Challenge #2

Detecting Antisemitic Hate Speech and Conspiracy Theories: From Raw Data to Smart Detection

By: Saisha Siram, Adit Syed, Mark Vinokur (Team 2)

Introduction: Training and Evaluation of ML Models in Detecting Antisemitic Content

Following our initial work creating and annotating our sample dataset of scraped social media posts in Challenge 1, the focus of Challenge 2 was to build and evaluate a machine learning model capable of detecting antisemitic hate speech on X (formerly Twitter). Through the use of IU's professionally pre-annotated datasets grounded in the IHRA's Working Definition of Antisemitism, we aimed to develop and fine-tune a system that could generalize well across both overt and subtle expressions of antisemitism. Our approach attempts to balance ethical considerations such as avoiding censorship of legitimate criticism, alongside with technical priorities like model robustness and classification accuracy.

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."

Dataset Overview

We used two curated, pre-annotated datasets: *"Antisemitism on Twitter: A Dataset for Machine Learning and Text Analytics"* and *"Antisemitism on X: Trends in Counter-Speech and Israel-Related Discourse before and After October 7."* Both datasets include real posts from X filtered using keywords decided by annotators associated with ISCA, including terms like: "Israel", "Jews", and more explicit slurs or conspiracy-related phrases such as "ZioNazi". Importantly, these posts were annotated by several reviewers, with final labels reflecting inter annotator agreement. The datasets include binary antisemitic/non antisemitic classifications, with some posts further categorized into subtypes calls to violence, conspiracies, or Holocaust distortion.

Training Methodology

Selected Model: Our team selected the [twitter-roberta-base-offensive] model from HuggingFace as it is pre-trained on content moderation tasks involving offensive language found on X (formerly Twitter). We believe this model would offer high best accuracy and would be particularly effective for analyzing social media posts on X.

Method of Fine-Tuning: Our team used PyTorch to fine-tune our model.

Hyperparameters	Purpose
Batch Size: 8	Smaller batch sizes can lead to better generalization, i.e. better performance on unseen data
Epochs: 3	Multiple epochs allows the model to update the weights over the runs to optimize learning
Weight Decay: .01	Lower decay helps prevent overfitting and improve generalization
Learning Rate: 2e-5	Smaller learning rate leads to less loss throughout the fine-tuning process

Evaluation: Model runs and metrics were monitored and recorded with Weights & Biases: AI Developer Platform

Model Evaluation

Our team ran two trials of fine-tuning and testing the model to see which produced better accuracy

Trial 1: 'Antisemitism on X: Trends in Counter-Speech and Israel-Related Discourse Before and After October 7' Dataset

```
=== Classification Report ===
              precision    recall  f1-score   support

not_offensive       0.69      0.97      0.81       184
    offensive       0.83      0.27      0.41       108

     accuracy                           0.71       292
    macro avg       0.76      0.62      0.61       292
 weighted avg       0.74      0.71      0.66       292

=== Confusion Matrix ===
                   pred_not_offensive  pred_offensive
true_not_offensive                 178               6
true_offensive                      79              29
```

Trial 2: Small Sample from 'Antisemitism on Twitter: A Dataset for Machine Learning and Text Analytics' (Gold Standard)

```
=== Classification Report ===
              precision    recall  f1-score   support

not_offensive       0.90      0.96      0.93        89
    offensive       0.33      0.18      0.24        11

     accuracy                           0.87       100
    macro avg       0.62      0.57      0.58       100
 weighted avg       0.84      0.87      0.85       100

=== Confusion Matrix ===
[[85  4]
 [ 9  2]]
          pred_not  pred_off
true_not        85         4
true_off         9         2
```

Trial 1 achieved an overall accuracy of 71%, demonstrating strong performance in identifying non-offensive posts (recall = 0.97), but struggled significantly with detecting offensive content (recall = 0.27). The model struggled to correctly identify offensive posts, and misclassified 79 out of 108 offensive posts as non-offensive. In contrast, Trial 2 reached a higher overall accuracy of 87% and maintained strong performance on non offensive posts (precision = 0.90, recall = 0.96). However, its ability to detect offensive content declined further compared to Trial 1, with a recall of just 0.18 and an

F1 score of 0.24. It's important to note that this result is based on a very small number of offensive examples (n=11), making our evaluation for offensive detection less reliable.

