Detecting Anti-Jewish Messages on Social Media. Building an Annotated Corpus That Can Serve as A Preliminary Gold Standard

Gunther Jikeli,[1] Deepika Awasthi,[1] David Axelrod,[1] Daniel Miehling,[2] Pauravi Wagh,[1] and Weejoeng Joeng[1]

Indiana University,[1,] Technical University Berlin[2]
{gjikeli, dawasthi, daaxelro, damieh, pwagh, weejeong}@iu.edu

Abstract

Hate speech detection in online environments faces numerous challenges. One of them is that hate speech has fundamental target-specific elements. Although certain characteristics are common to many forms of hate speech, forms directed against one group, such as Jews, can be very different from forms directed against Muslims, Roma, members of the LGBTQ+ community, and bullying victims. Due to the heterogeneity of hate forms, we suggest approaching forms piecemeal and building labeled datasets that are specific to target groups. These datasets can then be combined into an aggregate dataset that increasingly reflects the diversity of hate speech found for a given language. Another challenge is the subjectivity of annotators and heterogeneous labeling.

We created a labeled dataset of 4,137 antisemitic and non-antisemitic tweets, using a detailed definition and a specially designed annotation portal. The annotation was done by expert annotators who discussed their disagreements of each tweet. The dataset is built on representative samples of tweets containing more common keywords (such as "Jews") and keywords most likely to be used in antisemitic contexts (such as the term "kikes"). The dataset will be made available to the scientific community with the publication of this paper and will be updated with additional tweets and labels as the project continues.

The paper describes the dataset, the labeling process, the infrastructure that was built for this project, some of the challenges that we faced, and an evaluation of the inter-coder reliability. The goal is to provide a detailed description of the labeled dataset to serve as a preliminary gold standard and a model for creating similar datasets.

Introduction

Online hate speech has increasingly been the focus of public debate and social media platforms have pledged to remove hate speech from their platforms. Since the "Unite the Right" rally in Charlottesville, Virginia, in August 2017 and the public outcry following the killing of one counter-protestor, major platforms, such as Facebook and Twitter have suspended a portion of accounts violating their updated terms of service. More systematic suspensions and deletions of accounts came after the 2019 attack at two mosques in Christchurch, New Zealand where the terrorist killed 51 people and live-streamed the killings. Another push came after the violent riots on Capitol Hill in January 2021. These efforts to remove hateful content have been imperfect and it has become evident that better mechanisms, improved algorithms, and more transparency needs to be put in place to deal with harmful content on social media. Additionally, hate speech detection can be used in multi-dimensional context analysis to detect extremist narratives used by terrorists for preventative applications (Kursuncu et al. 2020).

Antisemitism is a core element of ideologies that are closely related to hate speech, such as white nationalism (Ward 2019) and jihadism (Rickenbacher 2021).

Related Work

Zannettou et al. (2020) present a framework for quantitative analyses of online antisemitism. Their quantitative measurement of slurs and antisemitic memes works well for fringe communities, such as 4chan and Gab, where users frequently use such explicit expressions of hate and where even the word "Jew" can be a strong indicator of hate speech. However, that is less useful for mainstream social media platforms, such as Twitter. For example, many users

on mainstream platforms call out others' use of antisemitic slurs, potentially leading to false-positive classification.

Another approach to hate speech detection is training algorithms with labeled datasets (gold standards). More and more such gold standards on hate speech are publicly available. However, automated hate speech detection is still challenging. The largest publicly available labeled dataset on hate speech was created by Gomez et al. (2020).[2] 150,000 tweets (including each one of 51 terms from Hatebase.org) were labeled in six categories by three different workers from the crowdsourcing platform Amazon Mechanical Turk: No attacks against any community, racist, sexist, homophobic, religion-based attacks or attacks against other communities. They used majority voting to determine the category of each tweet and experimented using hate scores for each tweet to account for the different votes by the three annotators instead of binary labels. However, they found that a major challenge for their models is the discrepancy between annotations due to subjective judgement. Their best performing model, using text and images, had a mean accuracy of 68.5 percent and an F-score of 0.702, not outperforming models based on text only.

Davidson et al. (2017) created another large dataset. 24,802 tweets were labeled by at least three human annotators and classified in three groups: hate speech, offensive but not hate speech, or neither offensive nor hate speech. The intercoder-agreement score provided by CF was 92 percent. They used the majority decision for each tweet to assign a label. Unanimous decisions were considerably lower than 2/3 decisions. Malmasi and Zampieri (2017) applied a linear Support Vector Machine (SVM) classifier on 14,509 tweets from the same dataset that are publicly available via Crowd- Flower.[3] Based on a 4-gram model, their classifier achieves 78 percent accuracy.

Waseem (2016) created another publicly available dataset of 6,909 labeled tweets. One expert annotator and three amateur annotators classified tweets in four groups: non-hate speech, racism, sexism, and both (i.e., racism and sexism). Inter-annotator agreement among the amateur annotators was $k = 0.57$. Gambäck and Sikdar (2017) used Waseem's dataset to train four deep learning models (CNN) but their best performing model, based on word2vec embeddings does not substantially outperform on the binary classification by Waseem and Hovy (2016), with higher precision than recall, and a 78.3 percent F-score.

