ONLINE COMMENT TOXICITY PREDICTOR

By- Siddarth, Karim, Samantha, Akshay

Table of Contents

- Problem Statement
- Data Summary
- Model Methodology
- Models
- Model Deployment
- Future Work

Problem Statement

Online discussions have become an integral part of daily lives.

03

41%

of Americans have personally experienced some form of online harassment.*

20%

of Americans of those been harassed have experienced online harassment because of their political views.

Gender or racial/ethnic background are also the reasons for being harassed online. Social Media Platform combat these behaviours through their internal policies.

Reference: *Pew Research: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment
Reference: *https://scholars.org/contribution/countering-online-toxicity-and-hate-speech

Policies to Combat Behaviour

- Public figures vs private individuals
- Critical discussion on public news
- Block private individuals
- Person can self report.

- Allows self reporting
- Block people who are found harassing.

- Policy of "Do not make personal attacks anywhere in Wikipedia
- Notes that attacks may be removed and the those users being blocked

Self Reporting is difficult when someone is getting trolled (Common in political debates)
Challenge of creating effective blocking policies is two fold:

1. Identification of a toxic comment
2. Maintaining a fine balance between freedom of Speech and toxic comments.

Data Summary

1.5 M

Total Toxicity Annotated Comments

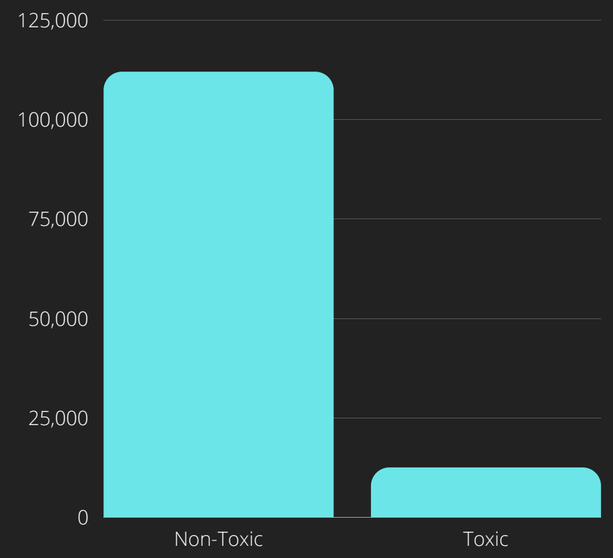
160K

Unique Comments Annotated by different Annotators

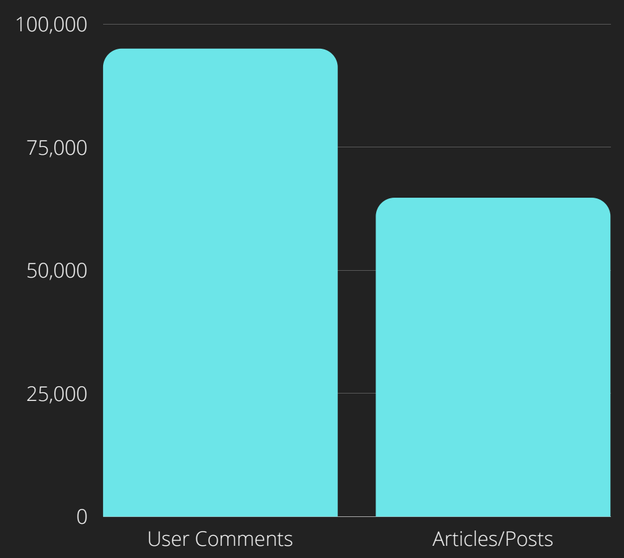
3.5K

Crowsourced Annotators with different demographics who labelled the data

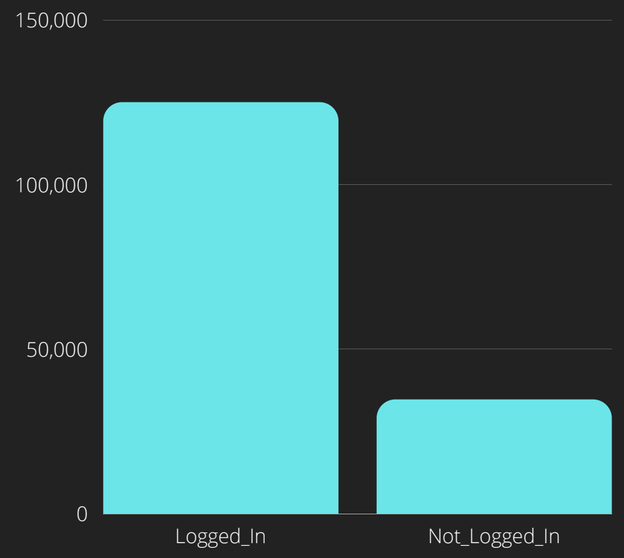
Toxic vs Non Toxic



User Comments vs Articles/Posts

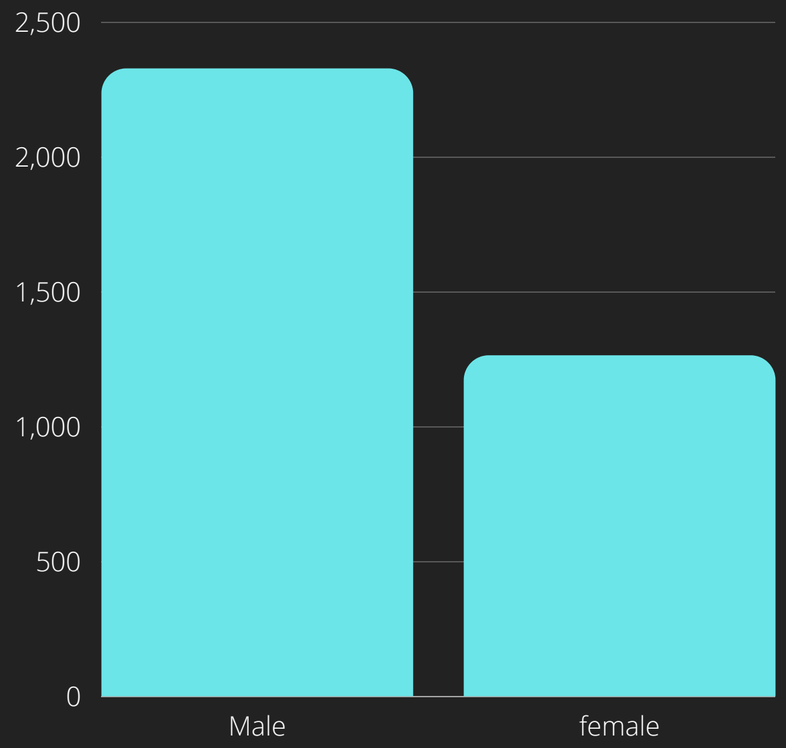


Logged in User vs Not Logged In Users



Annotators Summary

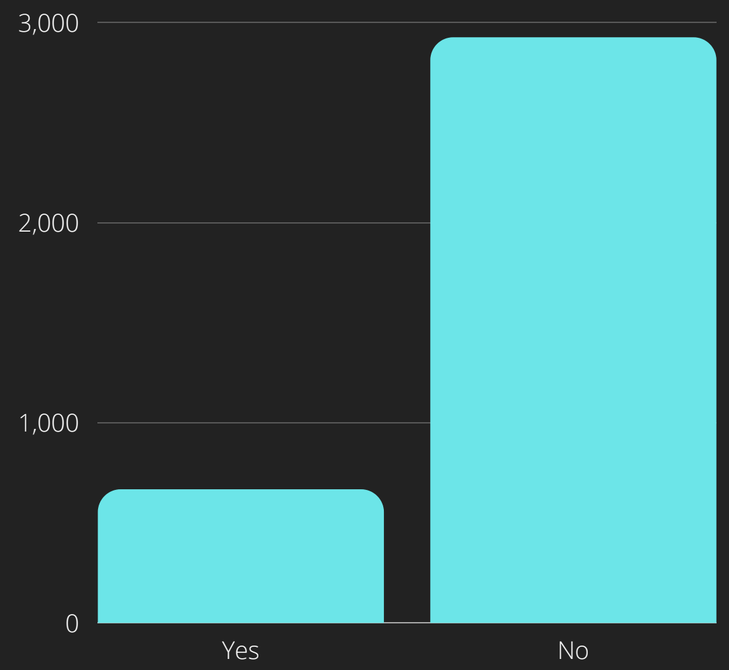
Male vs Female Annotators



73% of the times a comment was found toxic by Females but not Males

33% of the times a comment was found toxic by Non English first language users but not by English as first language

