Group 5 Final Project Progress Presentation

Valerie Chu, Allan Choun, Rohit Suresh, Hunter Jacobson, Vishal Saminathan

Topic Idea: Content Moderation Against Hate Speech

- Our main idea revolves around developing an AI model that flags texts as inappropriate through a list of words in a real dataset.
- Can be used in a broader application for identifying hate speech, more specifically phrases
- Flags the content and lets the moderator know, then the moderator can make the final decision, factoring in a human element
- Can be integrated into websites and other forums that foster human interaction - increased scalability

Dataset

- Our dataset is from a 2018 GitHub repository. It contains text from random sets of posts on Stormfront, a White Supremacist Forum.
- The text is stored as .txt files, each of which contains a sentence that has been manually labelled as containing hate speech or not in a separate csv file.
- We are given randomly chosen training and testing data at an 80/20 split, which was taken from the total selection of 10,944 text files.

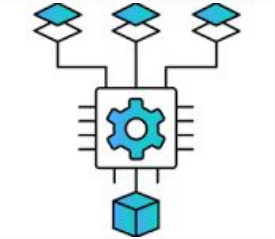
Ethical Analysis

- High importance in understanding that ethics are crucial in AI to ensure fair and unbiased content moderation
- Dataset may not fully represent all demographics
    - Address biases by actively identifying and correcting them
    - Modify algorithms that may unfairly target or fail to detect hate speech in specific groups
- Continuous Model Training: Regularly update and train the model on updated datasets to keep pace with evolving language use and societal norms.

Pros	Cons
Real world example helps in researching how effective content moderation algorithms are and ethical implications	Risk of overfitting to specific types of hate speech that are overrepresented
Improves AI model's ability to identify and address explicit forms of hate speech	Comments can be taken out of context culturally or socially

Project Plan

- Phase 1
  - Research based on the provided, ‘Hate Speech Dataset from a White Supremacy Forum’
  - Clean data to remove special characters and stopwords
  - Tokenize and vectorize the text
- Phase 2
  - Train a Logistic Regression/Naive Bayes model
  - Split the data into training and test sets
- Phase 3
  - Examine if certain groups are disproportionately flagged
  - Deduct how to reduce bias and how false positives/negatives could impact users



Thank you for listening!