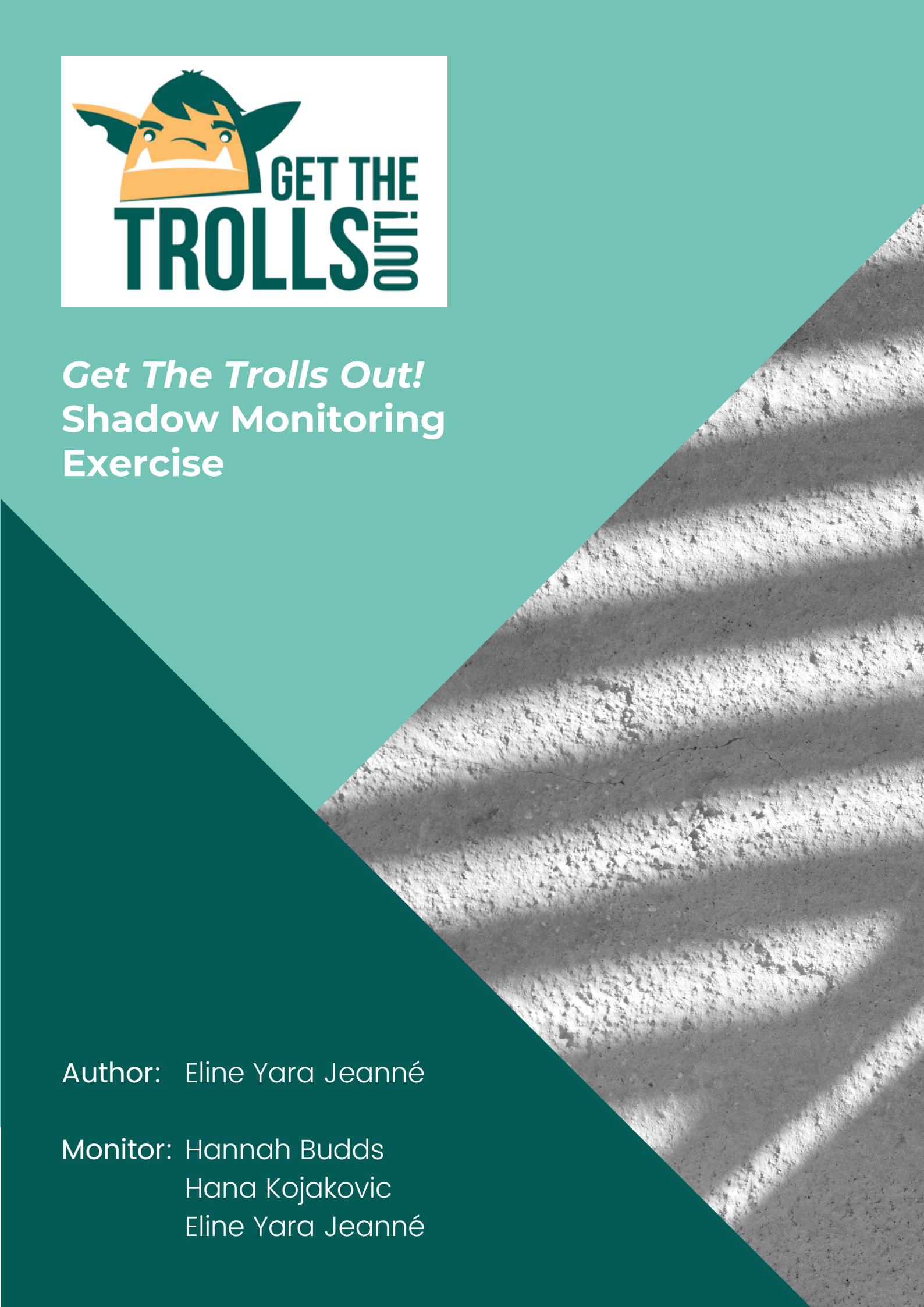



# *Get The Trolls Out!* Shadow Monitoring Exercise

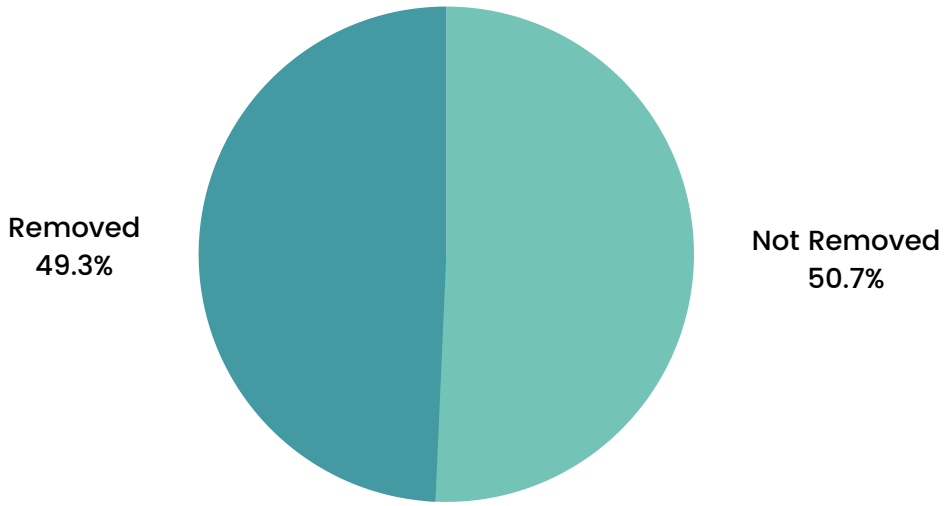
Author: Eline Yara Jeanné

Monitor: Hannah Budds
Hana Kojakovic
Eline Yara Jeanné

From February 15th – 28th a team of 3 monitors from the Media Diversity Institute, as part of the Get The Trolls Out! project, monitored a select few social media platforms for religious hate speech, focusing on calls for violence. The aim of the exercise was two-fold: to monitor the amount and type of hateful content present on social media platforms, and to observe the rate of removal. The platforms monitored as part of the exercise included Twitter, Facebook, YouTube, TikTok and Instagram.

Monitoring was conducted based on a set of key words collected and used through the Get The Trolls Out! project. Once a piece of content was found on a social media platform which went against Community Guidelines, the case was noted internally and then reported directly to the platform. The team then monitored how long it took for the platforms to take action in terms of removals.

