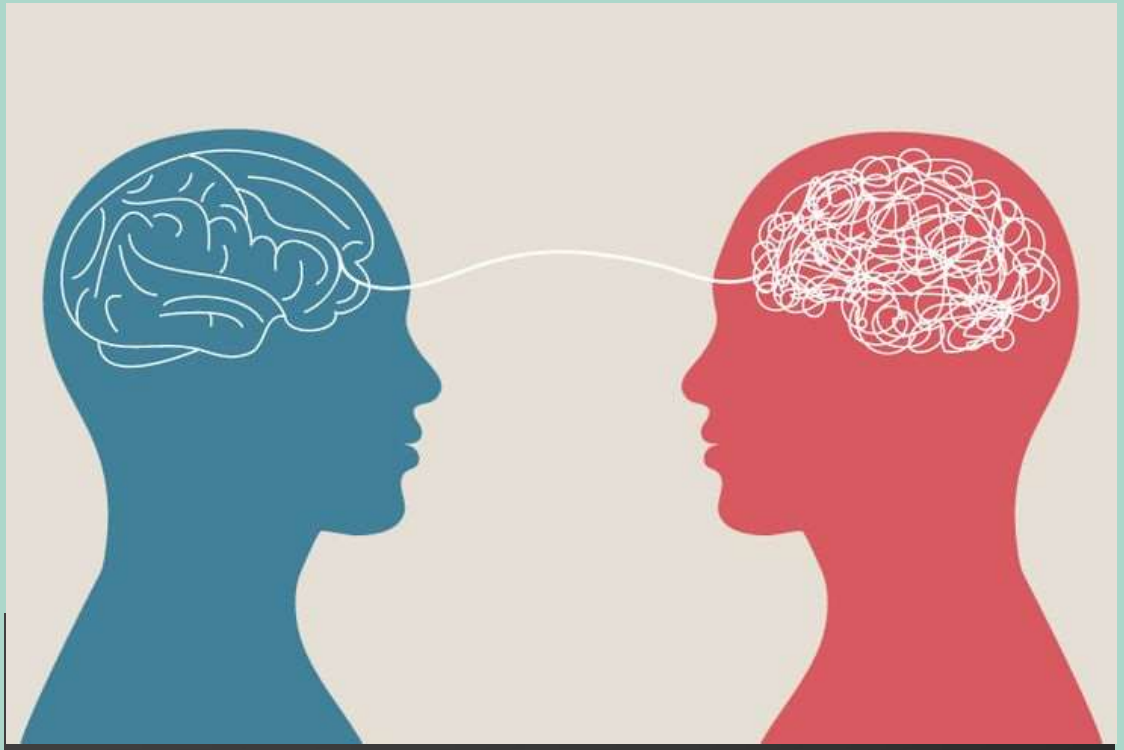
Toxic Language Detection System

GROUP MEMBERS

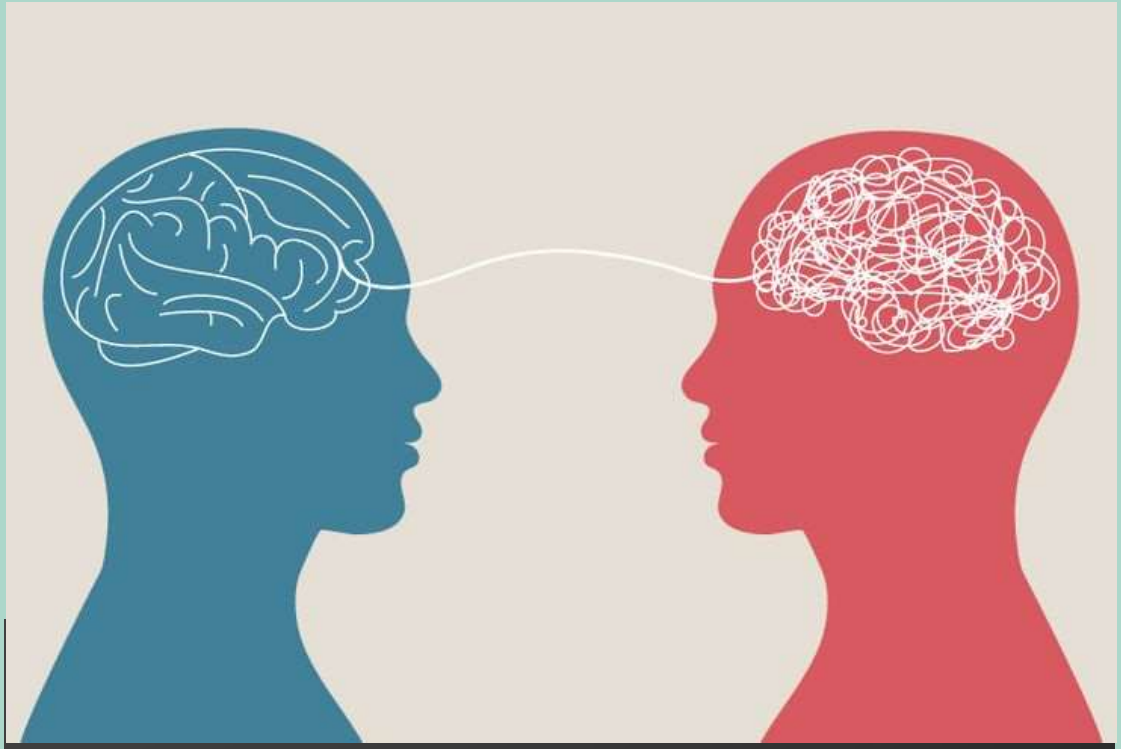
OVERVIEW

- The Toxigen dataset contains user generated sentiments made online targeting minority groups.
- The objective is to analyze the dataset and develop strategies to detect toxic sentiments, improving the accuracy of toxic language detection system



BUSINESS OBJECTIVE

- The primary objective of this project is to develop and refine strategies for detecting toxic sentiments. This will involve building and evaluating classification models to identify toxic language, and ultimately develop accuracy in detecting harmful/negative content.



Problem Statement

Online Toxicity

Negative online interactions can damage the public image leading to decreased engagement from the community

Consequences

Psychological harm, mental health issues, low self-esteem, societal division, damage to the public image of minority groups.

Examples

Cyberbullying, hate speech, harassment, trolling.

Importance

Accurate detection of toxic language without bias is essential for a healthier online environment.

Mitigation Strategies for Online Toxicity

Community Guidelines:

Example: Instagram uses AI to detect and separate harmful comments.

Action: Enforce guidelines to prevent and address toxic behavior.

Counselling Services:

·**Impact:** Address mental health issues caused by online harassment.

·**Action:** Provide access to counseling and create safe spaces for affected individuals.

Stricter Legal Frameworks

·**Action**: Implement legal consequences for cyberbullying and hate speech.

·**Impact**: Reduce the occurrence of online toxicity through enforcement.

Data Privacy

Action: Ensure datasets and models respect user privacy and comply with data protection regulations.

STEPS

Data collection

ToxiGen (toxigen.csv)

Machine-generated dataset for improving toxic language detection.

Pre-processing

Data cleaning, checking null values, check for duplicates, EDA, and checking the datatypes, removing special characters, stopwords, tokenization, normalization

Modelling and Deployment

Implementation of Recurrent Neural Networks (RNNs) as one of the machine learning models used in NLP, for classifying comments as negative or positive.

Analysis and insights

Generation of insights and recommendations for improving toxicity detection systems.

Hypothesis Testing:

