Title: Implicit Bias in Hate Speech: Reading Between the Vectors
Author: Merrilee Montgomery

The Problem: Hate speech is defined as abusive or threatening speech or writing that expresses prejudice on the basis of ethnicity, religion, sexual orientation, or similar identifying grounds. Because hate speech online can contribute to violent crime offline [2,4] and can constitute sufficient threat for hate crime [5], methods to measuring hate speech online should be explored more. Most NLP studies of hate speech on the internet have focused on identifying *explicit hate speech*, which is usually identified as containing key phrases and words. However, hate speech can also be *implicit* using coded or indirect language such as sarcasm, metaphor, or circumlocution to promote prejudicial views. This study focuses on the latter category of hate speech. This study seeks to identify a quantitative difference between implicit hate speech and non-hate speech texts using vector representations of words.

The Data: This study compares implicit hate speech and non-hate speech data collected by El-Sherief et al.[3] The implicit hate speech data was collected from Twitter hate groups posts. The implicit hate speech is categorized as promoting white grievance, incitement to violence, inferiority language, irony, stereotypes/misinformation, and threatening/intimidation. These seven categories are referred to as "hate types" in this study.

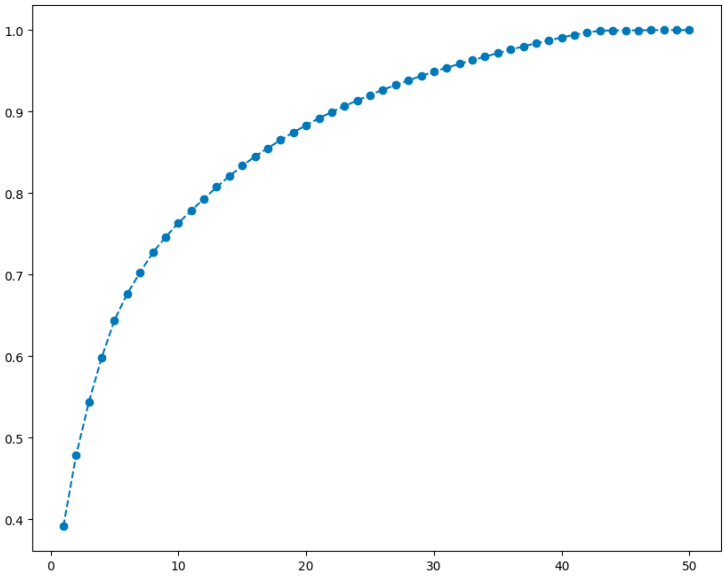
The Methodology: This study hopes to identify a quantifiable difference between hate speech and non-hate speech. After trying GLOVE embeddings of these words, I have concluded that using a transformer, such as Sentence-BERT [1] would be better. A transformer would better lend itself to comparing syntax and in general has more information for embedding entire sentences. BERT reads based on tokens, making it better suited to social media text, which uses informal terminology often.

I will use pytorch and the sentence models from sbert.net for the classification, and pandas, matplotlib.pyplot, and seaborn for data management and visualization. I will use Sentence-BERT to generate vector embeddings of implicit hate speech and control non-hate speech. This study will experiment with dimension reduction as needed and, beyond simply identifying hate speech, classify it according to hate type. Finally, this study will scrape posts from the subreddit "r/NationalFascism" and take text from AP News to try to use hate speech detection to determine which source a text passage is from. While tweets are short, these passages are longer. I am curious to see how this compares to the tweets dataset.

Intermediate/Preliminary Experiments

For each implicit hate post, I averaged all GLOVE-embedding vectors associated with each word across words such that each post was measured with 50 vectors. I did Principle Component analysis across those fifty metrics for seven components and found that the brute force-ish mean of all vectors for each sentence did not lend itself as nicely to principle component analysis as I had hoped. The graph below shows the number of components on the x axis corresponding to the ratio of variance explained on the y-axis.

I also attempted to do k-means clustering on the implicit hate post data based on these principal components. This k-means clustering did not work well, as it used Euclidean clustering instead the cosine clustering that is standard for comparing embeddings. I had a far-fetched hope that the implicit hate speech data could cluster around the hate types. It did not. I would like to try again using S-BERT encodings of sentences, implementing a cosine clustering algorithm on top of Sklearn as needed.



Division of Labor: I (Merrilee Montgomery) am doing this project alone.

Related Research:

Hertzbergm, N., Sayeed, A., Breitholtz, E.,Cooper, R., Lindgren, E., Rettenegger, G., Ronnerstrand, B. "Distributional properties of political dogwhistle representations in Swedish BERT." *Proceedings of the Sixth Workshop on Online Abuse and Harms,* pp. 170-175. 2022. Association for Computational linguistics.

This study uses BERT to generate an either 3 parameter or 768 parameter representation of the survey responses. These representations were then clustered to determine whether in-group survey responses that correctly responded to the dog whistle prompts were and out-group survey responses that did not clustered separately. This study showed that the BERT-generated representations of the survey responses could be separated. I will be using incorporate a modified Sentence-BERT generated representation of text.

