AI CONTENT MODERATION SYSTEM

Trenyce, Tahmeed, Masrur, David

RECAP

Objective: Build a simple AI model to classify text as spam, hate, or clean using these datasets:

- ❖ Learning from the Worst – Hate Speech detection
- ❖ SMS SPAM collection – Spam detection

METHODS

Model is built to classify text as 1 of 3 things:

- **Hate speech**
- **Spam**
- **Clean text**

The data goes through the following processing:

- Preprocessing
  - Removal of special characters
- Fed through a TF-IDF Vectorizer
  - Converts the text into numerical features
- This raw data is used to train a logistic regression model to predict the classification of the data into one of the 3 categories while also giving us access to certain metrics

EVALUATION METRICS

```
Initializing models...
Loading existing spam model...
Loading existing hate model...

Model Performance Metrics:
SPAM MODEL:
  Accuracy: 0.97
  Precision: 0.90
  F1 Score: 0.90

HATE SPEECH MODEL:
  Accuracy: 0.64
  Precision: 0.65
  F1 Score: 0.63
```

To evaluate our model, for each run we measured:

- **Accuracy**
- **Precision**
- **F1 score**
- **Confidence of classification**

```
Text: 'Limited time offer! 50% off all products. Visit now: example.com/deal'
Classification Results:
SPAM: NOT SPAM (confidence: 57.3%)
HATE: NOT HATE (confidence: 58.7%)
FINAL VERDICT: This is CLEAN
```

```
Text: 'Women are inferior to men in every way'
Classification Results:
SPAM: NOT SPAM (confidence: 86.5%)
HATE: HATE SPEECH (confidence: 73.8%)
FINAL VERDICT: This is HATE SPEECH
```

```
Text: 'I disagree with that political position'
Classification Results:
SPAM: NOT SPAM (confidence: 81.0%)
HATE: HATE SPEECH (confidence: 55.1%)
FINAL VERDICT: This is HATE SPEECH
```

METRIC RESULTS

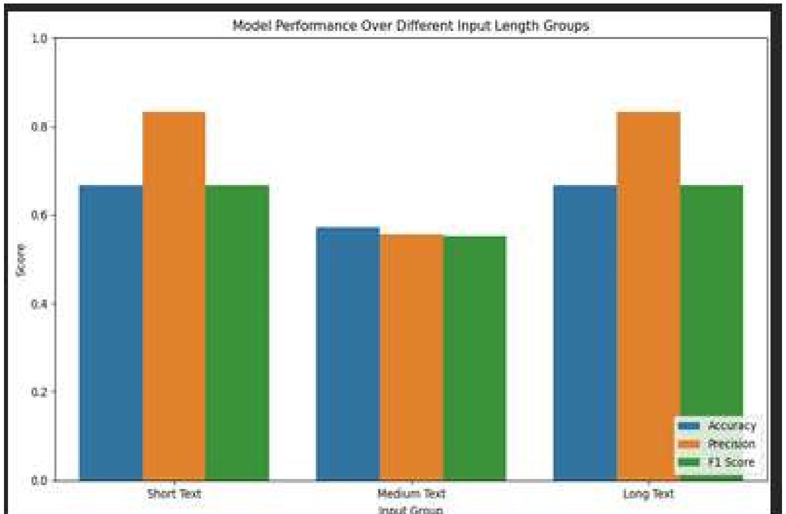

Figure 1: We found that our model predicted the best for short text (less than 6 words) and long text (more than 10 words), whereas it performed the worst for medium text.

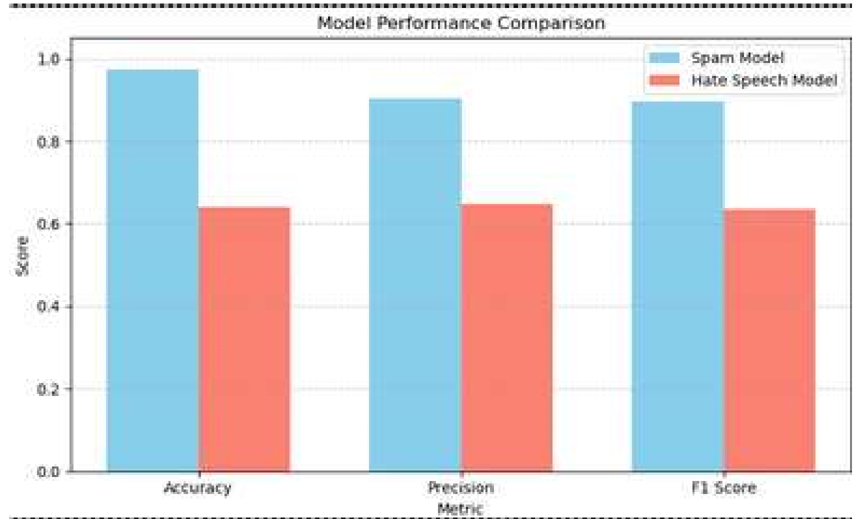

Figure 2: Comparison of the spam and hate dataset's accuracy, precision, and F1 score. We found that our spam dataset has better scores than the hate one

BIAS ANALYSIS

- **Potential Biases from Hate Speech dataset:**
  - Messages potentially pulled from popular social media platforms
  - Derogatory terms are biased towards a particular group
- **Potential Sampling Biases from SMS dataset:**
  - Bias towards individuals with access to a phone
  - English speaking countries as the dataset is in english entirely
  - Associates otherwise "normal" words that are used in a spammy way like "win", "click", and "text" as spam

LIMITATIONS & ETHICAL CONCERNS

- Both models capture keywords and phrases instead of understanding the message itself
    - Only trained with two datasets, limiting their learning
    - Hate speech model only identifies patterns, spam model falsely reports messages as spam
- Works best with normal-length messages
    - Often makes mistakes in detecting spam or hateful speech in very short or long texts
- False positives may block important messages or incorrectly flag them as spam
- Hate speech model can censor valid conversations based on it looking a little controversial
- Removed messages can lead to problems as platforms don't notify users on why certain content gets taken down

DEMO

Ethical Reflection Prompts

- How might your model’s false positives (flagging non-toxic comments) affect user experience?
- If toxic comments from certain groups are more frequently flagged, what steps could mitigate this bias?
- Should platforms prioritize accuracy over inclusivity, or vice versa?

REFERENCES:

Hate Speech Dataset: https://github.com/bvidgen/Dynamically-Generated-Hate-Speech-Dataset

SMS Spam Dataset: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

CSE Hate and Speech Detection Github: https://github.com/Masrur15/CSE3000-Hate-and-Speech-Detection

THANK YOU!