September 2022

40,000
Tweets flagged as high potential for being antisemitic

→

2,810
Sampled Tweets reviewed for Adopt IHRA Coalition

→

1,079
Confirmed antisemitic Tweets

Engagement Score and Potential Reach

35,000
Total engagement score of entire data set

+7 million
Total potential reach of entire data set

Top Violated Twitter Rule: Hateful Conduct

84%
violated Twitter's rules

21%
Rate of removal **before** CyberWell reporting

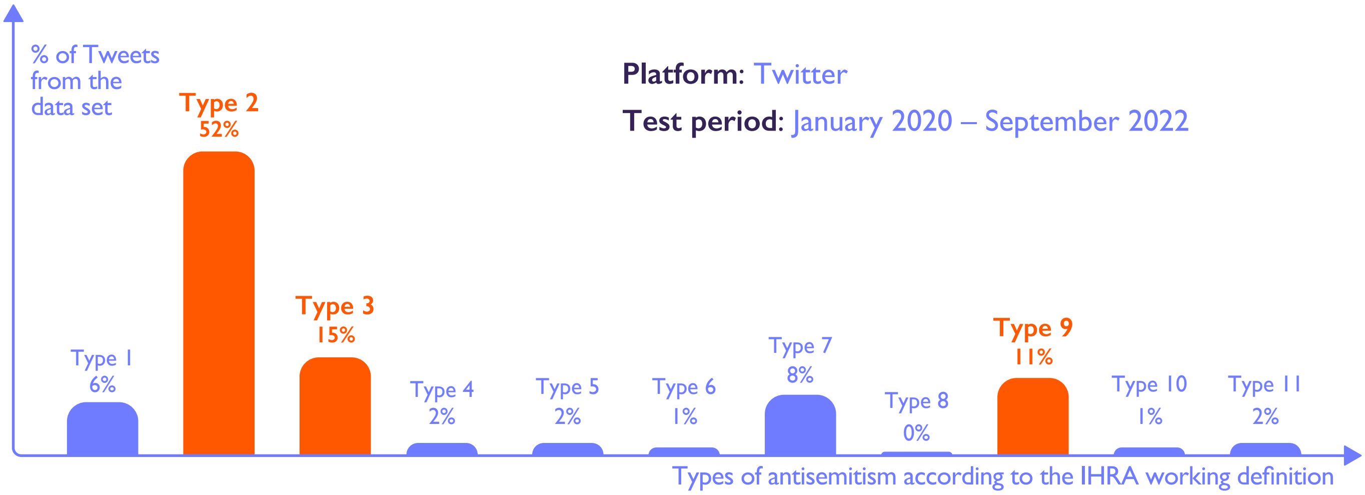
35%
Rate of removal **after** CyberWell reporting

Twitter's Hateful Conduct policy is consistent with a larger theme that is recognized & prohibited by most social media platforms' hate speech policy, Dehumanizing & Stereotypical Hate Content - Content that attacks or dehumanizes individuals or groups based on their protected characteristics. Prohibited content in this category includes speech or imagery in the form of generalizations or comparison to unqualified behavioral statements and/or stereotypes (i.e. tropes). Digital platforms have policies against this content because it can embrace a hateful ideology, inspire hatred or fear of said groups and even incite violent acts against that group or its members. Some platforms, including Twitter, address Holocaust denial through this community standard.

Full Dataset: App.CyberWell.org/twitter.php

IHRA INSIGHTS

Top forms of Jew-hatred in the dataset based on the working definition. It is worth noting, an antisemitic Tweet can be consistent with one or more forms of antisemitism as described in the working definition.



Platform: Twitter

Test period: January 2020 – September 2022

52%

of the Tweets are classically antisemitic, as outlined by the second criterion of the working definition (Type 2)

Criterion of the IHRA working definition: Type 2

“Making mendacious, dehumanizing, demonizing, 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.”

15%

of the Tweets are consistent with the third criterion of the working definition (Type 3)

Criterion of the IHRA working definition: Type 3

“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.”