Our aim is to build a labeled dataset of high accuracy and high inter-annotator agreement, targeting a certain form of hate speech (antisemitism) only. We hope that this might improve performance of models based on this dataset. Research on antisemitism is only at the beginning of the computational turn (Bruns 2020). However, another research group is working on a project that is similar to ours. Chandra et al. (2021) built a labeled dataset on antisemitism of 3,102 posts on Twitter and 3,509 posts on Gab (soon to be published[4]), focusing on posts that include both images and text and words related to Jews, such as 'Jewish', 'Hasidic', 'Hebrew', 'Semitic', 'Judaistic', 'israeli', 'yahudi', 'yehudi', and also slurs. Three annotators labeled posts as antisemitic or not and classified antisemitic posts in one of the four categories; political, economic, religious, or racial antisemitism, and also used the Working Definition of Antisemitism by the International Holocaust Remembrance Alliance (IHRA Definition). (Chandra et al., 2021) One of the main differences to our labeled dataset is that we use samples from a 2019 and 2020 dataset that includes ten percent of all tweets on a statistically relevant basis.

However, time-consuming manual annotation is the bottleneck for most supervised machine learning projects. Our project on antisemitic tweets is not different in principle from many other hate speech dataset projects that include defining a classification schema, labeling guidelines, gathering adequate data, pre-processing this data according to the task, training experts for labeling, and building a final corpus (Pustejovsky and Stubbs 2012).

Generating Our Corpus

Our dataset is an aggregate of samples taken from a Twitter database using a 10 percent stream of Twitter data and is managed by Indiana University's Observatory on Social Media (OSoMe). We can query this database, going back 36 months. The database is compliant with Twitter policy and removes deleted tweets on a regular basis. As such, we use live tweets only. The database allows us to build subsamples with keywords that are statistically representative of all tweets with these keywords. We then manually label the subsamples.

To build our preliminary gold standard, we would ideally use a completely randomized sample of tweets spanning all variations of antisemitic and non-antisemitic tweets. However, we are limited by time and the number of expert annotators who manually evaluate tweets. Labeling test samples showed that the evaluation of one tweet requires 1-2 minutes, on average, and that annotators can rarely do more than 100 tweets per day before quality suffer. As we wanted to have each tweet labeled by two experts, we aimed for a preliminary dataset of 5,000 tweets.

A completely randomized sample of 5,000 tweets would result in too few antisemitic messages and would thus have failed to include many varieties of antisemitic content. We therefore opted for keywords that would ensure gathering

[2] https://gombru.github.io/2019/10/09/MMHS/

[3] https://data.world/crowdflower/hate-speech-identification

[4] https://github.com/mohit3011/Online-Antisemitism-Detection-Using-MultimodalDeep-Learning

tweets that are thematically closer to discussions around Jews. We focused on two keywords that contain the greatest number of tweets related to Jews as a religious, ethnic, or political community: "Jews" and "Israel." We then added a few samples with more targeted keywords likely to generate a high percentage of antisemitic tweets: "kikes" and "ZioNazi*". Table 1 shows an overview of our queries and samples.

Sample	Keyword	Timespan	# Tweets in dataset
1	Jews	Jan.-Dec. 2019	1,230,801 as of 12/26/20
2	Jews	Jan.-Dec. 2019	1,230,801 as of 12/26/20
3	Jews	Jan.-Apr. 2020	238,965 as of 05/15/20
4	Jews	Jan.-Apr. 2020	238,965 as of 05/15/20
5	Jews	May-Aug. 2020	329,804 as of 09/01/20
6	Jews	May-Aug. 2020	329,804 as of 09/01/20
7	ZioNazi*	Jan.-Dec. 2019	2,079 as of 05/20/20
8	ZioNazi*	Jan.-Apr. 2020	342 as of 05/20/20
9	Israel	Jan.-Apr. 2020	834,349 as of 05/24/20
10	Israel	May-Aug. 2020	1,053,375 as of 09/01/20
11	kikes	Jan.-Dec. 2019	1,332 as of 5/20/2020

Table 1: Samples and queries of raw data of corpus

From these query results, totaling 3,691,047 tweets across all queries, we generated randomized samples, aiming for 500 live tweets per sample. We first selected 2000 tweets from each query (if the total number was sufficiently large) by applying randomized reservoir sampling. We then selected the live tweets from those 2,000 tweets. We used Tweepy and the tweets' IDs to check if the tweet is live. This was necessary because tweets are deleted and suspended at any time after our initial query. From the remaining tweets we created subsamples of 500 tweets, resulting in 11 samples (see table 1). For the keyword "Jews" we generated two samples per timeframe. Reservoir sampling was applied each time. Ten samples had between 496 and 500 live tweets at the time of sampling and after we uploaded them to our "Annotation Portal." The sample "ZioNazi* (January to April 2020) had only 232 live tweets at the time of sampling due to the overall small number of tweets from that keyword. For the "kikes" query, we added a filter that deletes (some) non-English tweets, using Google's language-detection library. Non-text tweets, such as those only containing URLs were exempt from language detection.

