AI CONTENT MODERATION SYSTEM

Trenyce, Tahmeed, Masrur, David

OBJECTIVES

•Build a simple AI model to classify text as appropriate or inappropriate (e.g., toxic, offensive, or spam) using a real dataset

- Analyze its performance, discuss ethical concerns, and propose improvements for fairness and transparency

•A successful AI content moderation system would solve the problems of users being exposed to inappropriate content

- Prevents self-harm and suicide by flagging content that promotes suicidal thoughts
- Filtering out any racist, sexist, and prejudiced language
- Tracking and deleting spam/phishing messages

  Detecting suspicious links would prevent users from accessing risky sites and getting tracked by hackers

DATASET USED: "LEARNING FROM THE WORST"

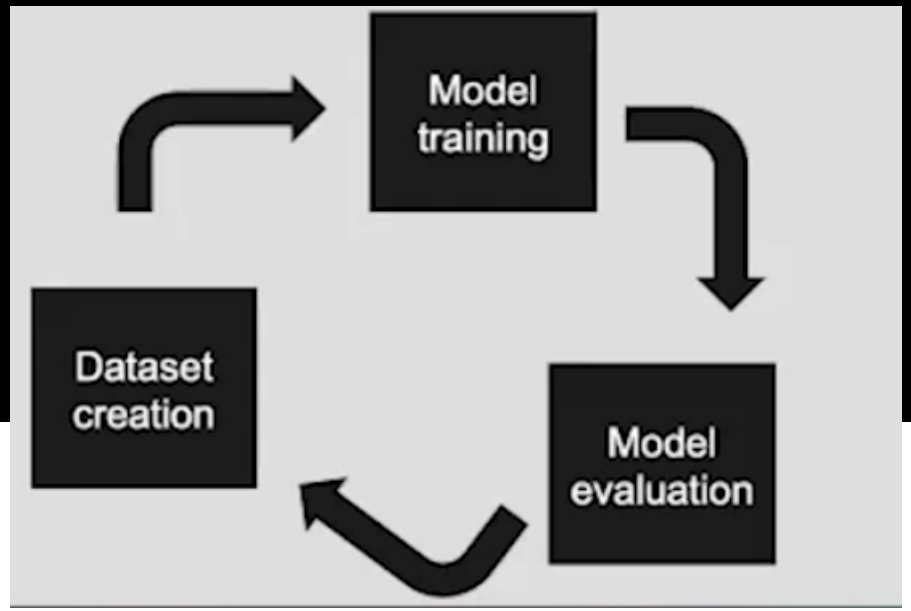
Goal: To train a model for detecting online hate speech through dynamic data generation

Entries: 41,255

- 53% is Hate speech
- This is usually around 20%
- 6 labels of hate to clarify data further

Label	Type	Total
Hate	Not given	7,197
	Animosity	3,439
	Dehumanization	906
	Derogation	9,907
	Support	207
	Threatening	606
	Total	22,262
Not Hate	/	18,993
All	TOTAL	41,255

DATASET: PART 2



Trained in 4 rounds:

Initial model was trained on:

- 400,000 entries from hatespeechdata.com

Each round consists of:

- Annotation

  The model is tricked by presenting it with adversarial content

- Model training:

  Retrained with data from the previous round and the current round

This form of modeling paradigm is what makes this dataset dynamic!

WHY WE CHOSE THIS DATASET:

- Large Sample Size of over 40,000
- Balanced data with around 50% hate and 50% not
- Excellent information, examples, and well-maintained code

DATASET: SMS SPAM COLLECTION

The SMS Spam Collection is a public set of SMS labeled messages that have been collected for mobile phone spam research.

- A collection of 425 SMS spam messages manually extracted from the Grumbletext Web site.
- A subset of 3,375 SMS randomly chosen ham (non-spam) messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore.
- These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available.
- It has 1,002 SMS ham messages and 425 spam messages

LABELING DATA & PRE-PROCESSING:

We have 3 categories that the data can fit into:

- Hate Speech (Characterized by offensive language)
- Spam (Characterized by promotional messages and phishing attempts)
- Normal (We included this to help with classification of the data)

Pre-Processing Steps:

- Convert all text to lowercase to standardize across datasets
- Remove punctation and special characters (we deemed this unnecessary for the classification categories)
- Split sentences into individual words
- Use Term Frequency-Inverse Document Frequency:
  - By converting words into numerical values, we get to identify which words are most common across datasets and analyze their relevance to the classification category

ROADMAP FORWARD

Model Training/Evaluation:

- Use Logical Regression as the ML model of choice
  - Logical Regression is a statistical method used for predicting the probability of a binary outcome based on input features
  - This model will work well after we assign numeric weight to the words in our dataset
- After pre-processing the data, it will be split into 80% training and 20% testing sets
  - Commonly used in ML; Offers a good balance of training the data and using the other 20% to test our prediction accuracy

ROADMAP FORWARD CONT.

Evaluation Metrics:

- Accuracy: How often is the data being properly classified?
- Precision: How many false positives and false negatives are we getting?
- F1-Score: This will help us to minimize the occurrence of false positive and negatives in our dataset

ETHICAL ANALYSIS

Ethical Dilemmas:

- AI systems tend to discriminate against marginalized groups if the training data is unbiased or unrepresentative
- We want to make sure that we balance preventing harm and protect ones' freedom of speech
- We value others' right to privacy and want to make sure users are informed about how their data is being collected and analyzed

Potential Fixes:

- Plan to implement context-sensitive moderation for different race and gender groups
- We chose from datasets where user information isn't uploaded with data. Also tried to choose from data sets that had transparency on where it was collected from
- We chose diverse datasets that have spam and hate speech data sets that are targeted towards different groups of people

SOFTWARE AND TOOLS

Jupyter Notebook: Chosen for its interactive environment, allowing for easy implementation, testing, and publication of code with visualizations and explanations in markdown.

Python: Used due to its rich ecosystem of libraries for text processing and machine learning.

CountVectorizer & TfidfVectorizer (from scikit-learn): Essential for tokenizing text and converting it into numerical features, enabling further analysis and machine learning applications.

Matplotlib: Selected for its ability to generate insightful visualizations, helping to analyze patterns and trends in the processed data.

Pandas: Used to efficiently handle, manipulate, and structure datasets, ensuring seamless integration with other libraries.

Logistic Regression: A simple yet effective baseline model for text classification, providing fast training and easy interpretability when using TF-IDF features. It also performs well on medium-sized datasets, like ours.

MINIMUM VIABLE PROJECT (MVP)

- Train a Logistic Regression model.
- Use Vectorizer for text processing.
- Evaluate its accuracy, precision, F1 score
- Conduct basic bias analysis on dataset distribution.
- Write an ethical reflection
- Prepare a presentation

WEEKLY PLAN

Week 1: Select & clean dataset (remove special characters, uppercase).

Week 2: Convert text to numerical features, split dataset.

Week 3: Train Logistic Regression model, evaluate with key metrics.

Week 4: Perform bias analysis, identify false positives/negatives, draft ethical reflection.

Week 5: Test model improvements, create visual data.

Week 6: Finalize presentation

THANK YOU!