The 14th International TAU Seminar on Contemporary Antisemitism
Ein-Gedi, November 25-27, 2019

Using Big Data to Monitor Anti-Semitism

Aryeh Tuchman
Associate Director, Center on Extremism
Anti-Defamation League

Dear friends, I regret not being able to attend the seminar this year and share these findings in person. Although the theme of our seminar is governmental approaches to fighting anti-Semitism, as you know, the First Amendment of the United States Constitution makes it illegal for the U.S. government to regulate spoken or written speech, including electronic speech, except under extreme circumstances. For this reason, monitoring anti-Semitism online has been relegated largely to the private sector, academia, and NGOs. Even police, including the FBI, have not developed a culture of defining and monitoring anti-Semitism per se, but rather focus on the detection of criminal activity, and turn to NGOs, including most prominently the ADL, for help when they have questions about anti-Semites and their ideologies.

As a result, it falls to organizations like ADL to pioneer new approaches and methods to monitoring anti-Semitism. By developing techniques that shed light on the nature of anti-Semitism and its prevalence online, we are able to name and shame both anti-Semites and the technology companies who host their content in our world which has been so transformed by social media; our research is also used to generate best practices for use by technology companies and publishers to oppose anti-Semitism and ensure that their platforms are not abused by those seeking to spread hatred and racism.

This is the story of one of ADL's early efforts to assess the nature and prevalence of anti-Semitism online. Although I had hoped to share this with you in an interactive format which could serve to further illuminate what is possible in this type of research, I hope you will forgive me for limiting this to briefer overview of our approach. I am eager to discuss any reactions you may have to this, and you may reach me at atuchman@adl.org. Thank you so much to my colleague, Carole Nuriel, for presenting these findings to you.

Overview

In mid-2016, the Center on Extremism (COE) at the Anti-Defamation League embarked on a project to quantify the amount of anti-Semitism on a major social media platform. After exploring the various options, the COE decided to focus the project on Twitter. Over the course of a year, COE created a taxonomy of anti-Semitism, formulated queries to pull in tweets that could potentially express anti-

Semitism, and developed a methodology to eliminate false positives from the ensuring corpus of tweets and extrapolate the total number of genuinely anti-Semitic tweets using statistical methods. Although we had hoped that our analysis would allow us to quantify the prevalence of individual anti-Semitic themes in the final results, we discovered that our technique was not sufficiently sensitive to make those determinations with methodological rigor. We concluded our project in mid-2018 with a quantitative assessment of the amount of anti-Semitism on Twitter in the 12-month period from February 1, 2017 through January 31, 2018, and a qualitative review of select anti-Semitic themes which had arisen during our analysis. This presentation will set forth some of the practical and methodological challenges we faced and how we overcame them over the course of this two-year project, and will share some of the insights we gleaned which may inform future attempts to quantify anti-Semitism in the age of big data.

The challenge of choosing a platform:

The Center on Extremism has monitored anti-Semitism online since the early days of internet. Although previously we had to rely on a combination of human intelligence gathering and reviews of printed materials published by extremists and anti-Semites, the advent of the internet led extremists to reveal ever more details about their ideas and activities in online forums, websites, and email lists which COE analysts could observe. The rise of social media sites like MySpace and Facebook in the early 2000s presented a challenge because our monitoring work had to shift from focusing on specific online forums where anti-Semites congregated, to monitoring individual anti-Semites across an array of social media sites where they interacted both with each other and with the general public. In the 2010s the social media space continued to fragment into a plethora of sites and services including message boards like Reddit, image boards like 4chan and 8chan, video sharing sites like YouTube and Vimeo, micro-blogging sites like Twitter, and an array niche social media sites like VK, Gab, and Minds. Extremists and anti-Semites are active in all these spaces, and monitoring all of them poses an every-increasing strain on COE resources.

Although it is clear that in-depth monitoring of important individual extremists and anti-Semites will be necessary for the foreseeable future, in mid-2016 we set out to determine whether automated tools could be created which would routinely sift through large amounts of social media content and generate insights into both the quantity of anti-Semitism and the themes and trends in anti-Semitic content. We quickly settled on Twitter as a good test case for this effort because it would pose a minimum of technical challenges. Although other social media platforms – including most notably Facebook – make it difficult to harvest large amounts of data from other users, Twitter's relatively permissive terms of service allowed us to contract with a commercial social media monitoring company to tap into the Twitter "fire-hose" – the raw stream of tweets generated by the platform's users. We hoped that once we obtained access to large volumes of raw tweets that we would be able to develop quantitative methods of assessing that content for anti-Semitism.