Annotation

We used a web interface, our "Annotation Portal," for labeling our samples. This infrastructure was developed specifically for this project to improve the quality of annotation. The Portal shows the tweet and a clickable annotation form (Illustration 1).

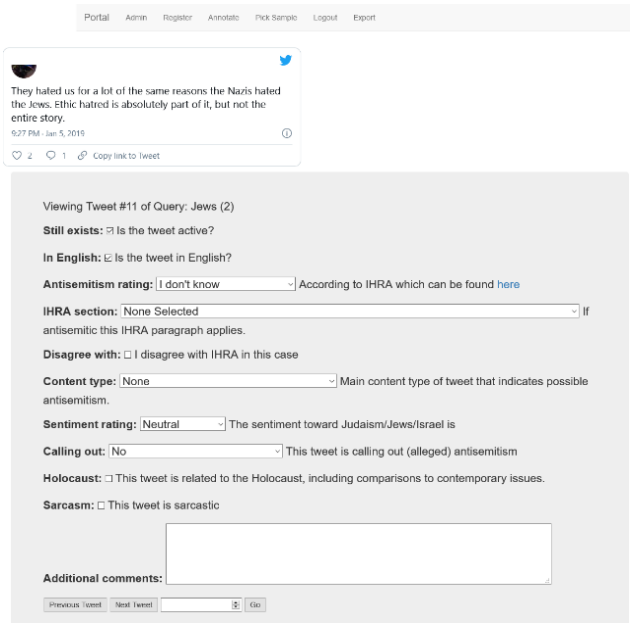


Illustration 1: Annotation Portal with tweet example[5]

As the goals is to have consistent labeling across the annotators assigned to a sample, the annotation form is designed to help make that decision as simple as possible and to focus on the application of the definition of antisemitism that we gave to the annotators.

The Annotation Portal pulls up live tweets by referencing their ID number. Our annotators first look at the tweet and if they are unsure about the meaning, they are asked to inspect the entire thread, replies, likes, and comments.

The tweet is visualized above the annotation form and annotators can click through all questions and then hit the "next" button. If they are still unsure of the meaning of the tweet they can click on the tweet and see it in the normal platform interface. They can also click on the user to find out more information about them or use Twitter's Advanced Search to see what other messages the user has tweeted.

[5] This is a screenshot of our updated form. The question about the content type was not used for the annotation of this dataset.

The first question in the annotation form is if the tweet is still live. Although we filtered for live tweets when we generated the samples, some tweets had been deleted after sample generation. This explains some of the discrepancies between the annotators as they did not annotate the samples at the same time. In some cases, messages from suspended users reappeared.

The second question is if the tweet is in English. Twitter is most prominent among English speakers and our keywords were all in English. However, in some samples we had non-English tweets that our annotators did not label.

Annotators had five options for the antisemitism rating according to a detailed definition of antisemitism: "confident not antisemitic; probably not antisemitic; I don't know; probably antisemitic; and confident antisemitic." We used the Working Definition of Antisemitism developed by the International Holocaust Remembrance Alliance along with a detailed description to make it usable for annotation (Jikeli et al. 2019). Annotators read this definition and explanation carefully and they were trained on some samples before they started the annotation process. All annotators had taken at least one university course on antisemitism or had a similar training.

If the annotators labeled the tweet as "probably antisemitic" or "confident antisemitic" according to the definition, they had to choose one of the twelve paragraphs of the definition that informed this decision. If none of the paragraphs applied, they were instructed to not label the tweet as antisemitic. They could, however, check a box that they disagreed with the definition on that tweet and put an explanation in a comment box. This further helped the annotators stick to a standard rather than their personal interpretations.

Asking the annotators to choose between a "very negative, negative, neutral, positive, or very positive" sentiment for the tweet with regard to Jews, Judaism, or Israel further helped the annotators apply the definition because they could express that the tweet had negative sentiment even if they were not able to find a section of the definition that applied. Many tweets were related to anti-Jewish sentiment, but they were in fact calling out antisemitism. Annotators could also label tweets that they understood to be sarcastic. Lastly, we wanted to know if a tweet was related to the Holocaust in some way.

Annotation Results

Five expert annotators of multiple faiths and genders went over eleven samples of tweets. Each sample was annotated by two annotators separately. After the annotation, we identified the tweets for which they disagreed in their antisemitism rating, that is, if one of the annotators had annotated a tweet as antisemitic (probably or confident) and the other had not. The raters then discussed their disagreements. In many cases, human error, or some oversight of some aspects of the tweet could be identified quickly and the annotators corrected their annotation. In other cases, a detailed discussion about which paragraph in the definition could be applied (or not), or a detailed discussion about the meaning of a certain tweet would clarify the matter and an agreement was reached. This discussion about meaning and context turned out to be particularly helpful for annotators who were initially unfamiliar with some political context or celebrity events in the U.S., Britain, India, or elsewhere. The annotators became increasingly familiar with the contexts as they often revolved around similar topics. The annotators went through three rounds of discussion and eventually came, in almost all cases, to an agreement. The tweets for which no agreement was found were removed from the final dataset.