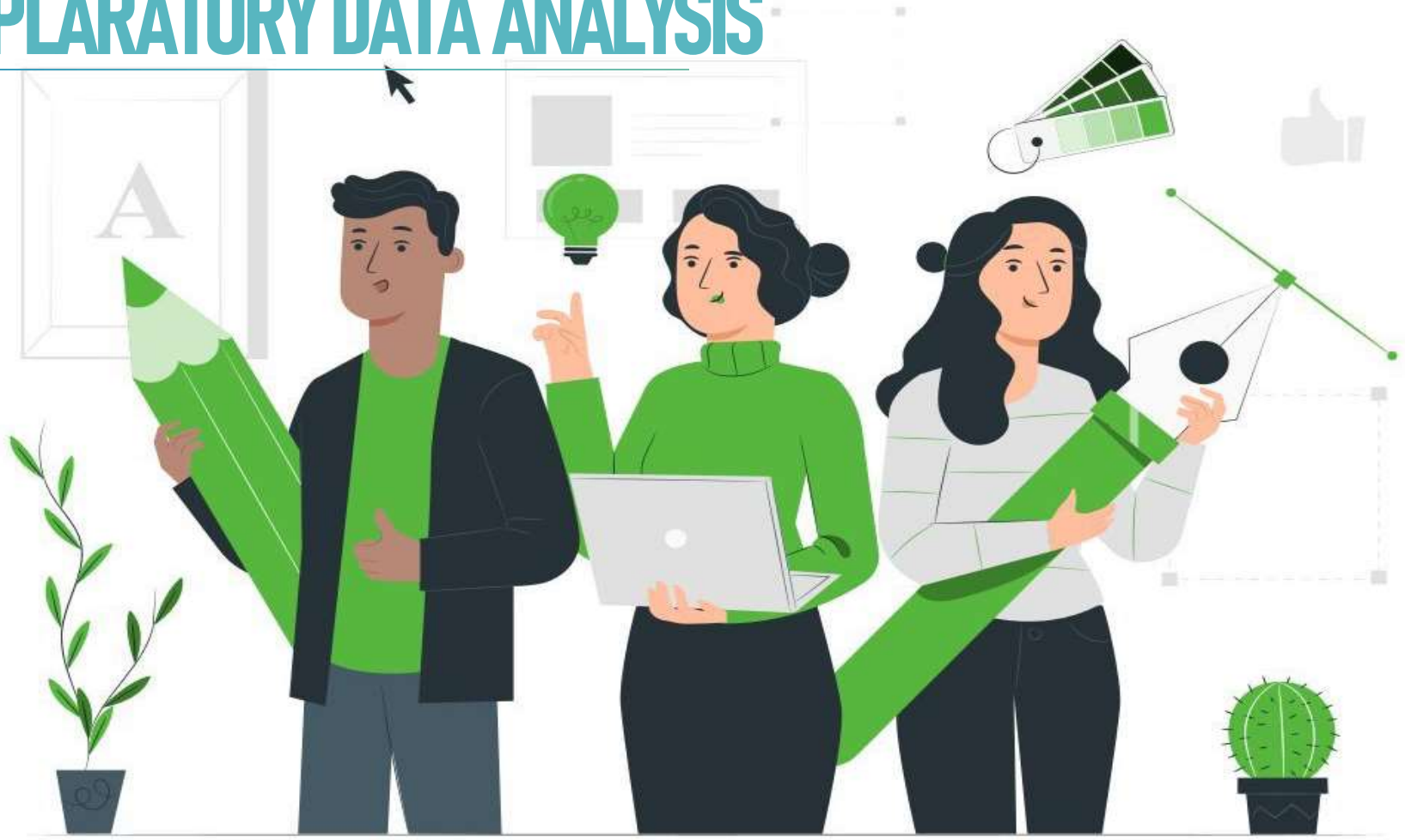
Hypothesis 1:

- **Statement:** Toxic comments are more likely to be associated with specific minority groups.
- **Test:** Comparative analysis of positive vs. negative comments across different minority groups.
- **Method:** Use boxplots, KDE plots, and heat maps.

Hypothesis 2:

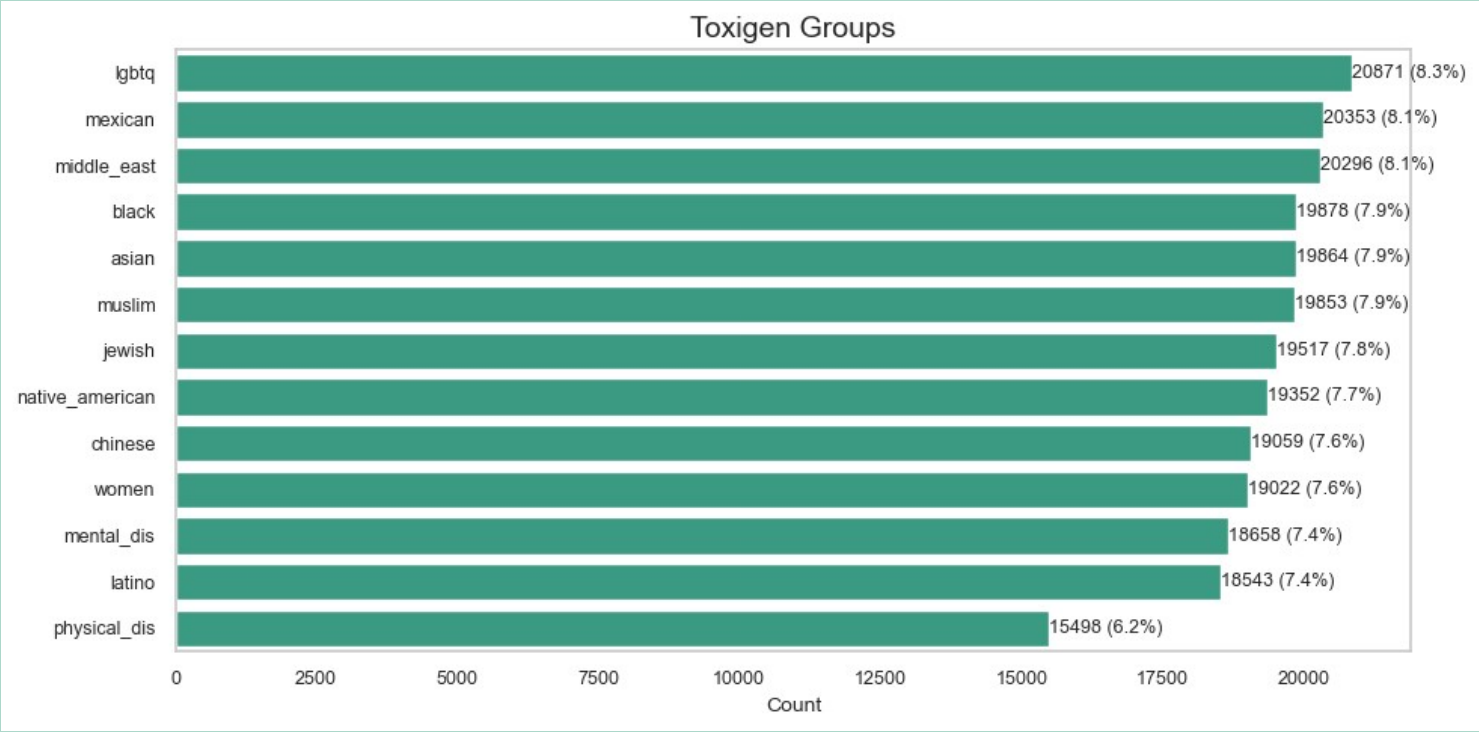
- **Statement:** Toxic reviews contain specific keywords or phrases.
- **Test:** Conduct keyword analysis to identify recurring terms and phrases.
- **Method:**
  Use text mining techniques such as word clouds to extract and analyze common keywords.

