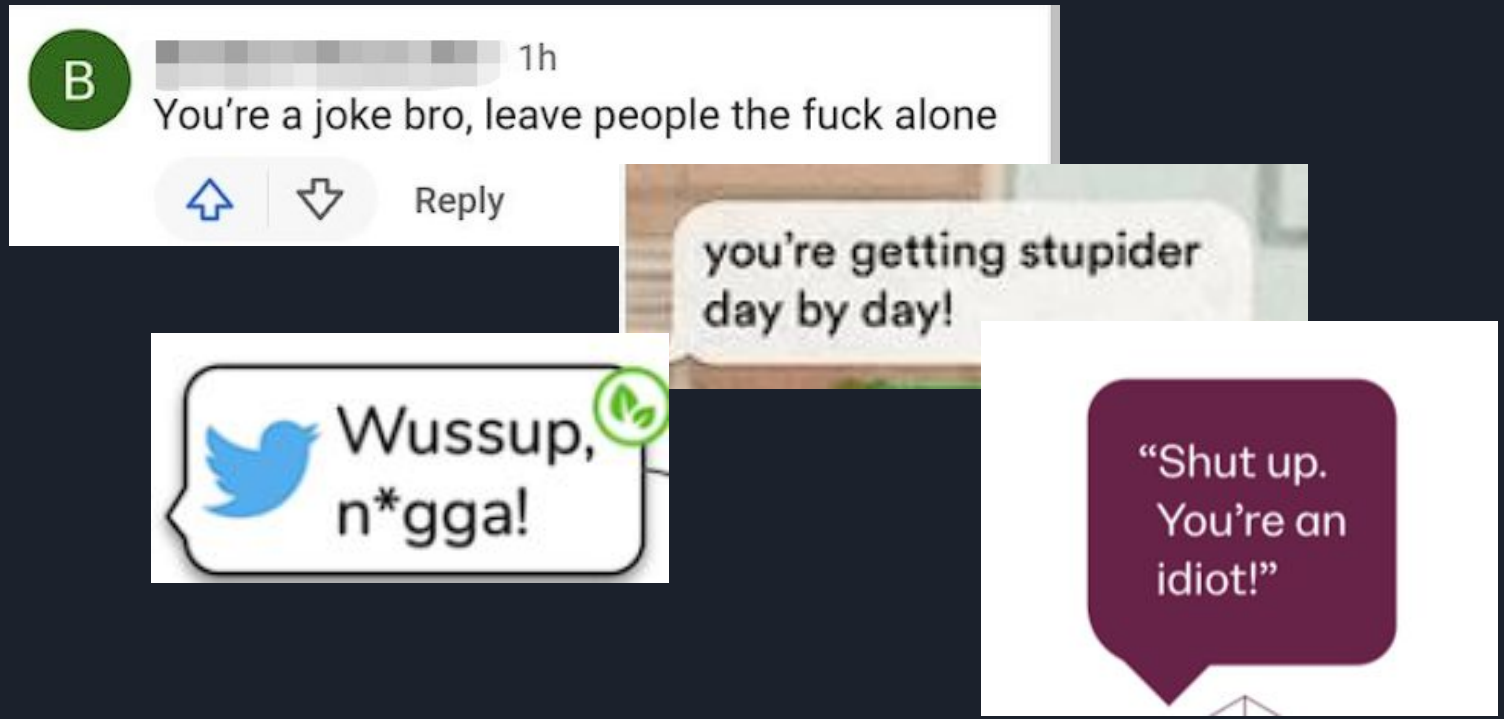
Toxic Speech Detection

#1: Data collection

Akansha Lalwani

What is Toxic Speech?



