Hate Speech on Social Media

Merrilee Montgomery

May 2, 2024

1 Abstract

Hate speech has become of concern online as criminally threatening, as a possible predictor of violent activity, and as an important contributor to radicalization. However, most studies have focused on *explicit hate speech* which clearly states the bias and intentions of the speaker, and have focused on identifying hate speech based on key derogatory phrases. Study is needed of *implicit hate speech*, which requires semantic and contextual understanding of language, making it more difficult for computers to identify and making it less likely for a speaker to be censored. Past study has classified hate speech in a variety of ways, including as implicit or explicit. This study acknowledges that hate speech may exist on a gradient from implicit to explicit and compares performance of a simple classification model to performance of two different regression models to determine whether hate speech can be understood as a gradient rather than discrete classes, with the possibility that some text's position on that gradient is a function of speech in context. This study finds that the model of text that considers both speech and context performs better than a simple linear regression and that its hidden representations of speech and context converge to be of equal magnitude and opposite signage. Overall, this study finds that the simple neural network classifier out performs the linear regressions, suggesting that hate speech is indeed discrete and unordered in categorization.

2 Introduction

The advent of social media in the 21st century has created a new space in which people express their opinions to include bias, hatred, and extremism. 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. Hate speech can contribute to radicalization, which can manifest as violent

activity offline [1,5]. Sufficiently threatening speech online can constitute grounds for hate crime charges [8]. For this reason, the security community, to include law enforcement, should be concerned with detecting hate speech online and evaluating how threatening it is.

Most natural language processing studies of internet hate speech have focused on identifying *explicit hate speech*, or speech that directly claims or clearly states the bias. This can be achieved by identifying certain key words and phrases. However, hate speech can also be *implicit*, forcing the reader to assume the bias in order to understand what is being said rather than clearly stating the bias. Implicit hate speech can be identified by use of coded or indirect language such as sarcasm, metaphor, or circumlocution to promote prejudicial views without overtly stating them. Implicit hate speech is uniquely difficult to identify. Implicit hate speech requires a semantic understanding of language and cannot be identified by simply identifying certain words. As social media platforms become better at identifying explicit hate speech, users may shift to implicit language to avoid censorship. For this reason, implicit hate speech presents a unique challenge for natural language processing while being important to the security community as a subset of hate speech.

3 Background

Identifying hate speech can be reduced to a classification task, which requires a vector embedding of the text to be classified. Past study has used SVM [2,6] and various BERT-derived [2,3,4] embeddings of sentences. It was noted [6] that models using SVM embeddings had difficulty interpreting irregular tokens such as acronyms, hashtags, and euphemisms that were not already in the SVM library.

Past studies have used specific derogatory words and phrases as indicators of hate speech [4,6]. Hertzbergm et al [4] collected survey responses to dog whistle prompts, embedded the responses with BERT, and found that distinct clusters emerged of those responses from individuals who clearly knew the derogatory meaning of the dog whistles and those individuals who did not. Magu et al [6] used a Long Short-Term Memory (LSTM) network and a Convolutional Neural Network (CNN) to try to separate texts containing phrases that could be either benign terms or derogatory references to certain people groups. However, approaches to hate speech recognition that rely on certain key phrases simply pushes hate speech and detection algorithms into an arms race, where new phrases can be invented to evade the algorithm until the algorithm catches up.

Beyond simply identifying hate speech [6,7], past study has also attempted to sort extremist speech by ideology [3] and into the explicit vs implicit categories [2]. El-Sherief et al [2] was a primary inspiration for this study. El-Sherief identifies the hallmarks of implicit hate speech to be the following: white grievance, incitement to violence, inferiority language, irony, stereotypes and misinformation, and threatening and intimidation. While some of these hallmarks are often present in explicit hate speech,they but are not always as obviously stated or clearly targeted in implicit hate speech. El-Shereif found that the BERT encoding of text consistently outperformed the SVM encodings. El-Sherief also identified the most common causes of mistakes in implicit hate identification. Models could not understand the semantics of coded hate symbols. Models ended up associating identity terms that were legitimately neutral tokens (such as *Jew* and *Black* with hate speech because they were used within hate speech so frequently. Finally, models had difficulties grasping relationships between statements used together to imply the overarching biased premise.