EXPLARATORY DATA ANALYSIS



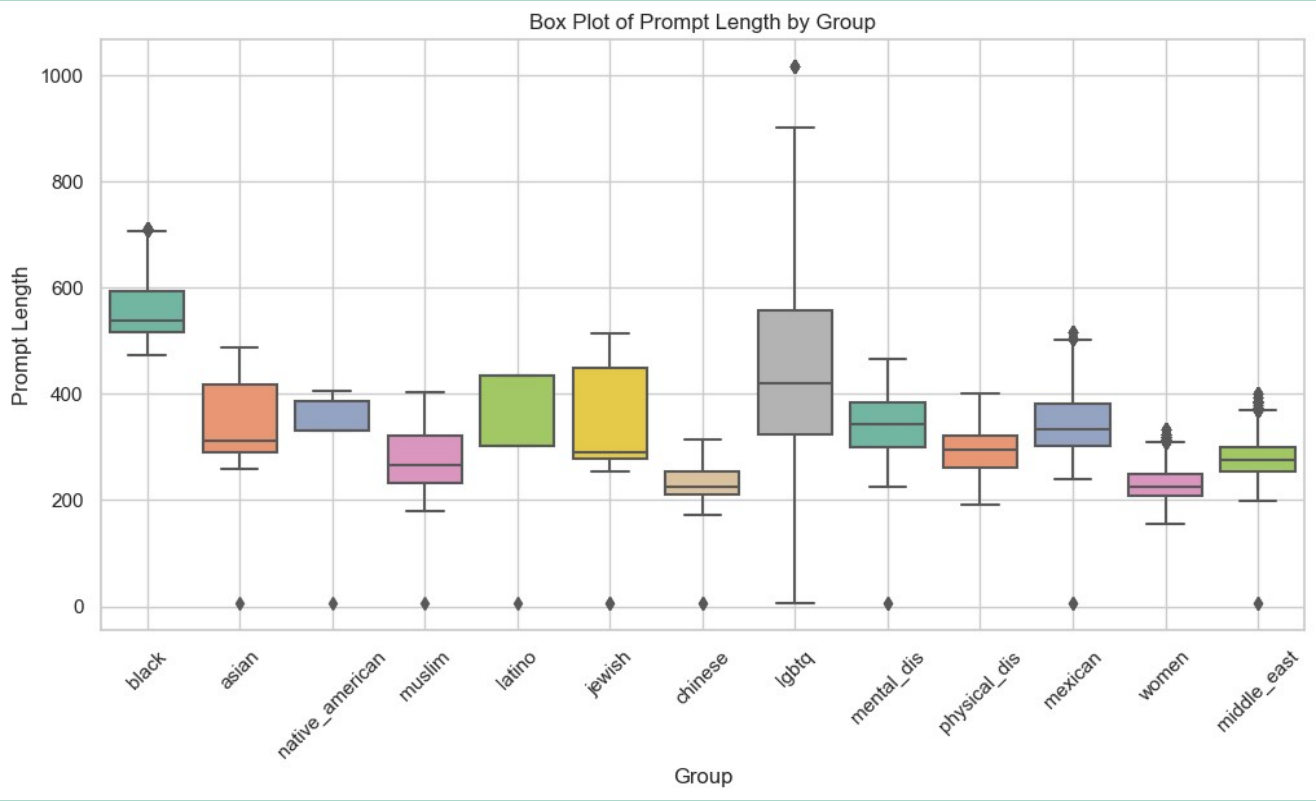
Sentiment Count per Group

- There is a total of 13 distinct groups.
- Lgbtq has the highest number of comments under its group and physically disabled has the least of comments under its group



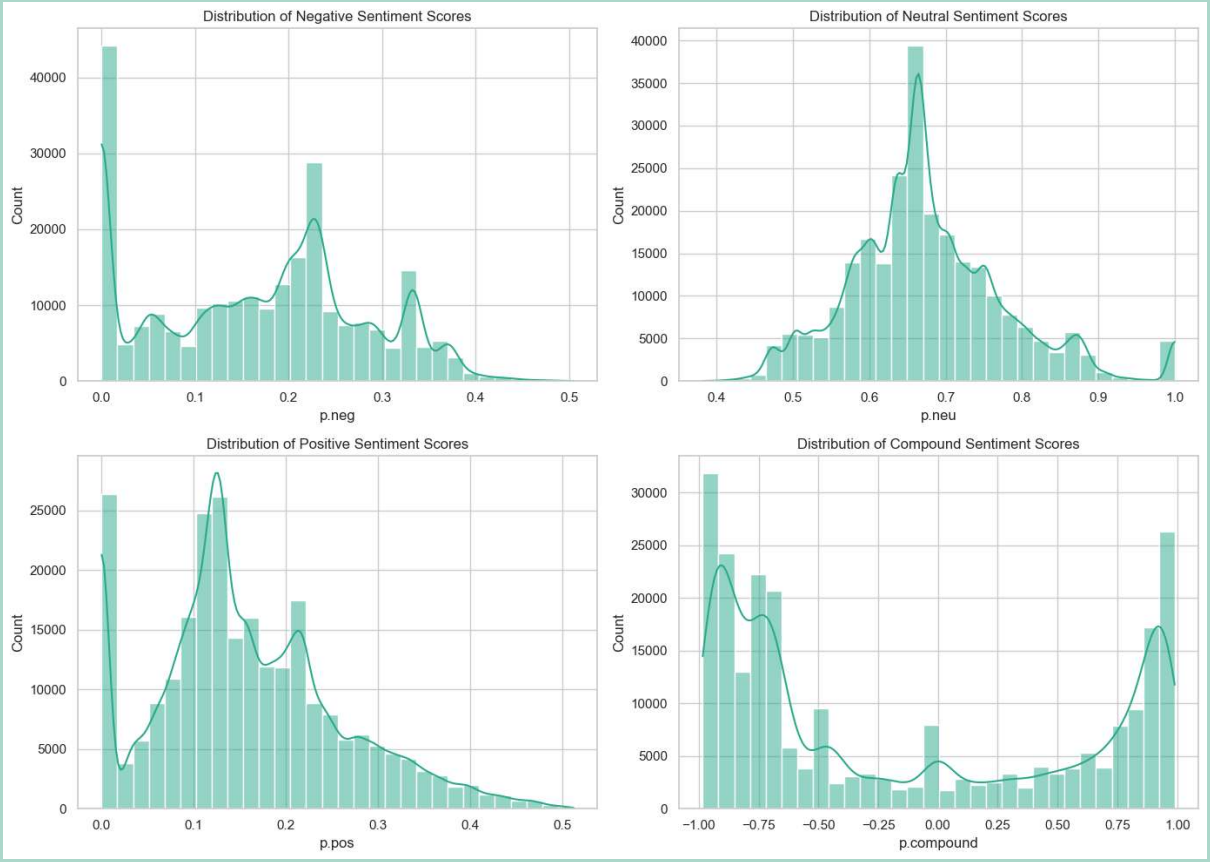
Sentiment (Prompt) Length per Group

The box plot shows that prompt lengths vary significantly across the groups, with the LGBTQ and Black groups having longer prompts, while Women and Muslims have shorter and more consistent prompt lengths.
Outliers are present in several groups, indicating some unusually long prompts compared to the majority.