Definition of Toxic Speech

“rude, disrespectful or unreasonable language that is likely to make someone leave a discussion"

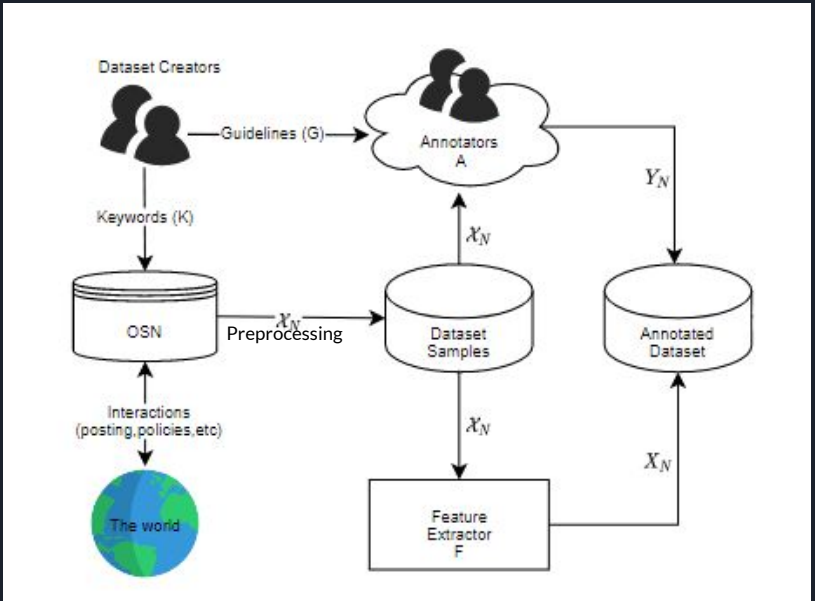
—Dixon et al.

Measuring and mitigating unintended bias in text classification

Core Challenge

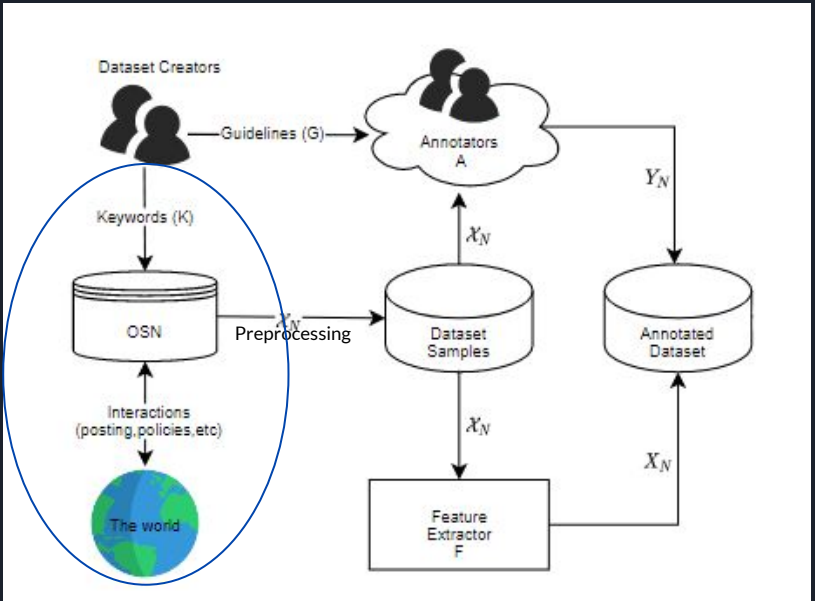
- Algorithms may **exhibit bias** - How to build a fair model
- Algorithms can **struggle to distinguish** between toxic and non-toxic comments - Model should capture the meaning and context behind the key word
- Algorithms are only as good as the data they're trained on - Volumes of harmful content online can make it difficult to **train accurate models**