4 Approach and Experiment

This study seeks to identify a quantitative difference between implicit hate speech and non-hate speech Twitter posts. These posts were collected and classified by El-Sherief [2]. This study generates 384-length vector representations of these Twitter posts using Sentence-BERT [9]. Sentence-BERT was chosen for embedding the Twitter posts for two reasons. BERT's token-level parsing of words and sentences will hopefully make for a more robust model with better capacity to understand the informal words, hashtags, acronyms, and euphemisms frequently used in online speech.

This study trained three different models for classifying the Twitter posts as not hate speech, implicit hate speech, or explicit hate speech. Each of the three models corresponded to some conceptual model of the relationships and barriers between benign, implicit hate, explicit hate texts. The first model was a simple probabilistic classifying neural-network consisting of a single 384x3 weights matrix parameter. The 384-length vector representing a text is passed through the weight matrix to get three scores, to which the Softmax function is applies to generate class probabilities. This first model corresponds to an understanding of these three categories of speech whereby all three can be independently evaluated. The text is assigned to the highest probability class.

The second model and third models test linear gradients of hate speech from not hate to implicit hate to explicit hate. While explicitness of the hate speech does not directly correspond to threateningness, implicit hate speech is so coded to obscure the threateningness of the statement. So, while threateningness of intention may not change, threateningness of the language may. For example, saying, "We will end them all soon" is an explicit threat and, depending on context, could institute threat of harm sufficient for a criminal charge. However, saying, "They will all be ended soon" creates a level of obscurity. For these regression models, the tweets were given scores of 0 (not hate), 1 (implicit hate), or 2 (explicit hate). The second model was a simple linear regression for which bias and Beta coefficients were learned to predict a score, for which a class was assigned based on which values (0,1, or 2) it was closest to. In this way, explicitness is measured on a simple gradient. For model three, Betas and biases were learned for two separate regressions. Two scores were learned for each tweet. The final score used for classification was calculated by subtracting the former from the latter. The intention for this third model is to calculate the explicitness of the speech as $explicitness = speech - context$

The training set consisted of 3000 labeled Twitter posts evenly distributed between the three classes. All models were trained over 1000 epochs on training sets of 300 randomly selected records. All models were tested on a nearly even (88 explicit hate; 89 of the other two) split set of 266 records that were kept the same across all models, and that the models had not seen before. The limiting factor for these testing and training sets was the number of explicit hate texts. There were only 1088 explicit hate texts in the corpus. The learning rate was initialized to 0.0001 for all models, but a scheduler was set to adjust the learning rate by $gamma = 0.3$ **every 100 steps. Finally, the Adam optimizer was used, and was set to perform back-propagation every epoch.** [H]

