2025 Datathon Submission – Challenge #1

Team Name/Tag: MagenCode
Team Members: Giuilio Zuckerman, Asher Rosenfeld, Oriel Atias, Eliana Woolf

1. Definition of Antisemitism

For the purposes of annotation and classification, we adopted the International Holocaust Remembrance Alliance (IHRA) working definition of antisemitism: “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.” We chose this definition because it is widely recognized by international bodies, including the United Nations, European Parliament, and numerous national governments. It provides a comprehensive and nuanced framework for identifying both overt and subtle forms of antisemitism, especially in politically charged or coded language online. Given its clear criteria and broad applicability, it was particularly well-suited for annotating social media content where antisemitic rhetoric may be implicit, embedded or coded in complex and subtle manners.

2. Scraping Method and Rationale

We scraped content from X (formerly Twitter) using the Bright Data platform. Rather than focusing on hashtags, user groups, or predefined topics, we targeted specific keywords that commonly appear in conversations where antisemitism may surface: “Jews”, “genocide”, “Israel”, “Palestine”, and “Gaza”. This ensured broad coverage and inclusivity, avoiding bias from hashtag- or group-based scraping. These keywords were selected because they are often used in contexts that can carry antisemitic undertones, particularly in global political discourse. In total, we scraped 355 tweets, aiming to include both antisemitic and non-antisemitic content for balanced training and evaluation.

3. Annotation Guidelines and Classification

We adopted the classification schema provided by the competition organizers, aligned with the IHRA definition. The schema offered detailed guidance for distinguishing antisemitic, borderline, and non-antisemitic content, helping ensure consistency in annotations. It also allowed annotators to flag ambiguous cases for further review.

4. Inter-Annotator Agreement (IAA)

To assess annotation reliability, a subset of 30 tweets (IDs 0–30 out of 355) was double-annotated by two independent team members. We calculated Cohen’s Kappa score, which was 0.54, indicating a moderate level of agreement. Discrepancies likely arose from subjective interpretations of borderline cases or ambiguity in the guidelines, suggesting the need for refined definitions or additional training.