Toxic Speech Detection

In-quarter project #3: Data use & revision

Akansha Lalwani (A59019733)
Mohammed Alblooshi (A59019984)
Kanlin Wang (A16182611)

Introduction

- **Purpose:**
  - Showcase outcomes and insights from our data science project
  - Highlight how these findings can contribute to increased justice in real-world contexts

- **Target audience:**
  - End-users without a data science background, other data scientists

Real-World Challenges and Model Utility

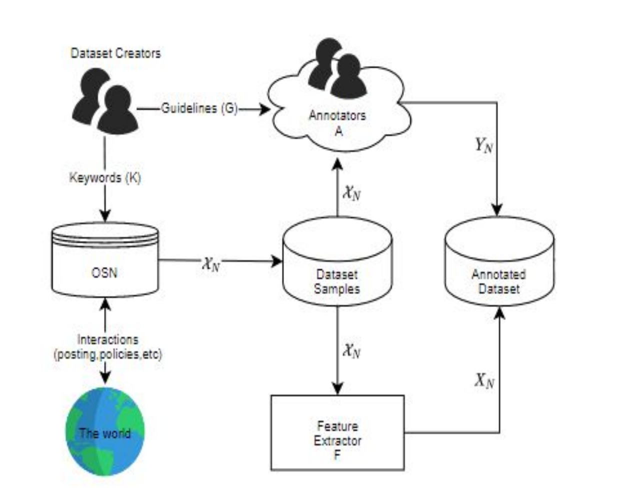
- **Bias:** Building a fair model that avoids favoring certain groups or demographics
- **Distinguishing toxic speech:** Ensuring accurate identification of harmful content
- **Volume of harmful content:** Managing the vast amount of toxic speech online

Background on Data and Models

- **Data Overview:**
  - Instagram comments from public profiles in California, Texas, and New York
  - Focus on comments written in English during November and December 2022
- **Data Collection Process:**
  - Implementing a system to collect relevant Instagram comments using the Instagram API
  - Compliance with privacy regulations and filtering out non-English comments, spam, advertisements, and irrelevant content

Model Types, Purpose, and Validation

- **Model Types and Purpose:**
  - Algorithms used: Naive Bayes, SVM, LSTM/GRU
  - Purpose: Classify and predict toxic speech in Instagram comments
- **Model Validation:**
  - Validation set: Subset of labeled data
  - Purpose: Fine-tuning model parameters and selecting the optimal model for accurate toxic speech detection



Appropriate Use and Limitations

- **Appropriate Use:**
  - Identifying and mitigating toxic speech in online communities for a safer environment
  - Supporting content moderation efforts to reduce harm and promote respectful dialogue
- **Tests/Rules for Non-Experts:**
  - Use the model to flag explicit forms of toxic speech, such as hate speech, threats, or harassment
  - Utilize the model as a tool to aid in content moderation decisions, providing insights for human judgment

To determine if the model can be used appropriately, non-experts should ask the following questions: Does the problem involve detecting toxic speech? Is the content primarily in English and within the contexts of California, Texas, or New York? If both answers are 'yes,' the model is likely suitable. Always remember to have a human moderator on standby for nuance interpretation.

Non-Appropriate Use

- **Contextual Interpretation:**
  - The model may not capture nuanced or context-dependent instances of toxicity where understanding underlying meaning and intent is crucial

- **Data Limitations:**
  - Biases in the training data can lead to biased behavior and inaccurate results
  - The model's effectiveness may be limited in detecting emerging or evolving forms of toxic speech

- **Ethical Considerations:**
  - Responsible use of the model is essential to avoid amplifying biases or restricting free speech
  - The model should not be the sole determinant of guilt or punishment

Determining Appropriate Use

- **Complexity of the Problem:**
  - Assess if the model aligns with the specific problem being addressed and if it provides meaningful insights
- **Contextual Understanding:**
  - Evaluate the need for human judgment in interpreting the model's output in different contexts
- **Potential Biases:**
  - Analyze the training data to identify and address biases and consider diverse perspectives
- **Impact and Consequences:**
  - Consider the ethical implications and potential consequences of using the model in specific scenarios

In a technical sense, ensure that the data fed into the model aligns with the parameters of the training data – English comments from public profiles in the mentioned states. Misaligned data might lead to inaccurate results due to the model's limitations and inherent bias.