English First Language Yes /No



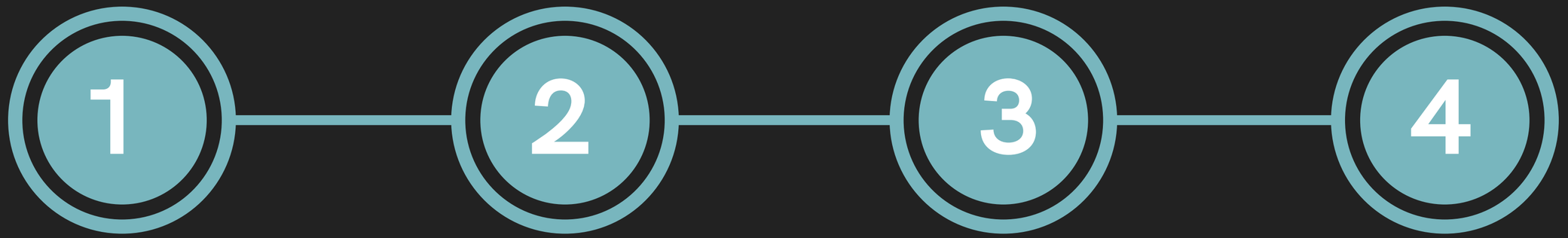
Basic NLP Model Methodology

Representation:

- TFIDF/Count Vector
- Tokenization
- Text to sequence
- Padding Sequences

Deployment:

- Prediction and model evaluation



Cleaning the text:

- Removing stop-words
- Removing irrelevant text to our analysis
- Stemming and Lemmatization

Modelling:

- Logistic Regression
- Naive Bayes
- MLP
- LSTM
- GRU

Word Representation Methods

Count Vectorizer

- Count Vectorizer is a way to convert a given set of strings into a frequency representation.
- Major Disadvantages of it are:
  - Its inability in identifying more important and less important words for analysis.
  - It will just consider words that are abundant in a corpus as the most statistically significant word.
  - It also doesn't identify the relationships between words such as linguistic similarity between words such as House and Home

Term Frequency — Inverse Document Frequency (TF-IDF)

$$\text{tfidf}_{i,j} = \text{tf}_{i,j} \times \log\left(\frac{\text{N}}{\text{df}_i}\right)$$

$\text{tf}_{i,j}$ = total number of occurences of i in j
df_i = total number of documents (speeches) containing i
N = total number of documents (speeches)

- Converts text documents to Matrix form, where each document is converted to a row of the TF-IDF matrix and each word is stored in a column vector.
- TF-IDF is better than Count Vectorizers because it not only focuses on the frequency of words present in the corpus but also provides the importance of the words.
- TFIDF is based on the logic that words that are too abundant in a corpus and words that are too rare are both not statistically important for finding a pattern.

Word Representation Methods

Global Vectors for Word Representation (GloVe)

- GloVe is an unsupervised learning algorithm for obtaining vector representations for words.
- Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.
- The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words.
- For eg: ice co-occurs more frequently with solid than it does with gas, whereas steam co-occurs more frequently with gas than it does with solid. But both words co-occur with their shared property water frequently.

Reference: *Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. [pdf] [bib]

Model 1: Logistic Regression

- We have used following methods to convert text to vector space
  - 1. TFIDF
  - 2. Count Vector
  - 3. GLOVE vector

08

Results

Word Representation	Precision		F1 Score	
	Train	Test	Train	Test
Count Vector	97.84%	62.45%	98.84%	72.74%
TFIDF	56.33%	49.37%	71.39%	65.12%
GloVe	65.81%	64.06%	74.49%	72.84%

Model 2: Naive Bayes

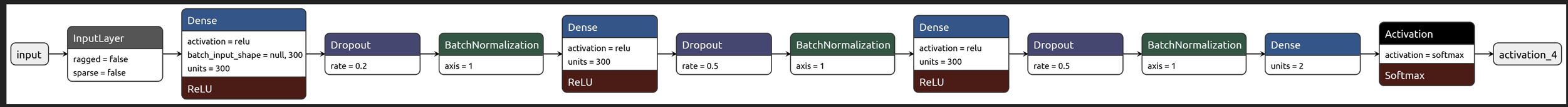
- We have used following methods to convert text to vector space
  - 1. TFIDF
  - 2. Count Vector
  - 3. GLOVE vector

08

Results

	Precision		F1 Score	
Word Representation	Train	Test	Train	Test
Count Vector	88.67%	52.52%	93.62%	60.58%
TFIDF	81.61%	56.68%	86.76%	66.52%
GloVe	11.15%	11.70%	22.05%	23.14%

