# Mini Project 3 Report

## Title: Adversarial Attack on Model trained on Hate Speech Datasets

**-Mohit Chandra**

**Type:** This project has to be presented as a poster

**Motivation:** The motivation for this project is to assess the different adversarial attack techniques that can be used on models trained for the task of hate speech detection. We analyze the drop in performance and some adversarial examples.

**Procedure:**

- We chose a general hate speech dataset (AbuseAnalyzer) for this task. This dataset was collected from an alt-right social media platform GAB.com and this dataset looks at different form of hatespeech (such as offensive language, xenophobia, homophobia, antisemitism, islamophobia etc).
- For this project we trained the model on the binary classification task of predicting whether a text post is hateful or not.
- We fine-tuned a BERT based classifier with 2 additional MLP layers on this dataset. Once trained, **we saved the model weights and the examples in the test set that were correctly classified as Hateful.**
- In total we evaluate our results on 765 test examples that were correctly classified as Hateful.
- In this project we only work with these correctly classified test set examples and we try to perturb these to confuse the classifier. If we are successful, we should be able to get some of these examples as 'Non-Hateful'.

**Adversarial Examples Generation Techniques:** In total we experiment with 5 different ways of generating the perturbed examples

- **Community Keyword Swapping:** In this method we tried swapping community based keywords from the sentences. In particular, we replaced, keywords related to the Jewish Community such as 'Jew', 'Jews' and 'Kike' that are heavily used on GAB. We replaced these keywords with 'White', 'Whites' and 'Caucassian'. The idea behind this technique is to check whether some examples are classified as hateful just because they are targeted to a specific community ? even if the content/intent remains unchanged.

- **Masking Bad Words with Signs/Symbols:** We also tried an approach in which we masked the racial slurs/bad words with symbols/signs. It is commonly observed that people denote specific communities with certain symbols. We replaced 'Jew', 'Jews' and 'Kike' with '(())' (which is a common symbol used to refer Jewish community).

- **Wordnet Based Similar Word Search:** In this method we try to randomly replace some percentage of words in each example with their wordnet based synonyms. We experiment with 20% and 50% tokens chosen for wordnet based swap.

- **Counter-Fitted Embedding Based Replacement:** We use RoBERTa based embeddings to replace tokens based on cosine similarity between the original and new sentence/text. We experiment with 20% and 50% tokens chosen for embedding swap.

- **Character Based Alterations:** In this method, we insert, replace or delete adjacent characters of randomly chosen tokens. We experiment with 20% and 50% tokens chosen for character swap.

**Results:**

- **Community Keyword Swapping**
  - Out of 765 test examples, 64 examples contained atleast one of the three terms -- 'Jews', 'Jew', 'Kike'
  - With this method we were able to get 3 examples misclassified -- i.e reduction of ~4.7% in accuracy.

- **Masking Bad Words with Signs/Symbols**
  - Out of 765 test examples, 64 examples contained atleast one of the three terms -- 'Jews', 'Jew', 'Kike'
  - With this method we were able to get 9 examples misclassified -- i.e reduction of ~14.1% in accuracy.

- **Wordnet Based Similar Word Search**
  - With 20% tokens being replaced we got an accuracy drop of ~6.1%
  - With 50% tokens being replaced we got an accuracy drop of ~11.6%

- **Counter-Fitted Embedding Based Replacement**
  - With 20% tokens being replaced we got an accuracy drop of ~3.5%
  - With 50% tokens being replaced we got an accuracy drop of ~7.3%

- **Character Based Alterations**
  - With characters in 20% tokens being replaced/added/deleted we got an accuracy drop of ~7.8%
  - With characters in 50% tokens being replaced/added/deleted we got an accuracy drop of ~29.9% [Although the tokens in this case didn't make much sense, so it might not be considered a valid perturbation]

## Example Analysis: **[Disclaimer: Offensive Language Ahead] apologies in advance**

**Here we present some of the perturbed examples that were misclassified by the classifier as 'Non-Hateful'.** Changed tokens are highlighted in yellow.

| **Methodology** | **Original Example** | **Perturbed Example** |
|---|---|---|
| Wordnet Based Similar Word Search (20%) | yes it is it's j*ws isn't it | yes it is it's h*brew isn't it |
| Wordnet Based Similar Word Search (20%) | the joke relates to the fact it kneeling while saluting would never never have occured in germany | the joke relates to the fact it kneeling while wassail would never never have pass in deutschland |
| Wordnet Based Similar Word Search (50%) | put a b*llet in them/they will never do it again | lay a s*oke in them/they will never do it again |
| Wordnet Based Similar Word Search (50%) | lol shut the f*ck up you look like the fat cuck from clerks | lol unopen the b*ng up you attend alike the rich cuck from sal*sclerk |
| Counter-Fitted Embedding Based Replacement (20%) | fuck twitter that is all | goddam twitter that is all |
| Counter-Fitted Embedding Based Replacement (20%) | ni**ers have been in the new world for 400 years but they haven't progressed one single day | d*rkies have been in the new international for 400 years but they haven't progressed one exclusive day |
| Counter-Fitted Embedding Based Replacement (50%) | lol you think very highly of yourself | jk you believe very exceedingly of yourself |
| Counter-Fitted Embedding Based Replacement (50%) | haha fresh off the boat ke*yans think they can influence topics ... hilarious | hahahaha fresh off the ship ke*yans reckon they can impacts items ... comic |
| Character Based Alterations (20%) | yeah that's a kike sorry | yeah that's a cike sorry |
| Character Based Alterations (20%) | j*ws orchestrating all this | jz*s orchestrating all this |
| Character Based Alterations (50%) | nope just full of ni**ers | nope just ful of n*gers |
| Character Based Alterations (50%) | ki*ky friedman high on jesus ki*ky friedman& the texas j*w boys | ki*ny friedman high on sjesus k*nyk friedman& the texas jqew bos |