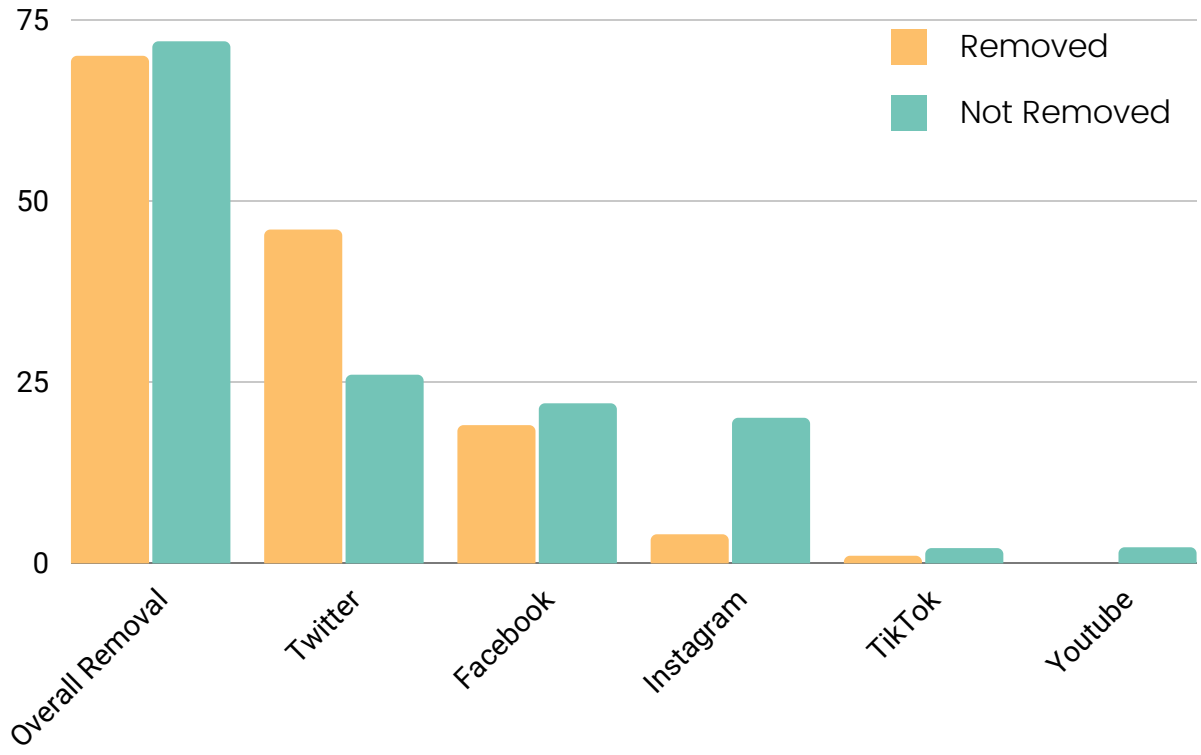
## Chart 1: Total Removal Rates (Facebook, Instagram, TikTok, Twitter and Youtube




In total, 142 cases of calls for violence on religious grounds were reported to various social media platforms over the two-week period. There was an overall removal rate of 49.30%

Overall, Twitter had the highest rate of removal at 63.90%, and YouTube had the lowest at 0%.