Model Overview and Performance

- **Model Architecture:**
    - The model utilizes a combination of natural language processing algorithms, such as Naive Bayes, Support Vector Machines (SVM), and deep learning models like LSTM or GRU
- **Training Data:**
    - The model was trained on a dataset of labeled Instagram comments, collected from public profiles in California, Texas, and New York during November and December 2022
- **Validation and Performance:**
    - A portion of the labeled data was used as a validation set to fine-tune the model parameters and select the optimal model for accurate toxic speech detection

Interpreting Model Outputs

- **Output Interpretation:**
  - Model outputs indicate if a comment is toxic or non-toxic
  - Toxic output: Contains harmful or offensive language
  - Non-toxic output: Does not contain explicit toxicity
- **Real-World Connection:**
  - Interpreting outputs addresses the challenge of toxic speech online
  - Positive output: Violates community guidelines or harms individuals
  - Negative output: Comment is not explicitly toxic
- **Context-Sensitive Interpretation:**
  - Consider platform, audience, and policies
  - Cultural norms impact interpretation
- **Guiding Non-Data Scientists:**
  - Evaluate outputs based on context, guidelines, impact on individuals and community
  - Align with platform values, policies
- **Human Review and Decision-Making:**
  - Model outputs provide insights but need human judgment
  - Consider context, biases, and subjective judgment

Ethical Data Retention Plan

- **Long-Term Data Retention:**
  - Data will be retained for one year for analysis and model refinement
- **Ethical Justification:**
  - Individuals have control over their data and can withdraw consent at any time
  - The ability to request data deletion ensures individuals' autonomy and data ownership
  - Secure storage and deletion processes minimize the risk of data exposure and potential misuse
- **Technical Details:**
  - Data will be securely stored using industry-standard encryption methods
  - Access controls and authentication mechanisms will be implemented to safeguard the data
  - Regular audits and monitoring will be conducted to ensure compliance with privacy regulations

Technically, our plan involves automated systems that delete data after a specified period, and a secured database to maintain the integrity and confidentiality of the data while it's retained.

Ethical Model Revision

- **Contexts Requiring Ethical Revision:**
  - **Emerging Forms of Toxicity:**
    - Language and social norms evolve, leading to the emergence of new forms of toxic speech
    - Ethical revision is needed to update the model to recognize and address these new toxic speech patterns
    - User reports and feedback are valuable in identifying and understanding emerging forms of toxicity
  - **Bias Mitigation:**
    - Bias can inadvertently be present in the model's predictions or decisions
    - Ethical revision is necessary to identify and mitigate biases, ensuring fair and equitable outcomes
    - User reports and ongoing monitoring play a crucial role in detecting biases and initiating necessary revisions

Ethical Model Revision

- **Ethical and Technical Aspects of Monitoring:**
  - **Ongoing monitoring of the model's performance is essential for:**
    - Detecting biases: Continuously assessing the model's predictions and decisions for potential biases to ensure fair and equitable outcomes.
    - Identifying false positives and false negatives: Regularly evaluating the model's performance in accurately classifying toxic and non-toxic speech.
  - **Ethical Justification:**
    - Ensuring effectiveness and fairness: Monitoring helps maintain the model's effectiveness in addressing toxic speech and promotes fair treatment of users.

Societal Scale Implications

As our model scales up to a societal level, potential issues could arise. The model's limitations, such as not fully comprehending non-English languages or dialects, could potentially impact its effectiveness.

Ethically, we risk enforcing a 'one size fits all' standard of toxic speech across diverse cultures and languages, potentially infringing on free speech rights. The magnitude of data and potential for misuse are also concerns that arise with large-scale use.

Thank you