Model 3: MLP



Embeddings: GloVe

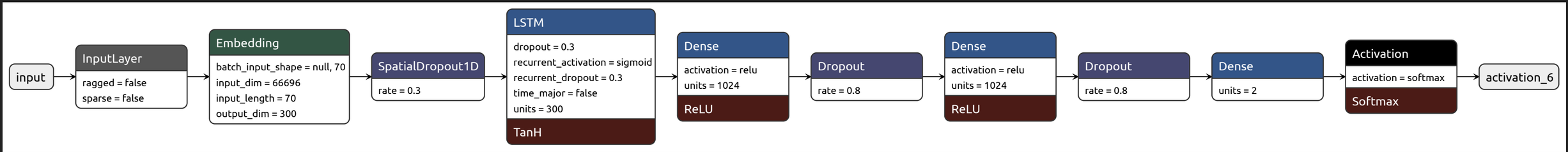
Loss: Categorical Crossentropy

Optimizer: Adam

Results

Word Representation	Precision Score		F1 Score	
	Train	Test	Train	Test
GloVe	81.94%	67.74%	87.09%	74.44%

Model 4: LSTM



Embeddings: GloVe

Loss: Categorical Crossentropy

Optimizer: Adam

Results

Word Representation	Precision Score		F1 Score	
	Train	Test	Train	Test
GloVe	82.54%	79.57%	83.50%	80.13%

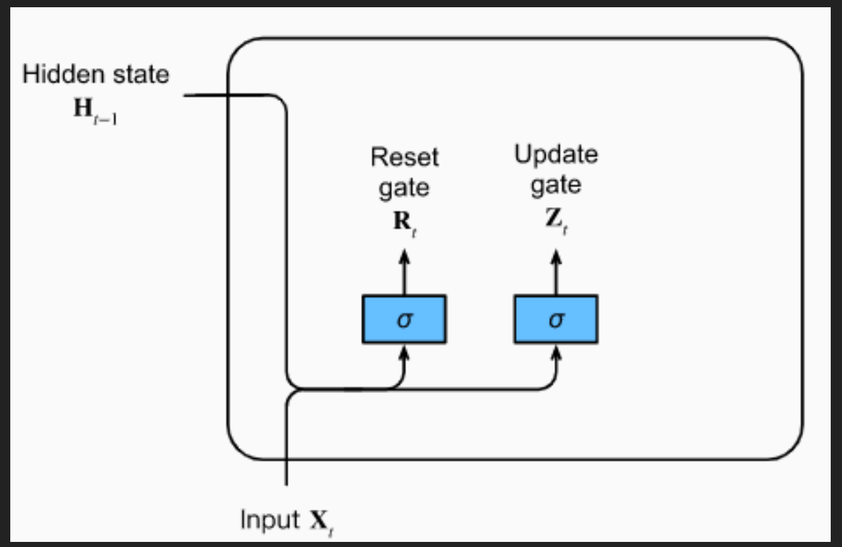
Model 5: GRU

83.39% 76.98% 86.34% 80.31%

Embeddings: GloVe

Loss: Categorical Crossentropy

Optimizer: Adam



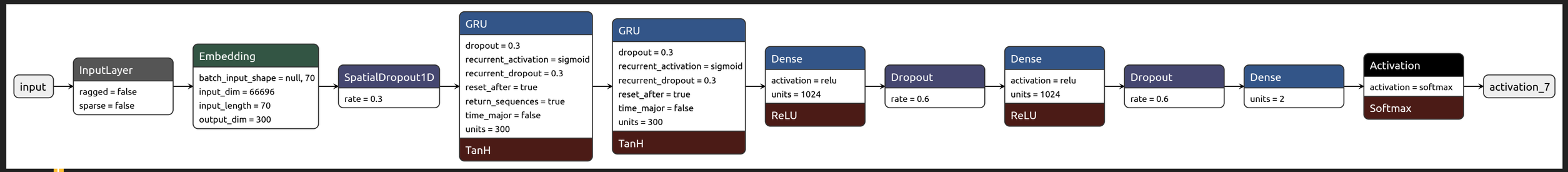
...