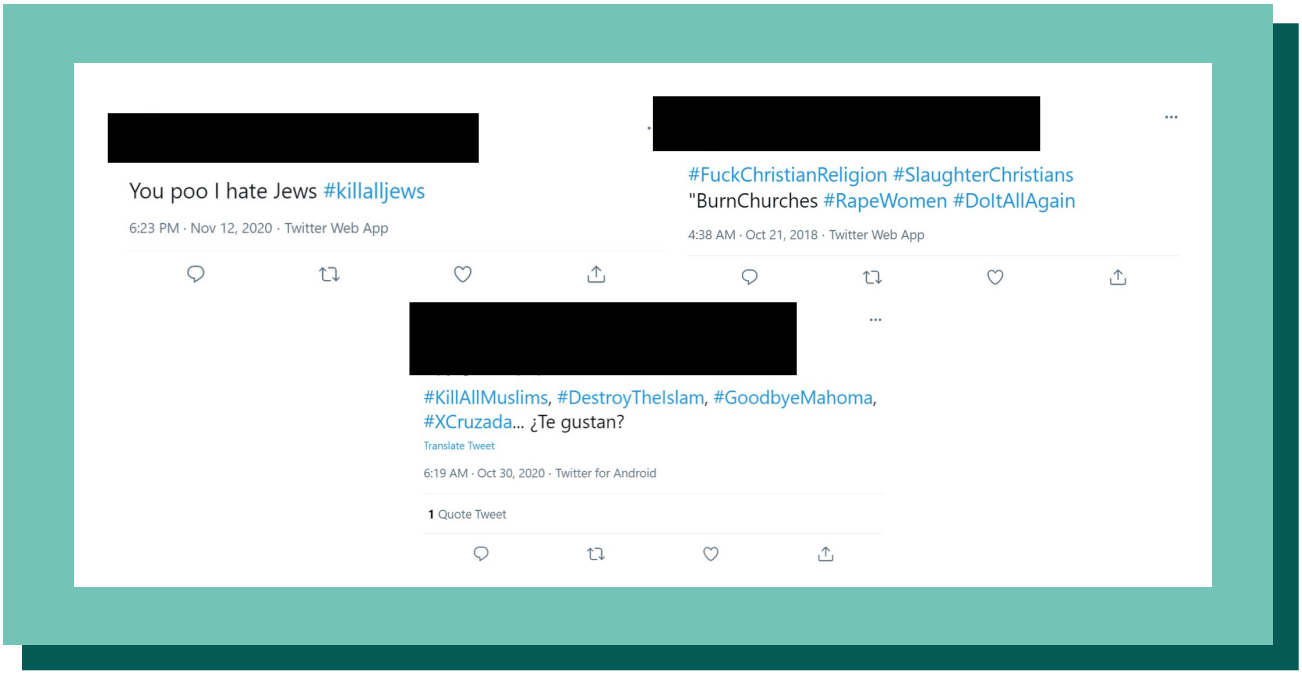
## Chart 2: Removal Rates




While this overall data shows us some interesting trends in terms of removal rates, there were some analytic observations made during the monitoring exercise which we feel show clear gaps in policy implementation on social media platforms. We found the violent hashtags were used without retribution, and also found that there was a clear antisemitic themes present on several social media platforms of people calling for "another Holocaust". We explain these trends in more detail, and with examples, below. Please note, all the examples shown in this report have not been removed at the time of writing this report (March 2021), despite being reported. We would also like to note that for a lot of the cases which we reported on Twitter, and were subsequently removed, we did not receive a notification of removal. This lack of transparency makes the reporting process less accessible.

# Violent Hashtags

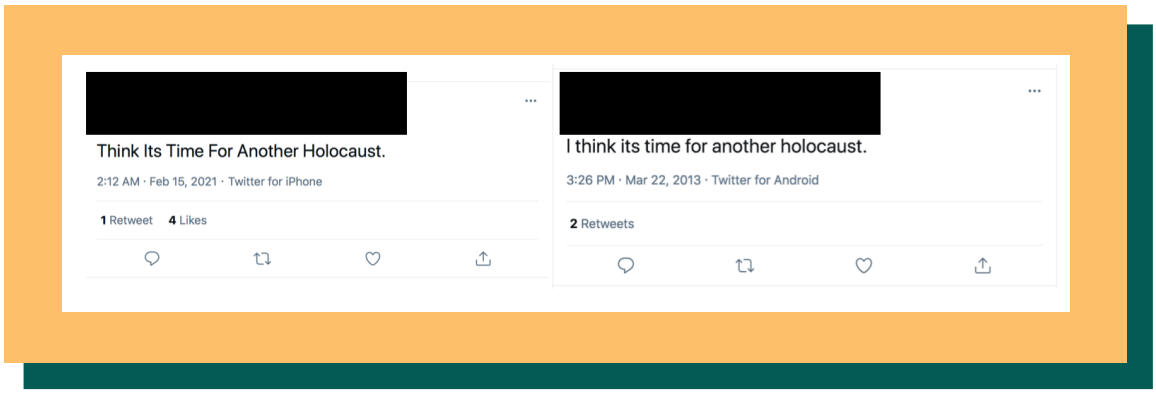
The monitoring team found several cases of extremely violent hashtags being used. Examples include #allmuslimsmustdie (on Instagram) on #killalljews (on Twitter) and #SlaughterChristians" (on Twitter). These are not hashtags which can be interpreted as violent: they are very clearly directly calling for violence towards different groups. We were shocked to find that some of the content using these hashtags was several months old and seems that to have gone undetected. Some examples below:



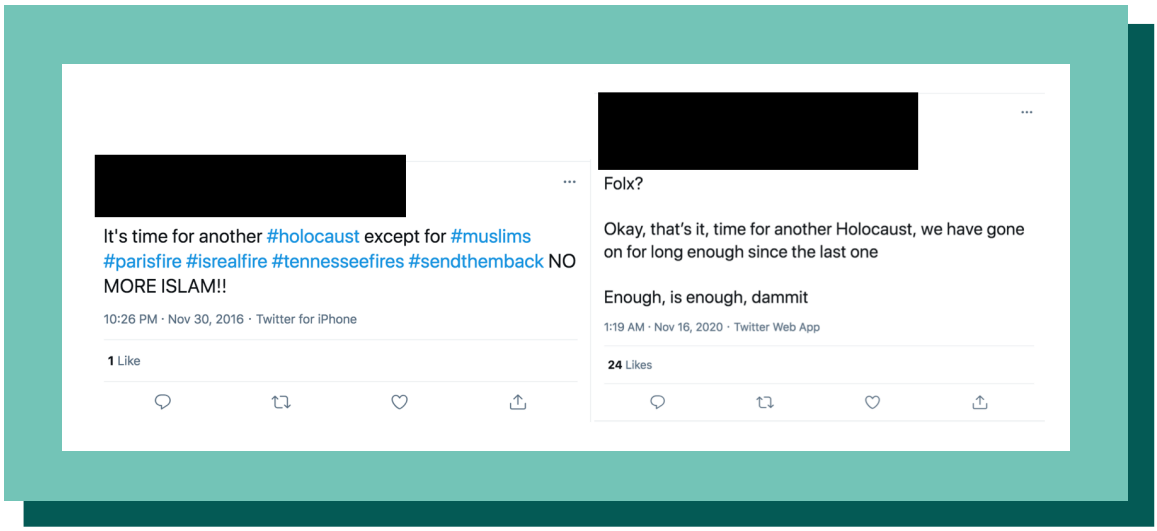

It is important to note that all of these cases were reported for removal, as they go directly against the community guidelines set by social media platforms. The problem is thus twofold: users are able to post content with calls for violence without detection and when this content is found and reported, it still remains online. As a monitoring team, we wonder whether hashtags are less detectable than content itself, and that thus people hide more extreme messages in hashtags. In the past, we have seen several different platforms take direct action with hashtags, such as banning certain QAnon hashtags. If hashtags related to conspiracy theories can be monitored strictly, then surely hashtags clearly calling for violence can be too.