Sample #	Number of tweets (Before Discussion)	Not annotated because deleted/ suspended/ foreign language	Annotated by	Percentage of antisemitic tweets
1	439	61	jg	6.2 %
1	455	45	dm	6.2 %
2	414	86	jg	7.5 %
2	428	72	dm	5.4 %
3	468	32	jg	12.2 %
3	466	34	js	9.9 %
4	429	70	jg	12.1 %
4	430	70	js	8.4 %
5	390	106	jg	12.1 %
5	405	91	sm	9.4 %
6	386	112	js	14.5 %
6	396	102	sm	10.9 %
7	348	152	js	88.2 %
7	391	109	dm	82.4 %
8	149	83	dm	83.9 %
8	140	92	sm	85 %
9	342	158	js	5.0 %
9	480	19	dm	0.4 %
10	431	69	js	13.2%
10	412	88	ks	14.6 %
11	295	205	dm	35.3 %
11	283	217	sm	7.9 %

Table 2: Annotation results before comparison

Table 2 shows the annotation results of the antisemitism rating for each sample and each annotator before the annotators discussed their discrepancies. The annotators did not annotate the samples at exactly the same time. This explains the discrepancy in the number of tweets that were not annotated because they were deleted or suspended at a given time. In this table and in our labeled dataset we use

binary categories and treat ratings of "confident not antisemitic; probably not antisemitic; and I don't know" as not antisemitic and "probably antisemitic and confident antisemitic" as antisemitic. For most samples, the annotators found a similar percentage of tweets to be antisemitic. However, the number of tweets that were rated as antisemitic by only one annotator and not the other was high for some samples, particularly sample 11 where 74 tweets that were rated differently (see table 3).

Sample	Keyword	Timespan	# of tweets in disagreement
1	Jews	Jan.-Dec. 2019	9
2	Jews	Jan.-Dec. 2019	20
3	Jews	Jan.-Apr. 2020	19
4	Jews	Jan.-Apr. 2020	36
5	Jews	May-Aug. 2020	43
6	Jews	May-Aug. 2020	34
7	ZioNazi*	Jan.-Dec. 2019	13
8	ZioNazi*	Jan.-Apr. 2020	4
9	Israel	Jan.-Apr. 2020	19
10	Israel	May-Aug. 2020	44
11	kikes	Jan.-Dec. 2019	74

Table 3: Number of tweets that annotators rated differently (antisemitic/ not antisemitic)

The high level of disagreement in sample 11 before discussion is related to the fact that this sample of tweets from the "kikes" query include many tweets by bots that are difficult to interpret. Many of the tweets with that keyword were about a famous soccer player, Enrique García Martínez and a famous baseball player, Enrique Javier Hernández, both nicknamed "kiké." However, most of those tweets were in Spanish and are not included in our Gold Standard. However, in some cases, this might have been confused with the antisemitic slur. Some nonsensical tweets from bots and the nickname kiké explain why the percentage of antisemitic tweets was relatively low for a sample queried with an antisemitic slur.

Inter-Rater Reliability Prior to Discussion

Inter-rater reliability shows how much homogeneity or consensus exists in the ratings and it can be seen as an indication of how diligent our annotators were in applying the given definition consistently. Two annotators evaluated each sample independently before discussing their disagreements in multiple rounds. We estimate inter-rater reliability prior to discussion with Cohen's kappa (Cohen 1960) and Gwet's AC_1 (Gwet 2008). The level of measurement is nominal (1 = not antisemitic and 2 = antisemitic). Cohen's kappa (k) measures the extent of agreement between two annotators on categorical variables. Commonly used criteria to interpret kappa coefficients are as follows: <0.00 as poor agreement, 0.00-0.20 as slight, 0.2-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and > 0.80 as almost perfect (Landis and Koch 1977). Kappa coefficients, however, have limitations because they are dependent on the prevalence (probability of rater's classification of tweet annotation into a category) and bias (frequency of raters' choice of a specific category). To overcome these issues, Gwet's AC_1 statistic is used, too. Both k statistic and AC_1 are developed based on percent agreement that is corrected for chance agreement employing different strategies but AC_1 statistic is affected less by the prevalence and bias. Interpretation of AC_1 coefficients is similar to the k statistic. Percent agreement between annotators is used for the agreement statistic.

