Progress Presentation

Sam Cultera, Malcolm Ferguson,
Josef Karpinski, Alexander Manos

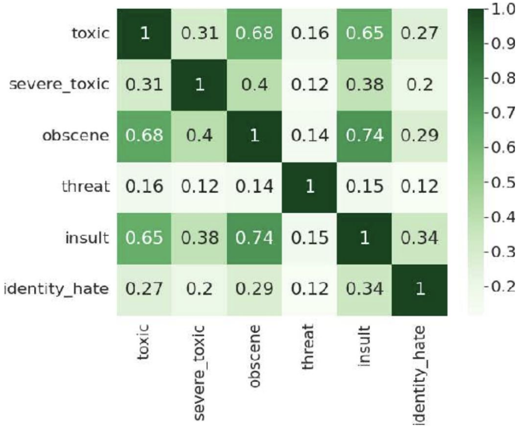
Project Overview

- Content Moderation Filter
    - Select a dataset of content classified as various levels of offensive
    - Train a model to classify text
    - Analyze results and performance of model
    - Study ethical implications of using such a model to classify content

https://health.clevelandclinic.org/dangers-of-social-media-for-youth

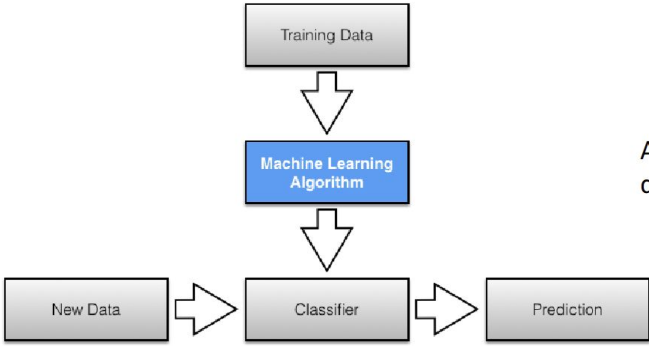
Selecting a Dataset

- Jigsaw Toxic Comment Classification Dataset
  - Purpose: Designed to classify comments as toxic or nontoxic
  - Data Type: English text comments from Wikipedia
  - Number of Samples: 150,000
  - Labels: Multi-class classification
    - Each comment is assigned 6 binary labels: (toxic, severe toxic, obscene, threat, insult, identity hate)
- Why this dataset?
  - Well-labeled data, ideal for supervised learning
  - Real world use case
  - Diverse in types of toxicity



Source: https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data

Plan



- Preprocess Dataset
    - Clean text (e.g. remove special characters)
    - Split data into training and test sets
    - Tokenize and convert text to numerical features
- Train Various Models
    - Classify text as toxic, not toxic, or any other label
    - Train simpler models (Logistic Regression, Random Forest)
    - Compare with more complex models (LLMs)
- Evaluate Performance
    - Use metrics such as accuracy, precision, recall, and F1-score

Ethical Considerations

- Over & Under Moderation

Over-Moderation: Excessive filtering could stifle free speech, preventing open discussions on controversial but necessary topics.

Under-Moderation: If the model is too lenient, it may allow harmful content to spread unchecked.

- False Positives and False Negatives

False Positives: Innocuous content being mistakenly flagged as toxic can frustrate users and limit expression.

False Negatives: Harmful content slipping through moderation could lead to harassment, misinformation, or community harm.