ROBERT A. AND SANDRA S.
BORNS JEWISH STUDIES PROGRAM

INSTITUTE FOR THE STUDY OF CONTEMPORARY ANTISEMITISM

Detecting Antisemitic Hate Speech & Conspiracy Fantasies – From Raw Data to Smart Detection – July 2025

Overview: In this datathon, your team will tackle two interconnected challenges that reflect real-world tasks in hate speech research: the creation of a labeled dataset and the application of machine learning techniques to detect hate speech. Your coordination, documentation, and critical reflection on the results will be assessed and evaluated by experts in this field.

#1 Challenge: Building and Annotationg a Dataset

Goal: Create a small but meaningful labeled dataset for hate speech detection.

Tasks:

1. **Data Collection:** Use the *Bright Data*[1] interface to scrape a minimum of atleast 100 relevant posts of user-generated content (e.g., tweets, posts, or comments) from X (fomerly known as Twitter).

2. **Sampling and Documentation:** Decide on your scraping focus (e.g., specific hashtags, keywords, seed list or user groups) and document your strategy. Describe how you targeted the issue, the motivations behind your decision, and why this content is relevant and potentially antisemitic. Keep in mind that your dataset should include antisemitic and non-antisemitic content. Machine Learning Models need labeled datasets that include both.

3. **Annotation:**

   - Apply a definition of antisemitism to annotate your dataset.
   - Apply a standardized annotation form (either use our scheme as prototype, adjust it or create your own scheme based on the type of content you are interested in).[2]
   - We recommend using our online *Annotation Portal*[3] for labeling.

4. **Deliverables include:**

   - A cleaned and annotated dataset, exported in `.csv` format, including only posts containing your selected keyword or hashtag.
   - A short dataset report (1–2 pages) that: Defines the labels used during annotation (e.g., antisemitic/non-antisemitic).

[1] https://brightdata.com/products/web-scraper/functions

[2] https://github.com/AnnotationPortal/DatathonandHackathon.github.io/blob/main/guides/annotation_scheme.md; see also: https://arxiv.org/abs/1910.01214

[3] https://annotate.osome.iu.edu/

- Describes the keyword/hashtag and time period selected, and why Summarizes the distribution of labeled examples.
- Explains how annotations were conducted and by whom.
- And Includes any challenges or limitations (e.g., platform constraints, ambiguous posts).[4]

#2 Challenge: Modeling and Evaluation

Goal: Use our pre-annotated gold standard dataset to build and evaluate a system to detect antisemitism.

Tasks:

1. **Dataset Access:** Download the provided annotated datasets:

   Antisemitism on Twitter: A Dataset for Machine Learning and Text Analytics[5]

   Antisemitism on X: A Dataset Tracking Trends in Counter-Speech and Israel-Related Discourse Before and After October 7[6]

2. **Modeling:** Use transformer-based models to build a classification system for detecting antisemitic content. This includes not only explicit hate speech but also coded language, conspiratorial narratives, Holocaust distortion.

   - We recommend fine-tuning a model from Hugging Face Transformers. You may choose from general-purpose or task-specific models:
     - `[twitter-roberta-base-offensive]`[7] — pre-trained on offensive language from Twitter/X
     - `[microsoft/mdeberta-v3-base]`[8] — strong multilingual classification baseline
     - `[bertweet-base]`[9] — designed for Twitter-style texts
     - `[GroNLP/hateBERT]`[10] — pre-trained on hate speech corpora from Reddit
   - Fine-tune your model using tools like Hugging Face's `Trainer` API or custom PyTorch code.
   - Use a standard train/test split (e.g., 80/20), and optionally a validation set (e.g., 10% of training) for early stopping or tuning.
   - Use a fixed random seed for reproducibility.

3. **Evaluation:** Your submission should include:

   - Standard classification metrics: **precision**, **recall**, **F1-score**, and a **confusion matrix** (on the test set).

[4] Our descriptions of the datasets published on Zenodo may be useful as template: https://zenodo.org/records/14448399

[5] https://zenodo.org/records/14448399

[6] https://zenodo.org/records/15025646

[7] https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive

[8] https://huggingface.co/microsoft/mdeberta-v3-base

[9] https://huggingface.co/docs/transformers/en/model_doc/bertweet

[10] https://huggingface.co/GroNLP/hateBERT

- A complete list of **hyperparameters** used (e.g., learning rate, batch size, epochs, weight decay).
- A short **error analysis** describing what types of posts your model struggles with (e.g., satire, rhetorical questions, mixed sentiment or transliterations).
- 3–5 **qualitative examples** of false positives and false negatives, with a brief explanation of why the model may have misclassified them.

  Note: *While prompt-based classification using large language models (e.g., ChatGPT, GPT-4) is a popular trend, this task focuses on* ***training and evaluating reproducible models****.*

#3 Resources and Structure

- A live tutorial will introduce you to our *Annotation Portal* and *Bright Data*.
- Breakout sessions will follow for Q&A, team coordination, and hands-on work.
- At the end of the datathon competition, each team will submit a report on their workflow and insights, including scripts and documentation of used tools.

#4 Scoring & Evaluation

Each team can earn a maximum of **100 points**, with an additional **20 bonus points** available for optional advanced tasks.

- **50 points – Task Performance:** Assessed based on the quality of data scraping, annotation consistency, and model performance on the evaluation dataset (e.g., precision, recall, F1-score).
- **50 points – Report and Presentation:** Awarded based on the clarity, structure, and completeness of the submitted report, insights into methodology, annotation rationale, and evidence of teamwork and collaboration.
- **+20 Bonus Points – Optional Advanced Contributions:** Bonus points may be awarded for:
  - Using advanced evaluation metrics such as **Krippendorff's Alpha** or **Cohen's Kappa** to report inter-annotator agreement
  - Applying your model to additional **unseen data** beyond the test set
  - Addressing the **social and ethical implications** of antisemitism detection and hate speech classification

The jury includes experts from machine learning, hate speech research, and digital humanities.

Any Further Questions? *Don't hesitate to reach out!*

For questions or technical support, reach out to the organizing team: Dr. Daniel Miehling (damieh@iu.edu), Prof. Gunther Jikli (gjkeli@iu.edu), and Rachel Kelly (rk18@iu.edu).

Remember: The competition not only demands curiosity, but also strong teamwork and critical thinking to succeed!