		Kappa Statistics (k)		Gwet's AC_1	
Sample	% Agreement	K co-efficient	95% CI	AC_1 co-efficient	95% CI
1	97.9	0.83	(0.71, 0.94)	0.98	(0.96,0.99)
2	95.1	0.6	(0.45, 0.76)	0.94	(0.92,0.97)
3	95.9	0.79	(0.7, 0.88)	0.95	(0.93,0.97)
4	91.5	0.54	(0.4, 0.67)	0.90	(0.86,0.93)
5	88.9	0.41	(0.26, 0.55)	0.86	(0.82,0.91)
6	91.1	0.6	(0.47, 0.72)	0.89	(0.85,0.93)
7	96.2	0.84	(0.75, 0.92)	0.95	(0.92,0.98)
8	97	0.89	(0.79, 1)	0.95	(0.92, 1)
9	94.4	-0.01	(-0.02, 0.003)	0.94	(0.91, 0.97)
10	89.2	0.56	(0.44, 0.67)	0.85	(0.81, 0.90)
11	73	0.3	(0.2, 0.41)	0.58	(0.48,0.68)

Table 4: Inter-annotator reliability before discussion

Table 4 presents percent agreement and inter-rater reliability for pre-discussion rating. Since two annotators are involved in annotating tweets, k statistic and Gwet's AC_1 statistic with 95 percent confidence intervals (CI) are computed to examine inter-rater reliability for each dataset. The overall percent agreement is over 80 percent for pre-discussion annotations for all datasets except for dataset 11. The overall inter-rater reliability is substantial to almost perfect (k >0.60/AC1 > 0.70) for most datasets. Note that when k coefficient is low due to skewed distribution of prevalence of antisemitism rating but Gwet's AC_1 and percent agreement are high, the inter-rater rating is considered reliable.

The samples 1 and 2 (Jews 2019-rep1and Jews 2019-rep2) have k coefficients of 0.83 and 0.60 and AC_1 of 0.98 and 0.94 respectively, indicating substantial to almost perfect agreement. The sample 3 (Jews2020Jan-Apr-rep1) has a k coefficient of 0.79 and AC_1 of 0.95, demonstrating almost perfect agreement. For the sample 4 (Jews2020Jan-Apr-rep2), the k coefficient of 0.54 offers moderate agreement but an AC_1 of 0.90 and a percent agreement of 91.5 indicate almost perfect agreement. This suggests the distribution of prevalence of antisemitism rating is skewed and thus k coefficient is low. The samples 5 and 6 (Jews2020.May-Aug.rep1and Jews2020.May-Aug.rep2) present moderate to almost perfect agreement (k 0.40-0.60/AC1>0.80). The samples 7 and 8 with the keyword ZioNazi* show almost perfect agreement (k > 0.80/AC1>0.90). The samples 9 and 10 (Israel2020Jan-Apr.rep1 and Israel2020May-Aug.rep3) have low k coefficients of -0.1 and 0.56 respectively, indicating the skewed distribution of prevalence of antisemitism rating while AC_1 coefficients of 0.94 and 0.85 suggest almost perfect agreement. For sample 11 (Kikes2019.rep1) although a percent agreement of 73 and an AC_1 of 0.58 indicate only moderate agreement, a k coefficient of 0.3 is low due to the skewed distribution of prevalence of antisemitism rating. The main factor of disagreement was the different labelling of posts of seemingly random words put together by bots. However, the annotators found agreement when discussing their different labeling. Tweets by (presumed) bots that include the slur "kikes" were eventually labeled consistently as antisemitic if the reader would infer that the slur stands for a group of people, that is, Jews, and is thus used as a slur, expressing hatred. If the text was nonsensical to a degree that it was not even clear that the slur denotes a group of people, then it was not labeled antisemitic.

The overall inter-rater agreement is high in almost all samples. In samples with a very low or very high percentage of antisemitic tweets (skewed distribution) this might be misleading. However, closer analysis of k statistic and Gwet's AC_1 suggest that the inter-rater rating is considered reliable. However, our goal was to create a univocal labeled dataset with no disagreement on the antisemitism rating. We achieved this by having the annotators discuss their disagreements in-depth, including going back to the tweets and finding out more about the context. This led to an agreement of 100 percent in almost all samples. Annotators could not agree on the rating of two tweets because there was not sufficient context. These were not included in the labeled dataset (Gold Standard).

Corpus and Raw Data Description

Our data comes from three distinct time periods, the entire year 2019, January to April 2020, and May to August 2020, drawing on a large dataset composed of a ten percent sample of all tweets. The corpus of tweets was drawn from eight queries with four different keywords, "Jews, Israel, kikes, and ZionNazi*", resulting in 3,691,047 tweets. We generated representative samples, two for each of the three queries with the keyword "Jews" and one for all other queries. This resulted in 11 samples with 5,224 tweets altogether. After the elimination of tweets that could not be annotated due to deletion, suspension, or because they were in a foreign language, 4,137 tweets remained in the dataset. 933 (22.55 percent) of these tweets were rated by two annotators as antisemitic according to the IHRA Definition of antisemitism. Table 5 shows the number of tweets of each sample of the Gold Standard that were annotated by two expert annotators and the percentage of antisemitic tweets. In some cases, the number of tweets is significantly lower than the original sample size of 500 tweets because tweets were in a foreign language or suspended or deleted when at least one of the annotators evaluated them. In some samples, such as sample 9 with the keyword "Israel," suspended tweets reappeared during the discussion process and could then be annotated. This explains the higher percentage of antisemitic tweets after comparison. Antisemitic tweets went live again.