11%

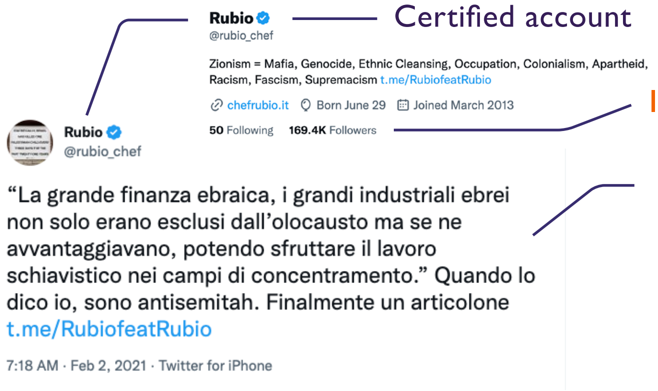
of the Tweets are consistent with the ninth criterion of the working definition (Type 9)

Criterion of the IHRA working definition: Type 9

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

Full Dataset: App.CyberWell.org/twitter.php

IHRA INSIGHTS - EXAMPLES



Rubio ✓
@rubio_chef

Zionism = Mafia, Genocide, Ethnic Cleansing, Occupation, Colonialism, Apartheid, Racism, Fascism, Supremacism t.me/RubiofeatRubio

chefrubio.it · Born June 29 · Joined March 2013

50 Following **169.4K** Followers

Rubio ✓
@rubio_chef

“La grande finanza ebraica, i grandi industriali ebrei non solo erano esclusi dall’olocausto ma se ne avvantaggiavano, potendo sfruttare il lavoro schiavistico nei campi di concentramento.” Quando lo dico io, sono antisemitah. Finalmente un articolone t.me/RubiofeatRubio

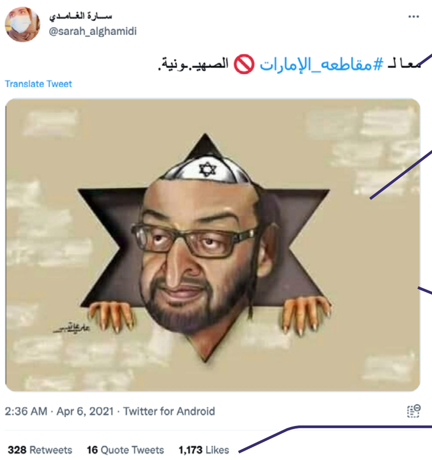
7:18 AM · Feb 2, 2021 · Twitter for iPhone

Certified account

169,4K Followers

“ "The great Jewish finance, the great Jewish industrialists were not only excluded from the holocaust but they took advantage of it, being able to exploit slave labor in the concentration camps". When I say it, I am anti-Semitic. Finally an article.”

Antisemitism Type 2



سـارة الغـامـدي
@sarah_alghamidi

معـا لـ #مقاطعه_الإمارات 🚫 الصهيـ.ـونية.

Translate Tweet

2:36 AM · Apr 6, 2021 · Twitter for Android

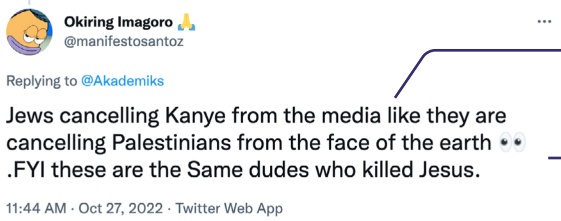
328 Retweets **16** Quote Tweets **1,173** Likes

“We stand with the boycott against the Zionist Emirates [UAE]”

Image of a stereopypical grotesque and evil Jew coming out of a Star of David, potentially, in context the "Jewifying" of a UAE leader after the Abraham accords.

Antisemitism Types 2 and 9

328 Retweets
1,173 Likes



Okiring Imagoro 🙏
@manifestosantoz

Replying to @Akademiks

Jews cancelling Kanye from the media like they are cancelling Palestinians from the face of the earth 👀 .FYI these are the Same dudes who killed Jesus.