Of course, we were aware that focusing our efforts on Twitter had its downsides as well. Although at the time Twitter did not release official statistics on its user base, we sensed that Twitter had far fewer users than larger platforms like Facebook. This was confirmed in February 2019 when Twitter shared that it

had 126 million active daily users, compared to 1.2 billion users of Facebook and almost 200 million users of Snapchat, according to an analysis by the Washington Post.[i] The smaller user base would limit the extent to which our finding could be used to illustrate broader societal trends.

Defining Anti-Semitism:

In order to quantify anti-Semitism it is of course first necessary to define it. As I explained at the outset of this presentation, the First Amendment to the US Constitution has discouraged the US government from developing and adopting a definition of anti-Semitism in our country. Although the State Department has adopted the IHRA definition of anti-Semitism as a tool for assessing the phenomenon in limited cases, especially overseas, no consensus has developed about defining anti-Semitism in domestic contexts. ADL itself has not adopted IHRA for our own purposes. For the purpose of this study, the Center on Extremism took an empirical approach to this problem. Rather than begin with one of the many pre-existing definitions of anti-Semitism which one may find, we developed our own taxonomy of how anti-Semitism manifests online based on the collective observations of COE experts who routinely work in this space. Our taxonomy of anti-Semitism included the following eleven facets:

1. Jews and Money: Terms indicating that Jews are cheap or greedy, that they engage in theft, cheating, or organized crime; that they control banks, the U.S. Federal Reserve, of international finance.
2. Jewish Disloyalty: Terms indicating that Jews are disloyal to the countries in which they live; that they have globalist sympathies or run an international conspiracy; that they are treacherous or untrustworthy.
3. Jews and Social/Political Movements: Terms suggesting that Jews are responsible for what the writers consider cultural degeneracy (e.g. pornography, feminism, homosexuality, drug use) or political extremism (e.g. Bolshevism, neoconservatism, etc.).
4. Jews and Judaism as inherently evil, including references to deicide or cursedness; affiliation with the devil or existential wickedness; practitioners of debased religious or magical rituals including pedophilia and ritual murder.
5. Terms indicating the presence of Holocaust denial.
6. Generic words and proper names surrounded by the echo symbol (three sets of parentheses). Enclosing a word in an echo symbol is a way for anti-Semites to express their belief that the subject within the parentheses is controlled by Jews (e.g. (((media))), (((Wall Street))), (((Trump))), etc.).
7. Terms suggesting that Jews function as a collective (e.g. “Jews always,” “Jews want,” “Jews incessantly,” etc.)
8. Anti-Jewish slurs (e.g. Kike, Heeb, Hymie, etc.)
9. Extensive list of phrases where words for Jews are paired with expletives or negative terms (e.g. “Fucking Jews,” “Jewish bastard,” “goddam synagogues”).
10. Names of known anti-Semites; titles of known anti-Semitic books, podcasts, and media; twitter handles of known anti-Semitic groups and individuals; statements of affirmation for known anti-Semites or their ideas (e.g. “Hitler was right”).
11. Terms pertaining to Zionism, including the use of “Zio-” as a prefix.

COE staff then converted the taxonomy into a 20,000-word Boolean query which we used to harvest tweets from the main Twitter firehose using the platform's API. Because of limits inherent in the architecture of the API and the commercial tool we had selected to harvest the tweets, we were forced to break up the query into distinct parts which were run sequentially; we saved the results of each part of the query in massive spreadsheets which we then deduplicated using custom scripts. The deduplication was necessary to ensure that a tweet which contained terms from more than one part of the master query was only included once in the final corpus of tweets.