Magu, R., Joshi, K.m Luo, J. "Detecting the Hate Code on Social Media." *Association for the Advancement of Artificial Intelligence.* https://arxiv.org/pdf/1703.05443.pdf

This study focuses on codewords specifically used to refer to groups of people for derogatory purposes without revealing that purpose is hate speech. This study used an SVM to classify documents containing these codewords as either derogatory or not. The SVM stuggled with acronyms and hashtags. My study will also be classifying semi-ambiguous text. I am choosing a type of BERT model in the hopes that tokenization will make the hashtags more recognizable for the model.

Mussiraliyeva, S., Omarov, B., Bolatbek, M., Ospanov, R., Baispay, G., Medetbek, Z., Yeltay, Z. "Applying Deep Learning for Extremism Detection." *Communications in Computer Science and Information Services*, vol. 1393. 2021. https://link.springer.com/chapter/10.1007/978-981-16-3660-8_56

This study points out that social media is a particularly useful source of data for predicting extremism, as social media has become a tool for extremist groups to communicate and to recruit. This study uses a Long Short-Term Memory (LSTM) network and a Convolutional Neural Network (CNN) to evaluate extremism in social network. This study uses

a simple binary classification text and a sigmoid function to separate the two classes during the prediction phase. However, I am curious what the gradient of ideology strength would be without the sigmoid function. While the sigmoid function forces separation between extremist and non-extremist samples, I am curious what patterns a simple clustering algorithm would reveal. I will employ clustering to try to identify the types of hate speech.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. 2021. Blow the Dog Whistle: A Chinese Dataset for Cant Understanding with Common Sense and World Knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2139–2145, Online. Association for Computational Linguistics.

This study adapts the game design for the board game *Decrypto* to assess the ability of four pretrained language models to understand semantic similarity. This study compared BERT, RoBERTa, Baidu ERNIE and ALBERT model performance to human performance in the online *Decrypto* game. The *Decrypto* game is a word guessing game played in two teams that, in single blind tasks, guess the order of a set of words based on related cue words. In double blind tasks, the teams must guess the opponents' words based on the track record of cue words that the opponents' team is using. This study used a pool of general words. I would like to adapt the game structure of this study to assess phrases that express an ideology, but I may not have time.

Gaikwad, M., Ahirrao, S. Kotecha, K., Abraham, A. "Multi-Ideology Multi-Class Extremism Classification Using Deep Leaning Techniques." *IEEE* . Vol 10. 2022. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9885202

Gaikwad et al. assemble a comprehensive dataset of extremist language from journal papers, newspapers, websites, and tweets. This dataset represents both Islamic-jihad and white-supremacist extremism. Uniquely, this study focuses on multi-class labelling beyond simple "extreme, non-extreme" or "extreme, non-extreme, neutral" categorizing. Instead of a gradient of extremism, this study uses BERT and BERT-model variants to separate text into propaganda, radicalization, and recruitment. While I like the idea of multi-class labelling, my study will approach hate speech as an expression of thought first, rather than propaganda. I will separate text by hate speech type using clustering as my multiple classes.

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., Choudhury, M., Yang, D. "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech." *UC San Diego, Georgia Institute of Technology.* 2021. https://arxiv.org/pdf/2109.05322.pdf

This study compares defines implicit and explicit hate speech and notes that most datasets of hate speech only contain explicit hate speech. This study provides a new dataset, which I will use in this study. This study used SVM and other forms of BERT, but not Sentence-BERT. I hope to build on the data and theoretical framework that this study provides

Reimers, N., Gurevych, I. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." *Ubiquitous Knowledge Processing Lab Department of Computer Science*, *Technicsche Universitat Darmstadt*. 2019. https://arxiv.org/pdf/1908.10084.pdf.

This is the BERT model that I will be using. It implements a pooling function after the BERT embedding to produce embeddings of the same length for multiple sentences so that sentences can be compared to each other.

Timeline: By deadline:
March 30: Load Sentence-BERT with data, save embeddings
April 6: Determine level of dimension reduction that is most useful for classifying hate speech.
April 13: Attempt Clustering, collect longer passages (AP News, "r/NationalFascism")
April 20: Classify longer passages
April 28: Write report

References:

1. Reimers, N., Gurevych, I. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." *Ubiquitous Knowledge Processing Lab Department of Computer Science*, *Technicsche Universitat Darmstadt*. 2019. https://arxiv.org/pdf/1908.10084.pdf.
2. Margolin, J., Pezenik, S. "It's become more difficult to identify motivations behind mass casualty attacks." *ABC News,* 2024. https://abcnews.go.com/US/become-increasingly-difficult-identify-motivations-mass-casualty-attacks/story?id=106338758
3. ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., Choudhury, M., Yang, D. "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech." *UC San Diego, Georgia Institute of Technology.* 2021. https://arxiv.org/pdf/2109.05322.pdf
4. Cahill, M., Taylor, J., Williams, M., Burnap, P., Javed, A., Liu, H., Sutherland, A. "Understanding Online Hate Speech as a Motivator and Predictor of Hate Crime." *U.S. Department of Justice Office of Justice Programs.* 2019. https://www.ojp.gov/library/publications/understanding-online-hate-speech-motivator-and-predictor-hate-crime
5. "Online Extremism: More Complete Information Needed about Hate Crimes that Occur on the Internet." *U.S. Government Accountability Office.* 2024. https://www.gao.gov/products/gao-24-105553