Challenge #1: Data Set Report

Detecting Antisemitic Hate Speech & Conspiracy Fantasies - From Raw Data to Smart Detection

By: Saisha Siram (Portal Manager), Adit Syed (Data Manager), Mark Vinokur (Annotator)

Overview: Social & Ethical Implications (Bonus)

This project focuses on building an annotated dataset of social media posts to support the development of an effective machine learning model for detecting antisemitism online. This work carries important social and ethical implications. While automated hate speech detection has the potential to curb online antisemitism, it must be implemented with care to preserve free expression, avoid political bias, and distinguish between legitimate criticism and hate. By creating a detailed and context-aware dataset, grounded in the IHRA Working Definition of Antisemitism, our goal is to train a machine learning model that is not only accurate but also applicable to real-world moderation, education, and policy settings.

The IHRA's current definition of antisemitism that we trained our ML model with:

"Antisemitism is a certain perception of Jews, which may be expressed as hatred toward Jews. Rhetorical and physical manifestations of antisemitism are directed toward Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities."

Methodology: Social Media Post Scraping

Dataset File: Parsed Tweets for Annotation

When scraping data from social media platforms such as X (previously known as Twitter) using BrightData, our main priority was building a nuanced dataset that proved context towards the evolution of hate speech both before the events of October 7th and after. The following figure depicts filters used through X's Advanced Search Criteria, along with their justifications, to pull meaningful profiles.

Advanced Search Criteria	Justification
Keywords: Jews, Israel **Exact Phrases:** Globalist-Agenda, anti-Jewish, Jewish conspiracy, Holocaust denial **Hashtags:** #Zionists, #HitlerWasRight, #Jews,	Each post was filtered using the general keywords of 'Jews' or 'Israel' to find relevant data. We then used more specific phrasing and hashtags such as "Holocaust", "#Zionist", etc. to ensure our dataset is balanced between antisemitic and non-antisemetic data, posts calling out antisemitism, etc.. Posts and users were then randomly selected for our data set once identified. The use of keywords/hashtags and random selection of posts is done in an effort to reduce sampling bias.

1. https://holocaustremembrance.com/resources/working-definition-antisemitism

#Antisemitism, #Israel, #Hamaz, #Palestine	
Time Frame: • September 1 - September 30 • October 1 - October 31 • November 1 - Nov 30 • Dec 1 - Dec 31 • Jan 1 - October	We collected data from both before October 7th, and after in order to track the evolution of antisemitism. We chose to divide our post sorting using compact time frames which followed soon after the event to further analyze the change in the rhetoric, style, and sentiment expressed through social media. We also chose to expand our time frame far beyond October 7th to get better insight into how other world events have shaped discourse and affected antisemitism in social media.
Minimum View Count: 200	This was done to ensure posts were relevant, not created by bots, and that people actually engaged in the content.

Using this advanced search criteria, we pulled 25 profiles from X, averaging 5 profiles per time period. Then using BrightData, we pulled posts from each user. Lastly we filtered the data using the following code to produce our final dataset ready for annotations.

Methodology: Data Annotations

Annotation Files: Mark Vinokur Annotations; Adit Syed Annotations

Annotations Labels: Our team followed the Annotation Scheme based on the IHRA Working Definition of Antisemitism (IHRA-WDA) provided to us. Few changes were made to this annotation scheme to remain aligned with the IHRA definition of antisemitism.

These labels include: Online, Antisemitic according to IHRA, IHRA Section, Content Type, Calling our Antisemitism. Our full Annotation Guideline can be found here: Annotation Guidelines/Questions

Annotations were conducted by two members of our team to ensure accuracy in how we are categorizing antisemitism, to account for human variance in perception of antisemitism, and to track and quantify interrater reliability. The following section uses statistical measures (Krippendorff's Alpha) to quantify this variance in annotations. Out of the total 300 posts that were included in our data set, each member of the annotation team annotated the first 100 posts, along with a select few posts to create a diverse dataset.

Challenges & Limitations: The main challenge our team faced was maintaining agreement in our annotations. Both annotators had different perceptions of antisemitism shaped by their own unique experiences and cultural backgrounds. The following section, calculating Krippendorff's Alpha, highlights this discrepancy.

Statistical Analysis: Krippendorff's Alpha (Bonus)

2. https://real-statistics.com/reliability/interrater-reliability/krippendorffs-alpha/krippendorffs-alpha-basic-concepts/
3. https://www.k-alpha.org/

IU Datathon 2025 | Team 2: Saisha Siram, Adit Syed, Mark Vinokur

Calculation Work: Krippendorff's Alpha Calculations

Krippendorff's alpha is a statistical measure of inter-annotator agreement that accounts for the possibility of agreement occurring by chance. Unlike simpler metrics like percent agreement, it works with any number of annotators, supports multiple label types (nominal, ordinal, etc.), and can handle missing data—making it well-suited for complex, subjective tasks like identifying antisemitism. We used Krippendorff's alpha to evaluate the consistency of our manual annotations. High alpha values (closer to 1) will indicate that our labels are dependable and that our model can learn from them with confidence. Krippendorff's alpha[2] is defined via the following formula:

$$\alpha = \frac{p_a - p_e}{1 - p_e}$$

Methodology: In order to calculate Krippendorrf's Alpha, we had to transform our entire qualitative dataset into a numerical one. This was done using the key that is shown in the Calculations Work above. For example, 'yes' became 1, 'no' became 0, and so on. We then created an agreement table, comparing both team members' responses. Since Krippendorff's Alpha is only used to evaluate 1 object at a time, we completed this process twice for two different annotation questions. Lastly, we used an online platform for the final calculations, evaluating at a 95% confidence interval.

Results:

Question 1 - Is the post antisemitic according to IHRA-WDA
Answers ranged from numerical values 0-4 indicating the level of perceived antisemitism
Krippendorff's Alpha (Nominal Scale) = .157
95% CI = [.019, .286]

Question 2 - Is the post calling out or reporting on antisemitism?
Answers ranges from 0 to 1 (yes or no)
Krippendorff's Alpha (Nominal Scale) = .171
95% CI = [-.078, .401]

Conclusion: Overall agreement between our annotations are low - further highlighting subjectivity in how different individuals perceive antisemitism. To improve our inter-annotator agreement, we can refine our annotation guidelines with clearer examples and introduce rationale tagging to explain labeling decisions. These steps will help us create a more consistent, reliable dataset for training our model.

2. https://real-statistics.com/reliability/interrater-reliability/krippendorffs-alpha/krippendorffs-alpha-basic-concepts/
3. https://www.k-alpha.org/