The annotators discussed all tweets that they rated differently. Only two tweets across all samples were removed from the dataset because the annotators could not agree if the tweets were antisemitic, lacking sufficient context.

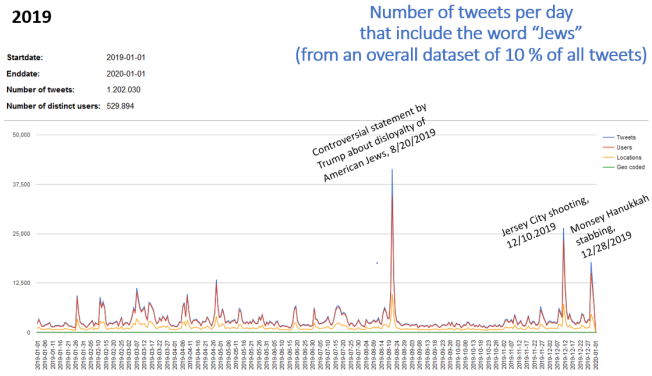
	Keyword	Timespan	Number of tweets in Gold Standard corpus	Percentage of antisemitic tweets
1	Jews	Jan.-Dec. 2019	439	6.2 %
2	Jews	Jan.-Dec. 2019	414	7.5 %
3	Jews	Jan.-Apr. 2020	469	11.9 %
4	Jews	Jan.-Apr. 2020	429	11.4 %
5	Jews	May-Aug. 2020	394	14.0 %
6	Jews	May-Aug. 2020	388	16.2 %
7	ZioNazi*	Jan.-Dec. 2019	374	88.8 %
8	ZioNazi*	Jan.-Apr. 2020	158	85.4 %
9	Israel	Jan.-Apr. 2020	344	10.2 %
10	Israel	May-Aug. 2020	431	13.0 %
11	kikes	Jan.-Dec. 2019	297	31.6 %
SUM		**Jan. 2019 to Aug. 2020**	**4,137**	**22.55 %**

Table 5: Samples of preliminary Gold Standard corpus

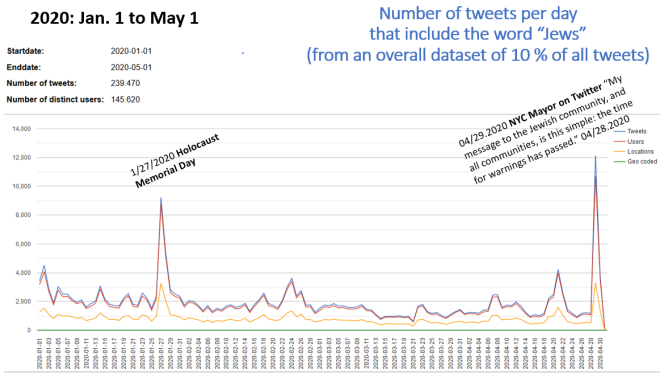
As expected, our samples with the generic keywords “Jews” and “Israel” had relatively low percentages of antisemitic tweets. The “Jews” query had between 6.2 and 16.2 percent with an average of 11.1 percent (281 tweets) for all samples and was lower in 2019 and higher in 2020.

The samples with the slur “ZioNazi*” also met expectations with a high percentage of antisemitic tweets, between 85.4 and 88.8 percent. By contrast, the sample with the slur “kikes” had less antisemitic tweets than expected with 31.7 percent. This is in large parts because many tweets referred to two celebrities whose nickname is “kike,” as discussed above.

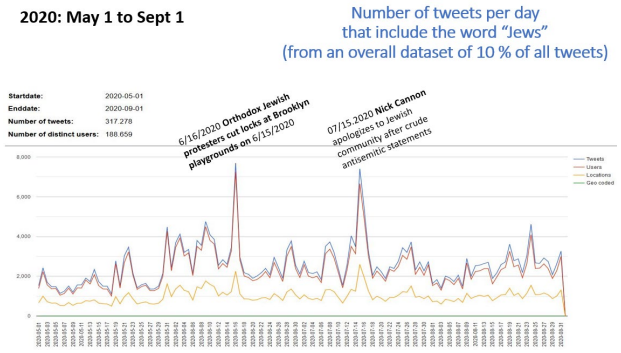
The majority of tweets in the dataset, 2533 tweets (61.2 percent), are tweets generated using the “Jews” query. We therefore focus on a description of these samples. Graphs 1-3 show the time series for the three periods of our queries.



Graph 1: Timeline of tweets with the word Jews in 2019



Graph 2: Timeline of tweets with the word Jews from January to April 2020



Graph 3: Timeline of tweets with the word Jews from May to August 2020

The three graphs show significant peaks that relate to specific events, such as, a statement by President Trump about disloyalty of American Jews on August 20, 2019; the shooting in Jersey City on December 10, 2019; the Monsey Hanukkah stabbing on December 28, 2019; Holocaust Memorial Day on January 27, 2020; a statement about Jews by NYC Mayor de Blasio on Twitter on April 28, 2020; a viral video of Orthodox Jewish protesters cutting locks at a Brooklyn playground on 6/15/2020; and a statement by Nick Cannon about Jews (and his apology) on July 15, 2020. These seven narratives are thus prominent in our original query and also in the samples that were drawn from this raw data.