Data Collection and Measurement System



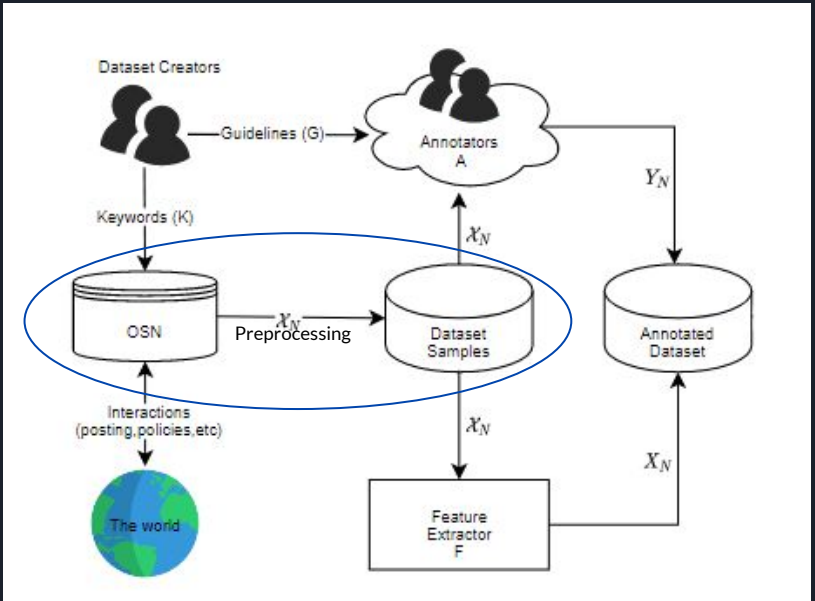
We have proposed a data collection and measurement system for toxic speech detection in Instagram comments, limited to the English language, and focused on California, Texas, and New York to cover diversity in the United States during November and December 2022.

Data Crawling and Scraping



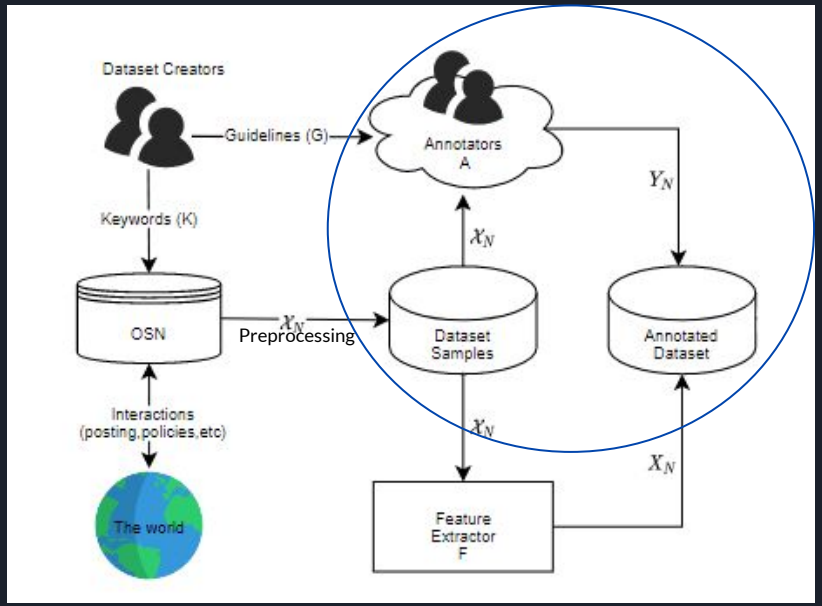
The first step would be to gather Instagram comments from public profiles in California, Texas, and New York during the specified time frame. We can use Instagram's API to collect this data, ensuring that we are only collecting public comments and not violating any privacy concerns. We would then filter the comments to only include those in the English language.

Data Preprocessing



The next step would be to preprocess the data, including text normalization, removing stop words, and stemming to reduce noise in the data. We would also remove any irrelevant comments that do not relate to the study, such as spam, advertisements, or comments in other languages.

Data Labeling



Toxic speech detection requires labeled data, meaning each comment needs to be classified as toxic or non-toxic. We could use a combination of human labelers and machine learning algorithms to label the data. For example, we could randomly sample a portion of the data and have human labelers annotate it as toxic or non-toxic, and then use this labeled data to train a machine learning algorithm to classify the rest of the comments.

