Antisemitism Detection in Social Media

Damir Cavar

Indiana University

October 2024

Agenda

- About
- Research Resources
- Goals
- Methods

About

- [Damir Cavar](#)
  - Associate Professor at Indiana University
    - Natural Language Processing (NLP), see [NLP-Lab](#)
    - AI
    - Quantum Computing (NLP, AI)

Literature and Resources

- Textbook
  - Jurafsky and Martin "Speech and Language Processing"

- Publications, a.o.:
  - Semantic Scholar – antisemitism or hate speech
  - arXiv – hate speech or antisemitism
  - ACL Anthology - antisemitism

Literature and Resources

- Code and Data
  - Datathon/Hackathon [2024 GitHub repo](#)
  - GitHub repo [dcavar](#) – [Python tutorials](#)
  - GitHub repos on [antisemitism](#)
  - Kaggle
    - [https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset](#)

Goals

- Process tweets (and other texts)
- Machine learning of classifiers
  - Bayesian models
  - Logistic Regression
  - Transformer Models
    - BERT (and derivatives)
- Large Language Models and Generative AI
  - OpenAI's GPT
  - Anthropic's Claude 3.x
- Word embeddings for ML Classifiers

Practical Tutorial

- Using Python
  - Jupyter Notebooks
  - API Endpoints provided by OpenAI, Anthropic, Hugging Face, VoyageAI
  - Sample data for general NLP, sentiment analysis, text classification, …