The top 10 influencers, defined as users most often mentioned in tweets with the word “Jews,” also provide some information about our dataset and prominent themes. In 2019, the top 10 influencers were: realDonaldTrump, IlhanMN, AuschwitzMuseum, _SJPeace_, AOC, Imamofpeace, RashidaTlaib, charliekirk11, jeremycorbyn, and LizaRosen101. “realDonaldTrump” had 32,173 mentions in the dataset, “IlhanMN” 21,967, and “AuschwitzMuseum” 15,999. The prominent role of “realDonaldTrump” in tweets about Jews in 2019 might be related to the highest peak that year that coincided with a statement by Trump about the alleged dual loyalty of American Jews.

Between January and April 2020, the top 10 influencers were the users AuschwitzMuseum, LizaRosen101, Imamofpeace, IlhanMN, realDonaldTrump, NYCMayor, TheRaDR, DavidAstinWalsh, DineshDSouza, and _DavidAsher. “AuschwitzMuseum” was mentioned most often. This might be related to the second highest peak in that period, related to Holocaust Memorial Day. The highest peak in that period is related to a statement by NYC Mayor de Blasio. “NYCMayor” was number six in the ranking of most mentioned users.

Illustration 2: Popular retweet

Between May and August 2020, the top 10 influencers were the users AuschwitzMuseum, realDonaldTrump, NYCMayor, SecPompeo, TheRaDR, jeremynewberger, DineshDSouza, CalebJHull, DonaldJTrumpJr, marklevinshow. The peaks in this time frame were less pronounced. However, the highest peak coincided with a video of Orthodox Jewish protesters cutting locks at a Brooklyn playground. Related tweets often mentioned "NYCMayor," the third most often mentioned user during that time period, as can be seen in illustration 2 showing the third most frequent retweet (1,606 times in our dataset from May to August 2020 with the keyword "Jews"). The second highest peak was related to statements by Nick Cannon. The user "NickCannon" was mentioned 1,122 times, the 15th most often mentioned user in that period. Most tweets in the samples with the keyword "Jews" were not antisemitic.

The original queries for the keyword "ZionNazi*" returned only 2,421 tweets. This is used only by a fringe group of Twitter users who are very hostile to Israel and Zionism. We consider this slur that conflates Zionism with Nazism as a strong indication for antisemitism according to the IHRA Working Definition. However, the annotators rated 11.2 and 13.6 percent of samples 7 and 8 as not antisemitic, either because they were calling out the slur or because the message was unclear.

The original queries for the "Israel" samples included a large number of tweets, totaling 1,887,724. Most tweets were on news about Israel but a significant percentage were antisemitic, 10.1 and 13.0 of samples 9 and 10 respectively.

Discussion

Our labeled dataset (a preliminary Gold Standard) of 4,137 tweets is built on representative samples of tweets including the common keywords "Jews" and "Israel" and keywords more likely to be used in antisemitic contexts, "kikes" and "ZioNazi*." It includes 933 tweets (22.55 percent) that are antisemitic according to the IHRA Working Definition of Antisemitism.

The majority of tweets (2,533) come from the "Jews" query. It is reasonable to assume that our dataset reflects discussions on Twitter about Jews well and covers the most prevalent topics, at least when the word "Jews" is directly implicated, and for the time period that the dataset covers: from January 2019 to August 2020.

281 tweets with the keyword "Jews" were rated as antisemitic. It is also reasonable to assume that they cover most relevant topics of antisemitic discussions about Jews on Twitter during that time period.

Additionally, the dataset includes 775 tweets with the keyword "Israel" from the first eight months of 2020, 91 of which were rated as antisemitic. This significantly increases the variety of topics that the corpus covers, however the sub-set of 91 antisemitic tweets with the keyword "Israel" might miss important topics of Israel-related forms of antisemitism. This is supplemented by two samples with the keyword "ZioNazi*" of 532 tweets altogether, with 467 being antisemitic. However, the variety of topics of this subset is very limited, and this word is not used very often by Twitter users as the query results with that keyword show. Future use of the labeled dataset should keep in mind that half of the antisemitic tweets of this dataset (50.1 percent) come from samples with the keyword "ZioNazi*", which is only marginally used, and which covers only a narrow range of topics.

The subset of tweets with the slur "kikes" might be particularly useful because it can help to distinguish antisemitic from non-antisemitic applicates of this string. The assumption that most messages that contain this string are antisemitic is wrong and depends on the context. In our sample, "only" 31.7 percent of such tweets were antisemitic.

Although the labeled dataset covers a large variety of topics, this preliminary gold standard needs to be updated going forward. Initial efforts will consist of adding samples from recent timeframes to account for evolving political or cultural situations and related discourse. Subsequent steps include adding content in other languages. We are currently working on a dataset of Tweets in German. We can also add key terms to expand topical coverage and help alleviate some of the class imbalance in our dataset. Lastly, class imbalance and emerging terms can be addressed by looking at content from known antisemites. As the dataset grows, our coverage of antisemitic users will help catch new slurs and coded language. We hope to be begin adding to the labeled dataset over the next month and publish these additions on GitHub (access upon request). We also provide access to our Annotation Portal to the academic community at https://annotationportal.com and invite scholars to annotate their own samples on that Portal.

