INDIANA UNIVERSITY
BLOOMINGTON

Automated Hate Speech Detection The Importance of Precise Datasets Including a Calling-Out-Bias Label

Damir Cavar, Gunther Jikeli, Sameer Karali

NLP-Lab & Observatory on Social Media, Indiana University Bloomington

Background

- **Automated hate speech detection:**
  - significant strides in recent years
  - use of Deep Learning techniques and large-scale training data
  - ML algorithms and DL networks employed to detect hate speech with high accuracy in test data-sets
  - language models like BERT, ELECTRA, Perspective API, and Topic Modeling have been developed to examine large data-sets containing toxic speech patterns and conspiracy-related content

- Hate speech detection: a challenging task for to several reasons
  1. **Datasets** on which models are trained are relatively small and do not encompass all variations of hate speech manifestations
  2. **Hate speech** lacks a common definition or uniform taxonomy
  3. **Hate speech** often includes a high degree of subjectivity, depending on cultural, social, and historical factors, making it difficult to identify and classify consistently
  4. **Correct classification** often requires more context than what is readily available, such as previous discussions in a thread or a history of ironic messages by particular users, leading to false positives.
  5. **Difficulties with AI models:** As demonstrated in a test with ChatGPT, the model correctly identifies antisemitic stereotypes within a message, it also classifies the entire message as antisemitic.

- History
  - Development of Annotation Portal for Social Media
  - Collection and Annotation of large number of social media posts
  - Two International Datathons and Hackathons, 2020 and 2023

Results International Datathon and Hackathon Competition on Hate Speech

- The teams received two labeled data-sets of tweets as follows:

1. *[5401] Messages with the keywords Asians, Blacks, Jews, Latinos, or Muslims classified as biased/non-biased and Calling Out biased/ not Calling Out biased (based on 75 percent of annotators' agreement)*
2. *[6153] Messages with the keywords "Jews, Israel, Kikes, or Zionazi*" classified as biased/non-biased and Calling Out biased/ not Calling Out biased (based on 100 percent of two annotators' agreement)*

- Teams conducted independent evaluations with objective of characterizing the quality, functionality, and performance of the solution to classify tweets and social media messages along two binary variables:
  - Bias
  - Calling out
    distinguishes biased tweets from those that call out bias, which is a particular challenge for automated detection
- goal: for the two data-sets, one solution separately for each of the two data-sets + one solution merging the two data-sets.

- 35 students from 10 countries were accepted into the program.
- They competed in teams for accurate annotation in the datathon and prediction of hate speech in the hackathon.
- Evaluation: in cooperation with IU's Data Science Club.

Results Datathon/Hackathon 2023

- **Six submissions from the six teams**
  1. One submission involved building two versions of a model and utilizing data pre-processing and data augmentation techniques using WordNet
     – Created three models that predict biased and calling out behavior across all six types of data
     – Prepared a GUI for efficient data handling.
  2. One submission utilized data pre-processing and data augmentation techniques using NLP
     – Created two models that predict biased and calling out behavior for all six types of data.
  3. One submission utilized data pre-processing and data augmentation techniques using NLP
     – Created only one models that predict biased behavior for one data set.

References

Jikeli, G., D. Cavar, D. Miehling (2019) *Annotating Antisemitic Online Content. Towards an Applicable Definition of Antisemitism.* arXiv:1910.01214

Jikeli, G., D. Cavar, W. Jeong, D. Miehling, P. Wagh, D. Pak (2022) Toward an AI Definition of Antisemitism? In M. Hübscher, S. von Mering (eds.) *Antisemitism on Social Media*, Routledge.

Jikeli, G., D. Axelrod, RK. Fischer, E. Forouzesh, W. Jeong, D. Miehling, K. Soemer (2022) Differences between antisemitic and non-antisemitic English language tweets. *Comput Math Organ Theory.* 2022 Sep 9:1-35.

Links

2020 Antisemitism Datathon and Hackathon event challenges students to grow socially and technologically and 2023 Datathon and Machine Learning Competition

Acknowledgements

Special thanks to Swaminathan, Paveethran and Patil Harshwardhan, members of the Data Science Club at IU Bloomington and to the participants of the Hackathon Datathon Competition.