# “Time for Another Holocaust”

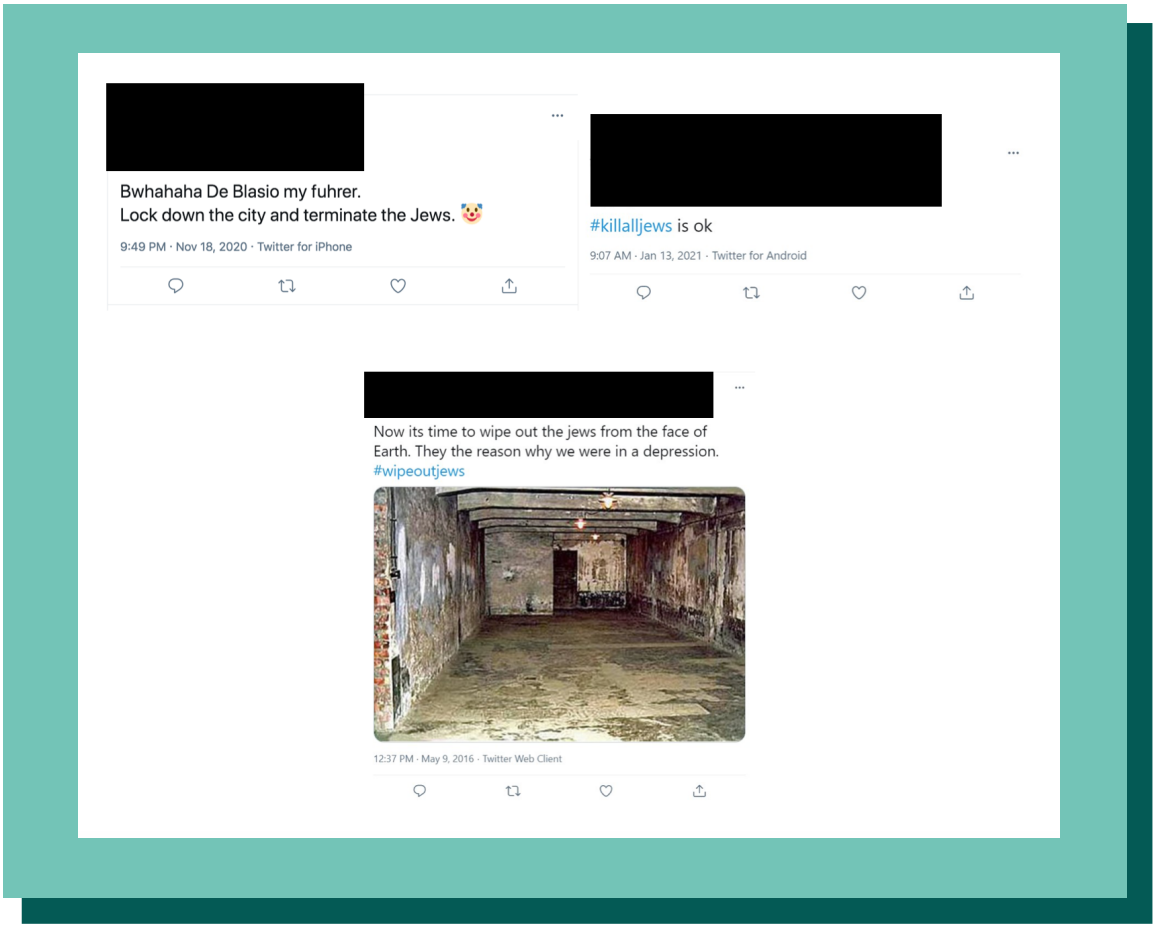
The monitoring team found a clear trend in antisemitic content on the social media platforms monitored. In fact, the majority of the content which we reported but was not removed was antisemitic. There was one clear trend spotted: there was a lot of content which called for “another Holocaust” or claimed that it was “time for another Holocaust”. This trend was mainly spotted on Twitter, though there were also some cases (which were removed after reporting) on Facebook and Instagram. The Holocaust was a horrific, and extremely violent, genocide targeting mainly Jewish people. Calling for “another Holocaust” is this not only antisemitic, but also a call for violence. Yet, a lot of this content is still online. Some examples below:




We also found that while a lot of the ‘Holocaust incitement’ was geared towards Jewish people, it was also used as a call for violence against other minority groups. Calling for genocide to be imposed on a group for people is a direct call for violence, and thus should be removed from platforms based on their own community standards. Some examples:



As noted, whilst a lot of the antisemitic content report was related to the Holocaust, there was also other clearly violent and hateful antisemitic content which we reported but was not removed. Many of these cases came from Twitter. Violent antisemitic content was also found and reported on platforms such as Facebook, Instagram and YouTube but the content was quickly removed. Some examples:



# Key recommendations

Based on this monitoring exercise, and knowledge more regular monitoring activities in the Get The Trolls Out! project, we would like to propose some recommendations for social media platforms:

- Monitoring of hashtags. Users should not be allowed to call for violence by using hashtags, as was shown through the examples in this report. Detection practices should be put in place to identify such hashtags and subsequently remove them and the associated content.

- Crackdown on Holocaust incitement. Calling for another Holocaust to take place is a call for violence. This is already included in the community standards of social media platforms; however, such content is still online, even after reporting.

- More transparency in the reporting process. It is too often the case that the person who has reported the content is not notified of the final decision. This makes the reporting process less accessible, and will ultimately deter people from reporting content in the future.

MEDIADIVERSITY INSTITUTE

GET THE TROLLS OUT!