# Hate Deception

Christopher Back  Vy Nguyen  Nathan Perez  Jonathan Serrano  Victoria Vynnychok  Joseph Wright
Dr. Latifur Khan  Dr. Yibo Hu  Dr. Yan Zhou

## Introduction

The rise of social media has led to an increased risk of political disinformation and hate speech being spread online, which can have further dire consequences [1]. The challenge for online platforms is to identify and remove hate speech effectively while also protecting free speech rights. To solve this issue, machine learning models have been developed to detect and combat these harmful practices, but can they be deceived?

Our defense system uses a machine learning model trained on a dataset of tweets using binary labels - hate speech or not hate speech. We worked with Dr. Yan Zhou and Dr. Latifur Khan to help chose a suitable dataset with enough features and to improve the accuracy of our defense model. However, we recognize that malicious actors may try to evade hate speech detection and bypass our defense system. With the help of Dr. Yibo Hu, our attack team used various techniques to modify the tweets in our dataset without changing their original meaning, helping us evaluate the robustness and effectiveness of our defense system in the face of adversarial attacks.

## Dataset

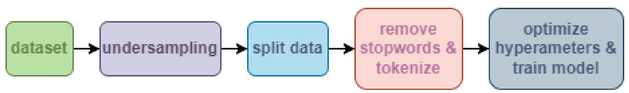
Our team's process for choosing a dataset for our project involved evaluating multiple datasets to find one that had the most useful features for testing and training our defense model. One challenge we faced during this process was the subjective nature of our topic, so we had to ensure that the dataset we chose followed Twitter's user guidelines about what constitutes hate speech and what does not. We were specifically interested in finding a dataset of tweets that were classified only as hate speech or not hate speech, as we were planning to create a binary classification model.

After reviewing several datasets, we eventually found one that was based on Twitter guidelines and was a tertiary classification model, classifying tweets as hate speech, not hate speech, or offensive language. To use this dataset for our purposes, we had to remove the "offensive" label to make it a binary classification model.

## Defense Model

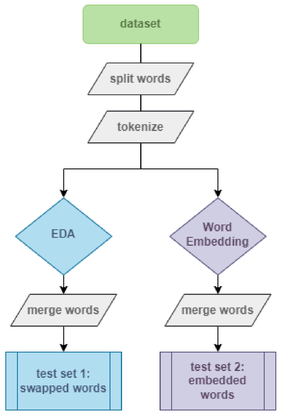
The attack system was tasked to classify text from our dataset to be either hate speech or non-hate speech with a high accuracy. We used three different modals using *sklearn* libraries, however we the prepared the data the same for all the methods. Since the ratio of hate and non-hate speech text was disproportional, we used random under-sampling to even out the ratio. We then split the random sample where 75% of the data would be used for training the modal, and 25% would be used to test the modal. Next, we tokenized and removed stopwords from the text. Finally, we use a grid search to optimize the hyper parameters when training each modal.

1. **Linear Model:** Uses a Passive Aggressive Classifier. This means that if the prediction is correct, keep the model (passive). If it is incorrect, make changes to the model (aggressive).
2. **Random Forest:** Creates a set of decision trees from a subset of the training data. It collects the predictions from all the decision trees to determine a final prediction.
3. **Neural Network:** Uses Muli-layer Perceptron classifier that utilizes a feed forward neural network for classification. Uses a logistic activation function, the Adam model for optimization, and an adaptive learning rate.




## Attack Model

Our goal for the attack system was to generate different variations of the tweets in our dataset using three different text attack methods: Easy Data Augmentation (EDA), Word Embedding, and Bidirectional Encoder Representations from Transformers (BERT).

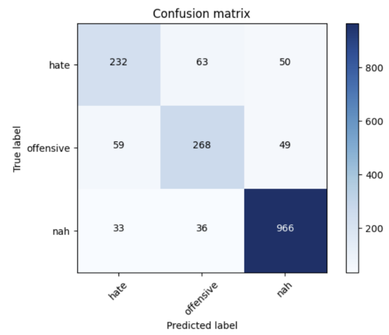



1. **EDA:** For the EDA type, we chose to use random swap, as it generated new sentence structures while keeping original words intact and while still being computationally efficient for our large dataset.
2. **Word Embedding:** For Word Embedding, we used a pre-trained Global Vectors for Word Representation (GloVe) model since it accurately captured global word relationships and is widely used and tested.

From each text attack method, we gathered a new dataset of modified tweets. These three test sets would be run through the defense team's model to see if we successfully fooled it.

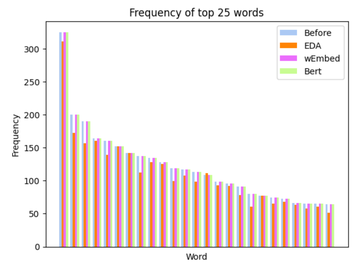
## Results

The accuracy of our linear defense model ended up at 93.47% and 91.07% for the random forest model. To further analyze the success of our defense team's model, we used a confusion matrix to visualize how accurate the predictions were. The diagonal (from top left to bottom right) represents the frequency of successful predictions made by the model.

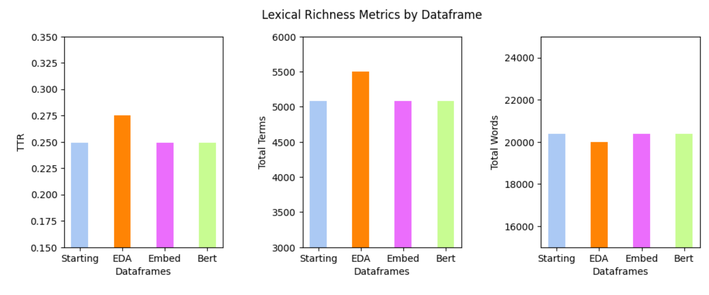



- The model appears to be performing well in correctly identifying tweets that are not hate speech, with 966 true negatives and only 33 false positives. However, the model struggles in correctly identifying tweets that are hate speech.
- The model seems to be confused between tweets that are offensive but not necessarily hate speech, and those that are actually hate speech. This is evident from the relatively high number of false positives and false negatives in the offensive category.

Along with our defense team's model accuracy, the attack team ran statistical analyses on the three test datasets of augmented tweets to compare them to the original dataset.




- **Frequency maintained:** The top 25 words that appeared the most in the tweets in the original dataset still appeared just as frequently after Word Embedding and BERT text attacks. EDA seemed to have less success in this, though.
- **Significance:** This may not necessarily a good or bad thing. It could indicate that the attacks were not effective enough in changing the original meaning. On the other hand, it suggests that attacks were subtle enough to not be detected.




Lexical richness measures the diversity and complexity of the vocabulary in texts. One defining characteristic of hateful language is that it is usually predictable, repetitive, and simplistic. Because of this, the attack team expected to have more lexical richness after the three text attacks, as this can suggest that the attacks successfully changed tweets previously detected as hate speech to not hate speech. The only text attack method that seemed to succeed in this was EDA, however.

## Conclusions

We solved Racism, and world hunger. We helped hate spread across the world easier.

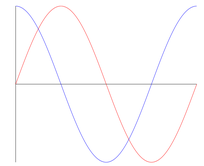


Figure 1. Another figure caption.

## References

[1] "'Urgent Need' for More Accountability from Social Media Giants to Curb Hate Speech: UN Experts | UN News." United Nations, United Nations, https://news.un.org/en/story/2023/01/1132232.



https://github.com/ACM-Research/DisinfoDeception$_{public}$