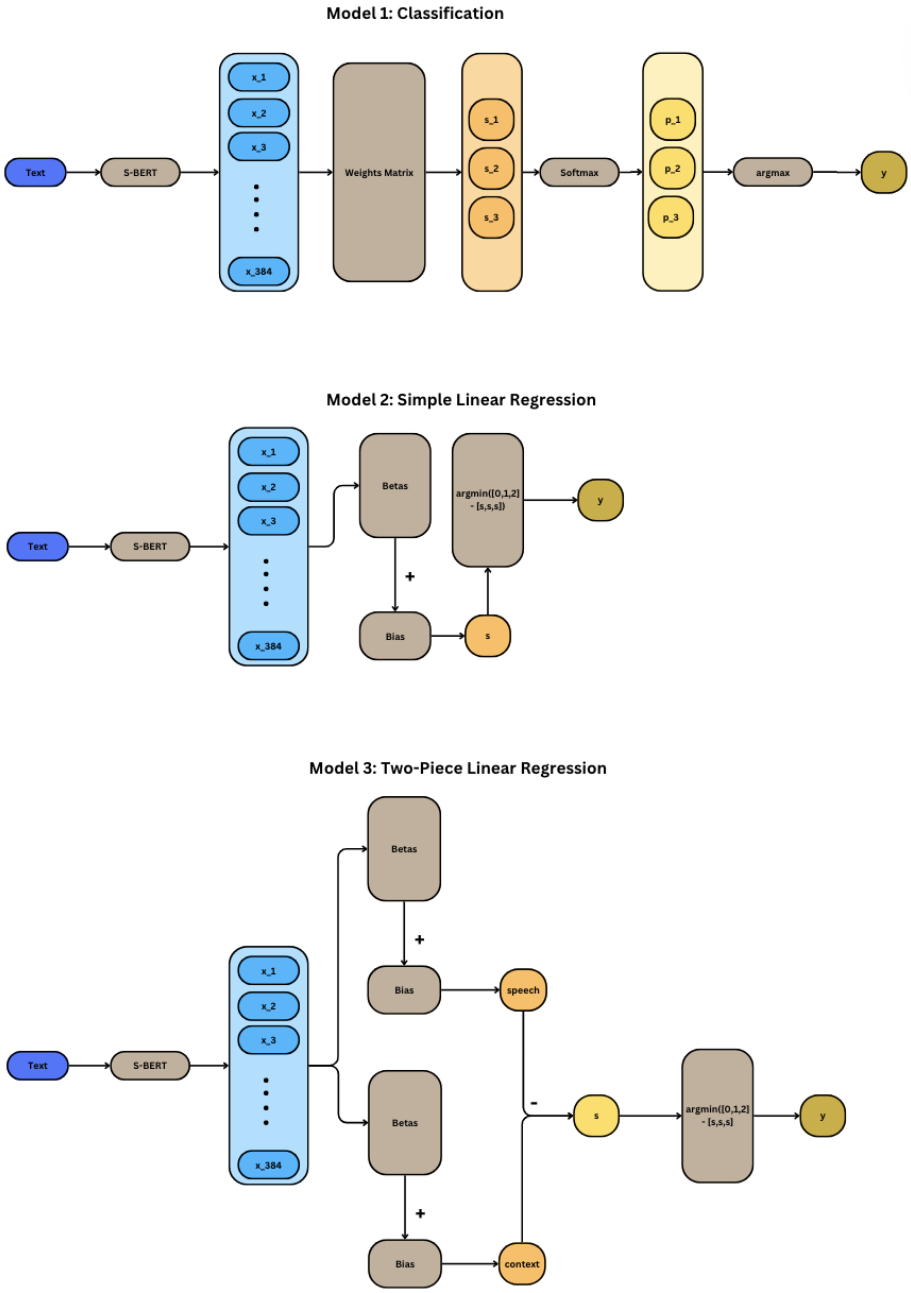
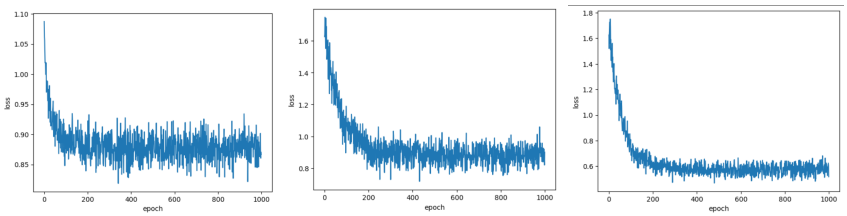


Figure 1: Predictive Models

5 Results



(a) Model 1 (b) Model 2 (c) Model 3

Figure 2: Loss Over Training Progression by Model

Although all three models were trained for 1000 epochs, the models all seemed to converge after 300 epochs. The finals biases 0.3263 for the simple linear model and 0.2973 and -0.2973 for the two-piece model.

	Actual Class		
Predicted	No Hate	Implicit	Explicit
No Hate	50	22	17
Implicit Hate	20	58	10
Explicit Hate	5	19	64

Table 1: Model 1 Confusion Matrix: Accuracy: 0.64

	Actual Class		
Predicted	No Hate	Implicit	Explicit
No Hate	79	68	34
Implicit Hate	10	21	53
Explicit Hate	0	0	1

Table 2: Model 2 Confusion Matrix: Accuracy: 0.38

	Actual Class		
Predicted	No Hate	Implicit	Explicit
No Hate	65	36	12
Implicit Hate	24	53	61
Explicit Hate	0	0	15

Table 3: Model 3 Confusion Matrix: Accuracy: 0.50

The confusion matrices make it apparent that Model 1, which calculates all probabilities of hate speech classes independently, performs the best overall. However, it is interesting to notice that Model 2, the two piece regression, performed better than the single linear regression despite the fact that the weights converged to cancel each other out. Further examination of the trained two piece regression model found that the speech and context variables also converged such that the values that corresponded to speech and context were always equal in magnitude but opposite in sign such that, when context was subtracted from speech, the model was really just doubling speech.

6 Conclusion

This study did not find evidence to support the hypothesis that hate speech exits on an ordered gradient from the three models that this study tested. Rather, the most accurate model learned probabilities for each of the three classes independently. However, the linear regressions used to test for this ordered gradient relationship did tend toward smaller values, leaving open possibility for future study.

7 Bibliography

1. 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

2. 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

3. 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

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

5. 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

6. 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

7. 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. 10.1007/978-981-16-3660-856.

8. "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

9. 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.