How is it useful for the core challenge?

Language specificity: By limiting the data collection to Instagram comments in the English language, the system ensures that the collected data is specific to the language used in the United States. This specificity is crucial in training machine learning algorithms to identify toxic speech accurately.

Diverse sample: Focusing on California, Texas, and New York ensures that the collected data covers a diverse range of demographics and geographic regions in the United States. This diversity is essential in identifying and addressing any potential biases in the data and ensuring that the toxic speech detection system is effective across different regions and communities.

Time specificity: Collecting data during November and December 2022 ensures that the data collected is timely and reflective of the current social and political climate in the United States. This is important as the prevalence and types of toxic speech can vary over time.

Preprocessing: This ensures that the machine learning model can focus on identifying patterns and features of toxic speech rather than being confused by irrelevant data.

Labeled data: Although we have the definition of the toxic speech, different people have different feelings for the same speech. To eliminate bias, we use a group of annotators to define whether the speech is toxic or not. First, we consider the annotators' background. We build the annotators group with a variety of political, religious, and social views. Secondly, one speech should be annotated by different annotators to get average points which can avoid bias. So that we can know is the same speech is toxic or not for different people. To avoid the model only define the toxic by key words, we also feature extractor to find more related features instead of the key words only.

Relevance of the data to the real world issue

- Enable researchers and developers to **identify patterns and trends in toxic speech**, such as the types of language used, the frequency of occurrence, and the specific groups targeted.
- Useful for **developing and training** more accurate and effective machine learning models for toxic speech detection.
- Help in **creating targeted interventions and support f**or those affected by toxic speech online.
- Provide insights into the prevalence and impact of toxic speech across different regions in the United States, **highlighting potential differences** in cultural norms, demographics, and online behavior.
- Inform **targeted policies and interventions** to address toxic speech and promote a more inclusive and respectful online community.

Ethical Concerns and Response

- Privacy
- Bias and Discrimination
- Free Speech
- Transparency and accountability
- Unintended consequences

Privacy

Ethical Concern:

- Collecting data for toxic speech detection raises privacy concerns
- Should respect individuals' privacy and comply with data protection regulations

Response:

- Anonymization techniques such as data masking, aggregation, and encryption can protect individuals' privacy
- Informed consent procedures should be implemented
- Individuals should be informed about the data being collected and how it will be used

Bias and discrimination

Ethical Concern:

- Toxic speech detection models may be biased and result in discriminatory outcomes
- Biases can come from training data, implicit assumptions, or other factors
- Design and test models to minimize the risk of bias and discrimination

Response:

- Careful data selection, pre-processing, and model training can help mitigate bias
- Use model validation techniques like fairness testing and sensitivity analysis to identify and correct biases

Free speech

Ethical Concern:

- May infringe on individuals' freedom of expression by potentially restricting legitimate speech
- Balance the need to protect people from harmful speech with the principles of free speech

Response:

- Ensure the system only flags explicitly harmful or threatening speech, not controversial or offensive speech
- Robust appeals processes should be in place to allow individuals to challenge any decisions made by the system

Transparency and accountability

Ethical Concern:

- Can be opaque, making it difficult to hold them accountable and address errors or biases
- Transparency and accountability are crucial in the development and deployment of these systems

Response:

- Provide clear documentation of the system's inner workings and decision-making processes
- Conduct regular audits and independent evaluations
- Ensure the system is operating as intended and identify any errors or biases

Unintended Consequences

Ethical Concern:

- Misidentification of harmless or even positive speech as toxic
- May lead to a chilling effect on free speech

Response:

- Anticipate and mitigate unintended consequences through careful planning, testing, and evaluation
- Design the system to take potential risks and harms into account
- Implement mechanisms to monitor the impact of the system and make adjustments as needed

Thank you