The challenge of false positives

Of course, the fact that a given tweet might contain terms present in our query does not conclusively prove that the tweet expresses anti-Semitic sentiment. In order to accurately gauge the number of anti-Semitic tweets, we had to find a way of eliminating false positives from our results. This was the biggest methodological and practical challenge we faced in this project. We found several categories of false positives:

- Non-anti-Semitic Homonyms: Many terms may have multiple meanings, only some of which may be anti-Semitic. For example, the term "kike," which has been used as an anti-Jewish slur in the United States since at least the early 20th century, is also an affectionate nickname given to Enrique Martinez, a star baseball player on the Los Angeles Dodgers, whose team played in the World Series which occurred during the period of our study. Similarly, the term "ZOG," which is used by some anti-Semites as an acronym for "Zionist Occupation Government," is also part of the name of a New York-based company called "ZOG Sports." Thus the mere presence of the term "Kike" or "ZOG" in a specific tweet could not be construed as proof that the tweet was promoting anti-Semitism. We experimented with ways of manually excluding non-anti-Semitic uses of these terms from our results by refining our Boolean query. In the case of "Kike" we modified the query to search for "(Kike) AND NOT (game OR world series OR left fielder)," and in the case of "ZOG" we modified the query to exclude "ZOG Sports." Although these modifications to our query significantly cut down the number of false positives, we found that using Boolean tools to eliminate non-anti-Semitic homonyms is a time-consuming process that requires constant oversight by researchers who must stay on top of the ever-shifting uses of language online.
- Ironic use of anti-Semitic terms: Upon sampling some of the millions of tweets which were pulled in by our query, we found many cases where people would consciously use anti-Semitic language in an ironic, self-referential way. For example, someone might write a tweet stating, "This cheap Jew just bought his first new car!" Although arguably the reference to himself as a cheap Jew could be seen as serving to perpetuate an anti-Semitic stereotype, in context that did not seem to be in the intention of the writer.
- Rejection of anti-Semitic terms: A large percentage of tweets which were captured by our query because they contained anti-Semitic terms, were clearly written in order to express opposition to anti-Semitism. This was most apparent in the aftermath of the 2017 white supremacist march in Charlottesville, VA, when Twitter was flooded with users who used their tweets to express their shock and dismay at the fact that white supremacists had chanted "Jews will not replace

us" at the event. The phrase itself had been included in our taxonomy of anti-Semitic language, but its actual use by white supremacists on twitter was overwhelmed by its use by anti-racists and other users who opposed anti-Semitism.

Initial, non-scientific assessments by COE staff of the results of our Boolean searches showed that false positives accounted for as much as 80% of the tweets which were identified as containing language which could signify the presence of anti-Semitism in some of the categories of our taxonomy. In order to continue our study we consulted with several statisticians and data scientists who helped us devise a method of statistically sampling the corpus of tweets pulled in by our Boolean query to generate a subset of tweets for further, manual review. These subsets would then be assessed by experts in the COE to determine the precise percentage of tweets which were not false positives and actually expressed anti-Semitic sentiment, with a margin of error of three percent. Once our experts assessed these subsets of tweets we could then extrapolate this percentage back to the larger general population, and calculate the total number of anti-Semitic tweets.

Conclusions

We believe that the two-pronged nature of our method in this study, which used automated methods to harvest an initial corpus of tweets which was then supplemented by careful human review by experts, has yielded a highly accurate count of anti-Semitic tweets in English during each of the fifty-two weeks covered by our study. In total, we estimate that approximately three million unique Twitter handles were responsible for the 4.2 million tweets.

However, the manual reviews of the content were extremely time consuming, and limited the number of questions we could ask of our finished results with scientific accuracy. When it came to identifying themes and trends within the identified anti-Semitic content, we were forced to rely on non-quantitative assessments by staff. Our complete findings were released in mid-2018 in a report entitled *Quantifying Hate: A Year of Anti-Semitism on Twitter*. The report is available on the ADL website and at the following shortcut: bit.ly/ADLTwitterReport

ADL learned valuable lessons about using big data to monitor anti-Semitism as a result of conducting this study. We surmounted several difficult methodological and technical challenges. We also came away from the study with a fresh taxonomy of anti-Semitic language, and an incredibly valuable dataset of manually coded tweets which tested positive and negative for anti-Semitism. These datasets formed the foundation for subsequent, ongoing experiments in using machine learning and artificial intelligence to monitor anti-Semitism in the age of big data.

[i] https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/