$$r_t = \sigma(x_t * U_r + H_{t-1} * W_r)$$

$$u_t = \sigma(x_t * U_u + H_{t-1} * W_u)$$

$$\hat{H}_t = \tanh(x_t * U_g + (r_t \circ H_{t-1}) * W_g)$$

$$H_t = u_t \circ H_{t-1} + (1-u_t) \circ \hat{H}_t$$



Results

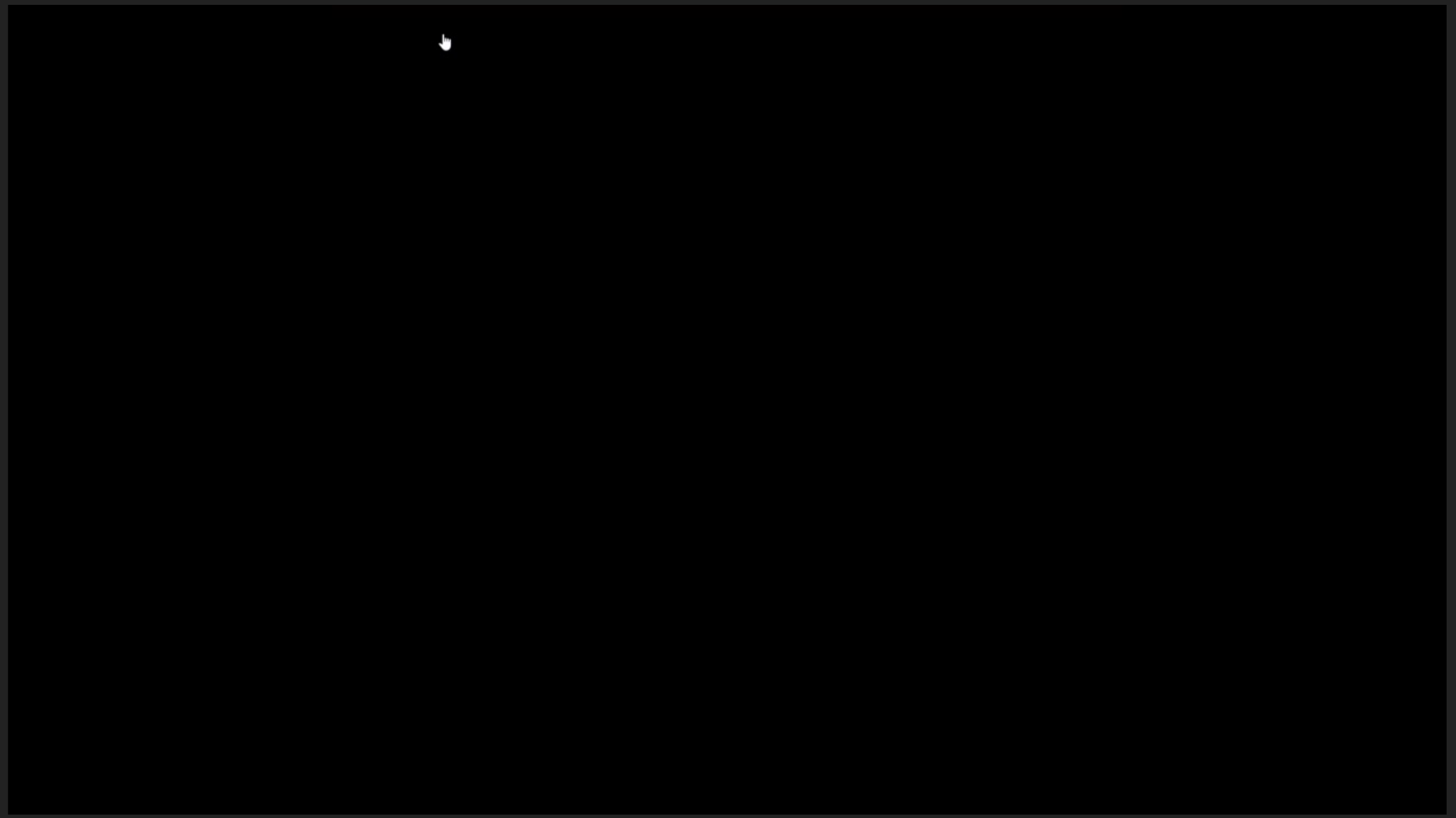
Word Representation	Precision Score		F1 Score	
	Train	Test	Train	Test
GloVe	83.39%	76.98%	86.34%	80.31%