11:44 AM · Oct 27, 2022 · Twitter Web App

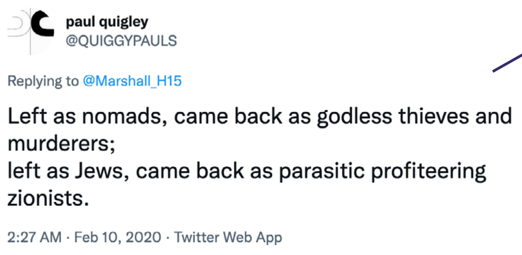
Kanye West related antisemitic Tweets

Antisemitism Type 3

Recent example from CyberWell database

Full Dataset: App.CyberWell.org/twitter.php

IHRA INSIGHTS - EXAMPLES



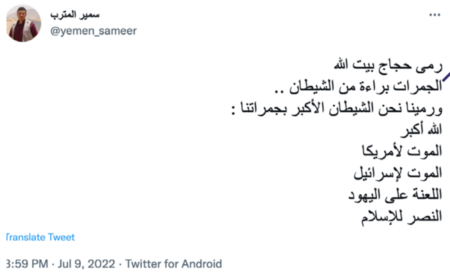
paul quigley
@QUIGGYPAULS

Replying to @Marshall_H15

Left as nomads, came back as godless thieves and murderers;
left as Jews, came back as parasitic profiteering zionists.

2:27 AM · Feb 10, 2020 · Twitter Web App

Antisemitism Types 2 and 9



سمير المترب
@yemen_sameer

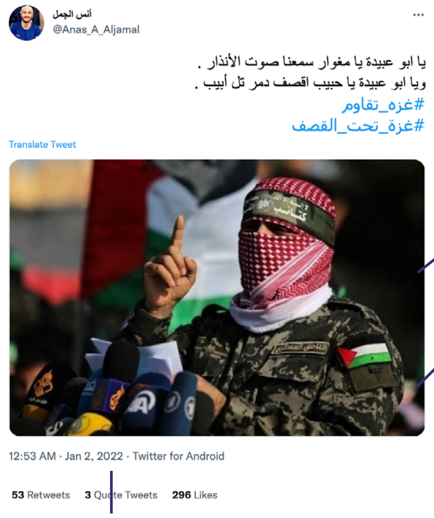
رمى حجاج بيت الله
الجمرات براءة من الشيطان ..
ورمينا نحن الشيطان الأكبر بجمراتنا :
الله أكبر
الموت لأمريكا
الموت لإسرائيل
اللعنة على اليهود
النصر للإسلام

Translate Tweet

3:59 PM · Jul 9, 2022 · Twitter for Android

Slogan of the Houthi movement in Yemen.
"The Stoning by the pilgrims in the House of God
Stoning of the Devil is a purification from the Devil
And we stone the biggest Devil:
Allah is the greatest
Death to America
Death to Israel
Damn / Curse the Jews
Victory to Islam"

Antisemitism Types 2 and 9



أنس الجمل
@Anas_A_Aljamal

يا ابو عبيدة يا مغوار سمعنا صوت الأنذار .
ويا ابو عبيدة يا حبيب اقصف دمر تل أبيب .
#غزه_تقاوم
#غزة_تحت_القصف

Translate Tweet

12:53 AM · Jan 2, 2022 · Twitter for Android

53 Retweets 3 Quote Tweets 296 Likes

Tweet from this year

"Abu Obeida the commando
We heard the alert Abu Obeida the beloved,
bomb and destroy Tel Aviv"

Image of the spokesmen of Al-Qassam brigades, the military wing of Hamas

Antisemitism Type 1

Criterion of the IHRA working definition

"Calling for, aiding, or justifying the killing or harming of Jews in the name of a radical ideology or an extremist view of religion."

Full Dataset: App.CyberWell.org/twitter.php