Our annotation process seems to be robust. The inter-rater reliability was very good before the annotators discussed their differences in rating. It was almost 100 percent after in-depth discussion and revisiting of the tweets in question. We consider training qualified annotators and the discussion process to be essential to producing an accurate and univocal gold standard.

The gold standard now awaits testing and to facilitate this we will make the labeled dataset available upon request.

Acknowledgements

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. We are grateful that we were able to use Indiana University's Observatory on Social Media (OSoMe) tool and data (Davis et al. 2016).

References

Bruns, Axel. 2020. "Big Social Data Approaches in Internet Studies: The Case of Twitter." In *Second International Handbook of Internet Research*, edited by Jeremy Hunsinger, Matthew M. Allen, and Lisbeth Klastrup, 65–81. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-024-1555-1_3.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20 (1): 37–46. https://doi.org/10.1177/001316446002000104.

Davidson, Thomas, Dana Warmskey, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." In *Proceedings of the Eleventh International Conference on Web and Social Media 2017*, Montreal, Canada. http://sdl.soc.cornell.edu/img/publication_pdf/hatespeechdetection.pdf.

Davis, Clayton A., Giovanni Luca Ciampaglia, Luca Maria Aiello, Keychul Chung, Michael D. Conover, Emilio Ferrara, Alessandro Flammini, et al. 2016. "OSoMe: The IUNI Observatory on Social Media." PeerJ Computer Science 2 (October): e87. https://doi.org/10.7717/peerj-cs.87.

Gambäck, Björn, and Utpal Kumar Sikdar. 2017. "Using Convolutional Neural Networks to Classify Hate-Speech." In Proceedings of the First Workshop on Abusive Language Online, 85–90. Vancouver, BC, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3013.

Gomez, Raul, Gibert Jaume, Gomez Lluis, Karatzas Dimosthenis. 2020. "Exploring Hate Speech Detection in Multimodal Publications." In P*roceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1470-1478

Gwet, Kilem Li. 2008. "Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement." *British Journal of Mathematical and Statistical Psychology* 61 (1): 29–48. https://doi.org/10.1348/000711006X126600.

Jaki, Sylvia, and De Smedt, Tom. 2019. "Right-wing German Hate Speech on Twitter: Analysis and Automatic Detection." ArXiv:1910.07518 [Cs.CL]

Jikeli, Gunther, Damir Cavar, and Daniel Miehling. 2019. "Annotating Antisemitic Online Content. Towards an Applicable Definition of Antisemitism." ArXiv:1910.01214 [Cs.CY], arXiv preprint. https://doi.org/10.5967/3r3m-na89.

Kursuncu, Ugur, Manas Gaur, Carlos Castillo, Amanuel Alambo, K. Thirunarayan, Valerie Shalin, Dilshod Achilov, I. Budak Arpinar, and Amit Sheth. 2020. "Modeling Islamist Extremist Communications on Social Media Using Contextual Dimensions: Religion, Ideology, and Hate." *ArXiv:1908.06520 [Cs],* October. http://arxiv.org/abs/1908.06520.

Malmasi, Shervin, and Zampieri Marcos. 2017. "Detecting Hate Speech in Social Media." ArXiv:1712.06427 [cs.CL] https://arxiv.org/abs/1712.06427.

Landis, J. Richard, and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159. https://doi.org/10.2307/2529310.

Rickenbacher, Daniel. 2021. "The Centrality of Antisemitism in the Islamic State's Ideology and Its Connection to Anti-Shiism." In *The Return of Religious Antisemitism?*, edited by Gunther Jikeli, 93–102. Basel: MDPI. https://doi.org/10.3390/books978-3-03943-498-5.

Ward, Eric. 2017. "Skin in the Game. How Antisemitism Animates White Nationalism." *Political Research Associates*. June 29, 2017. http://www.politicalresearch.org/2017/06/29/skin-in-the-game-how-antisemitism-animates-white-nationalism.

Pustejovsky, James, and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications.* O'Reilly Media, Inc.

Towns, John, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, et al. 2014. "XSEDE: Accelerating Scientific Discovery." Computing in Science & Engineering 16 (5): 62–74. https://doi.org/10.1109/MCSE.2014.80.

Waseem, Zeerak. 2016. "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter." In *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142. Austin, Texas: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-5618.

Waseem, Zeerak, and Dirk Hovy. 2016. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter." In *Proceedings of the NAACL Student Research Workshop*, 88–93. San Diego, California: Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-2013.

Zannettou, S., Finkelstein, J., Bradlyn, B., & Blackburn, J. (2020). A Quantitative Approach to Understanding Online Antisemitism. Proceedings of the International AAAI Conference on Web and

Social Media, 14(1), 786-797. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/7343.