Model Results

08

Model	Word Representation	Precision Score		F1 Score	
		Train	Test	Train	Test
Logistic Regression	Count Vector	97.84%	62.45%	98.84%	72.74%
	TFIDF	56.33%	49.37%	71.39%	65.12%
	GloVe	65.81%	64.06%	74.49%	72.84%
Naive Bayes	Count Vector	88.67%	52.52%	93.62%	60.58%
	TFIDF	81.61%	56.68%	86.76%	66.52%
	GloVe	11.15%	11.70%	22.05%	23.14%
MLP	GloVe	81.94%	67.74%	87.09%	74.44%
LSTM	GloVe	82.54%	79.57%	83.50%	80.13%
GRU	GloVe	83.39%	76.98%	86.34%	80.31%

Model Deployment



05

Future Work

05

Modelling Improvements:

- Reduce the gender, race, education level bias in the model to develop a more unbiased toxicity predictor. Try distribution predictions based on different characteristics of annotators.
- Try character level models to address the out of vocabulary problem common in online comments due to use of slangs and lack of edits.
- Toxicity can be various types such as: Personal Attack, Threat, Identity Hate, Insult etc. Thus next step should be classfying text into different types of toxicity.
- Try attention scoring models such as transformers and find out which type of words have attention scores for particular comment class.
- Develop a user toxicity score based on the overall comments made by the user.

Model Deployment Improvements:

- Create a real time (as one types) toxicity predictor.
- Identify and highlight the words that are making the comment toxic and provide real time suggestions to make it non-toxic/less toxic.

03

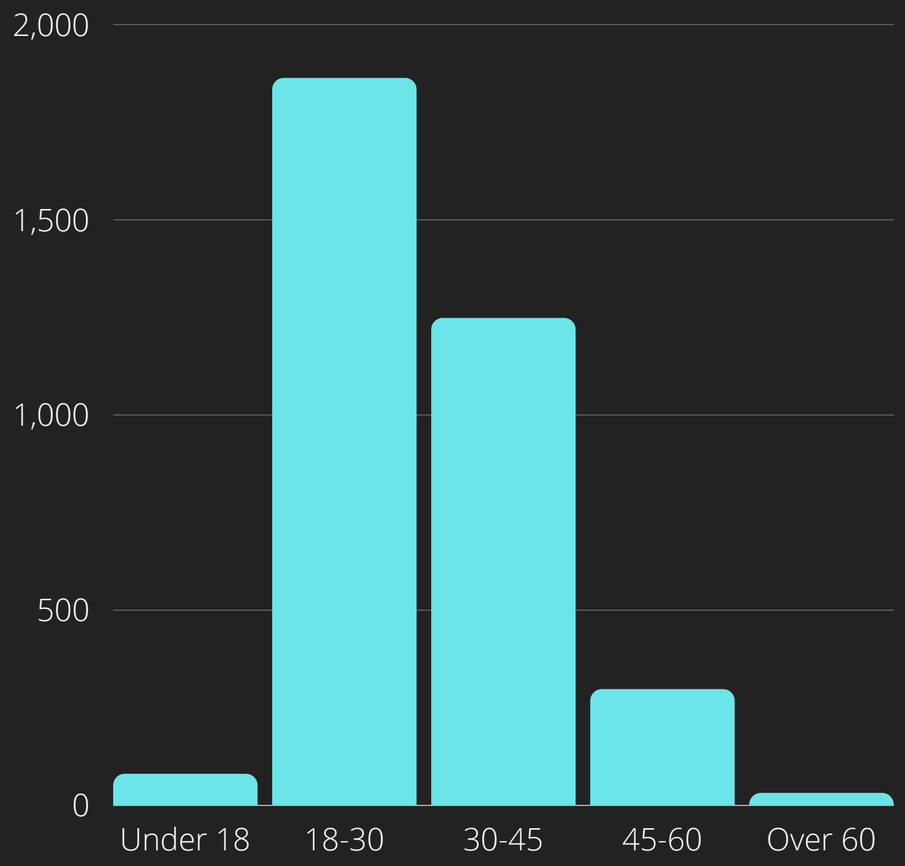
Thank You

References

- Pew Research: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/
- https://scholars.org/contribution/countering-online-toxicity-and-hate-speech
- Facebook Reference: *https://www.facebook.com/communitystandards/bullying
- Instagram Reference: *https://help.instagram.com/547601325292351
- Wikipedia Reference: *https://en.wikipedia.org/wiki/Wikipedia: No_personal_attacks
- https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6856482.pdf

Annotators Summary

Annotators Age Groups



Annotators Education Level