Qualitative Analysis: False Positives and Negatives

Sample False Positives (non-offensive labeled offensive):

Example 1: 'As a Jew, I must say fuck the Azov battalion supporting ADL.'

Example 2: This top black men an jews might be trying to tame white tyrants and end up raping them. It happens to girls in the south that charm things the town thinks they haven't earned. Nobody really believes in true love with group think.

Example 3: Israel Catches Terrorist from Murderous 2010 Attack on Gaza Fence | United with Israel https://t.co/xZUDryuMa0 BH, our enemies will always be captured and punished severely! Hang the swine by his ankles till the blood fills his head and explodes!!!!!!

Sample False Negatives (offensive labeled non-offensive)

Example 1: "The Protocols" and "The International jew" were forewarning agreements provided by the jews, notifying of the takeover. The same script from the Torah (in re to Canaan), nothing new. Muh 'they were right!' is the proof for the mission accomplished.

Example 2: 'The alleged Solomon's Temple does not even exist, it's a lie that only Jews believe it'

Example 3: 'Israel more like is fake'

Error Analysis

The model tended to produce a high number of false negatives for both trials. This can be attributed to the model's reliance on detecting explicit antisemitic language, such as violent threats or overt slurs. However, much of the antisemitism in our datasets was more nuanced, often expressed through coded language, sarcasm, conspiratorial framing, or dehumanizing generalizations without the use of obviously offensive terms. As a result, the model failed to flag many instances of harmful content that did not match the explicit patterns it had learned, highlighting the need for improved representation of subtle and implicit antisemitism during training.

Limitations and Future Work

Limitations

- Recall for "offensive" class is critically low (0.27 vs 0.97 for non-offensive).

- The model shows bias toward predicting "not offensive" (178 correct vs only 29 correct for offensive).
- Clear and overt antisemitic content being misclassified (e.g., "Evilest Jews" list with only 24% probability).
- Dangerous under-detection of hate speech.
- Some legitimate political speech flagged (e.g., criticism of ADL with 51% probability). Potential chilling effect on free expression

Future Work

- Prioritize samples similar to current false negatives. Include more edge cases of "coded" hate speech.
- Improve cross-linguistic detection.
- Multi-task learning to simultaneously detect hate speech categories.
- Separate precision/recall metrics for demographic groups (Jewish, Black, LGBTQ+ references).
- Secondary classifier to distinguish criticism from hate speech (e.g., "ADL policies" vs. "Jewish control").
- Explanations for high-stakes predictions (SHAP/LIME outputs).

Bonus: Unseen Data

```
=== Classification Report ===
              precision    recall  f1-score   support

not_offensive       0.93      1.00      0.96        38
    offensive       0.00      0.00      0.00         3

     accuracy                           0.93        41
    macro avg       0.46      0.50      0.48        41
 weighted avg       0.86      0.93      0.89        41

=== Confusion Matrix ===
                   pred_not_offensive  pred_offensive
true_not_offensive                  38               0
true_offensive                       3               0

Total examples: 41
False positives: 0
False negatives: 3
```

We retested our model using unseen data from our previously annotated dataset developed in Challenge 1, that was reannotated using the same label scheme. This dataset was small, containing approximately 40 posts. As shown, our model continues to perform well in identifying non-antisemitic posts. Since the dataset was incredibly small, and only contained 3 posts labelled as offensive, it is difficult to determine our models performance in identifying antisemitic posts from this dataset.

Conclusion

The model demonstrates a strong performance in identifying non-offensive content, achieving a 97% recall. However, it struggles considerably with detecting hate speech, correctly identifying only 27% of truly offensive examples. This disparity is reflected in the 79 false negatives, which included primarily antisemitic stereotypes, conspiracy theories, and dehumanizing rhetoric that the model failed to flag and recognize as harmful.

At the same time, the 6 false positives reveal oversensitivity to political statements and broader critiques that, while potentially provocative, are not inherently antisemitic. Although the model achieves

a 71% overall accuracy, this metric conceals serious limitations as the model fails to detect 73% of offensive content while incorrectly flagging legitimate discourse. These results suggest the model, in its current form, is not suitable for standalone deployment. However, it may serve effectively as a first-pass filter, provided that all “non-offensive” classifications involving identity-related topics are subject to mandatory review by a human moderator.