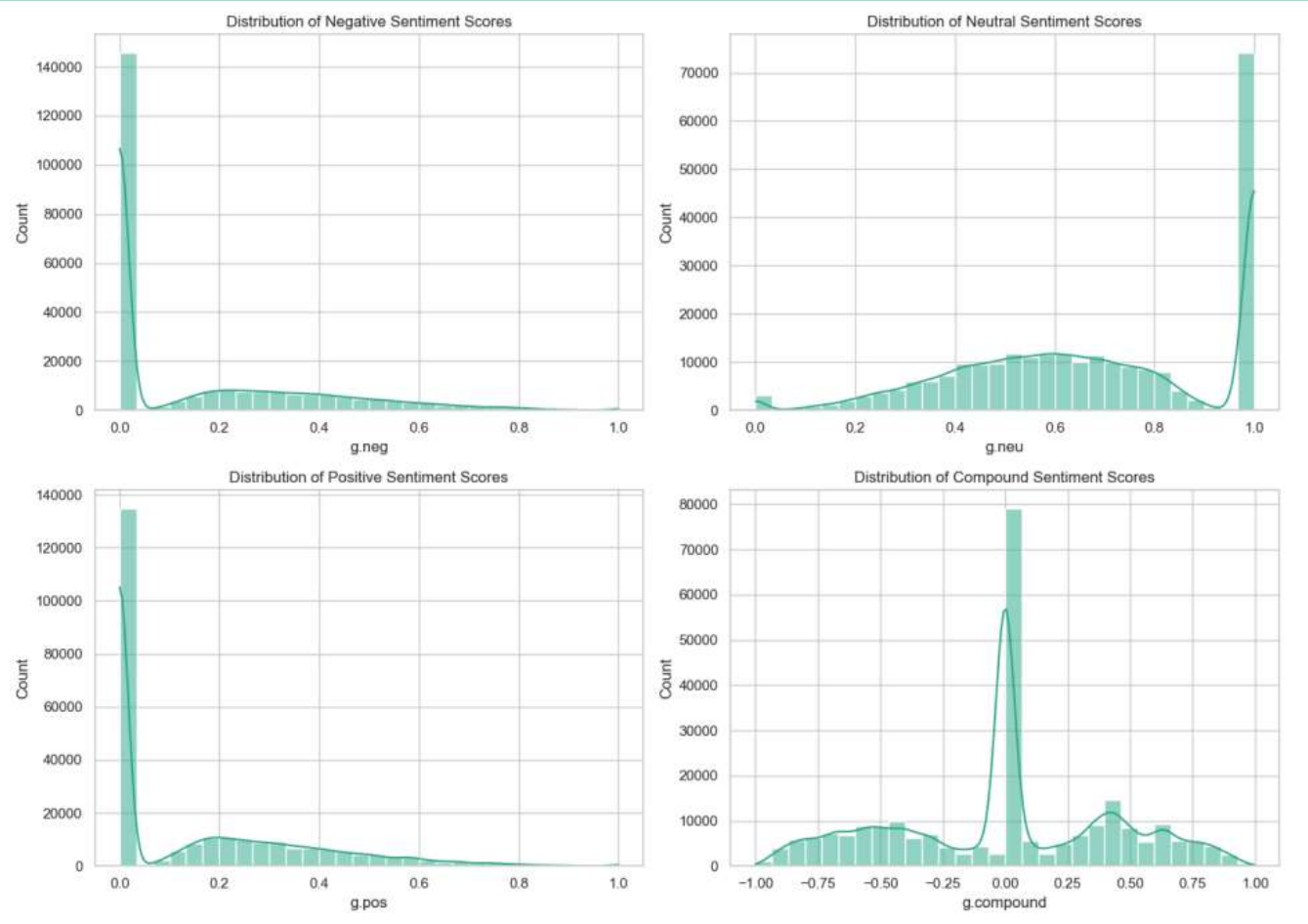
Sentiment Analysis

Prompted sentiment scores



Sentiment Analysis

Generated sentiment scores



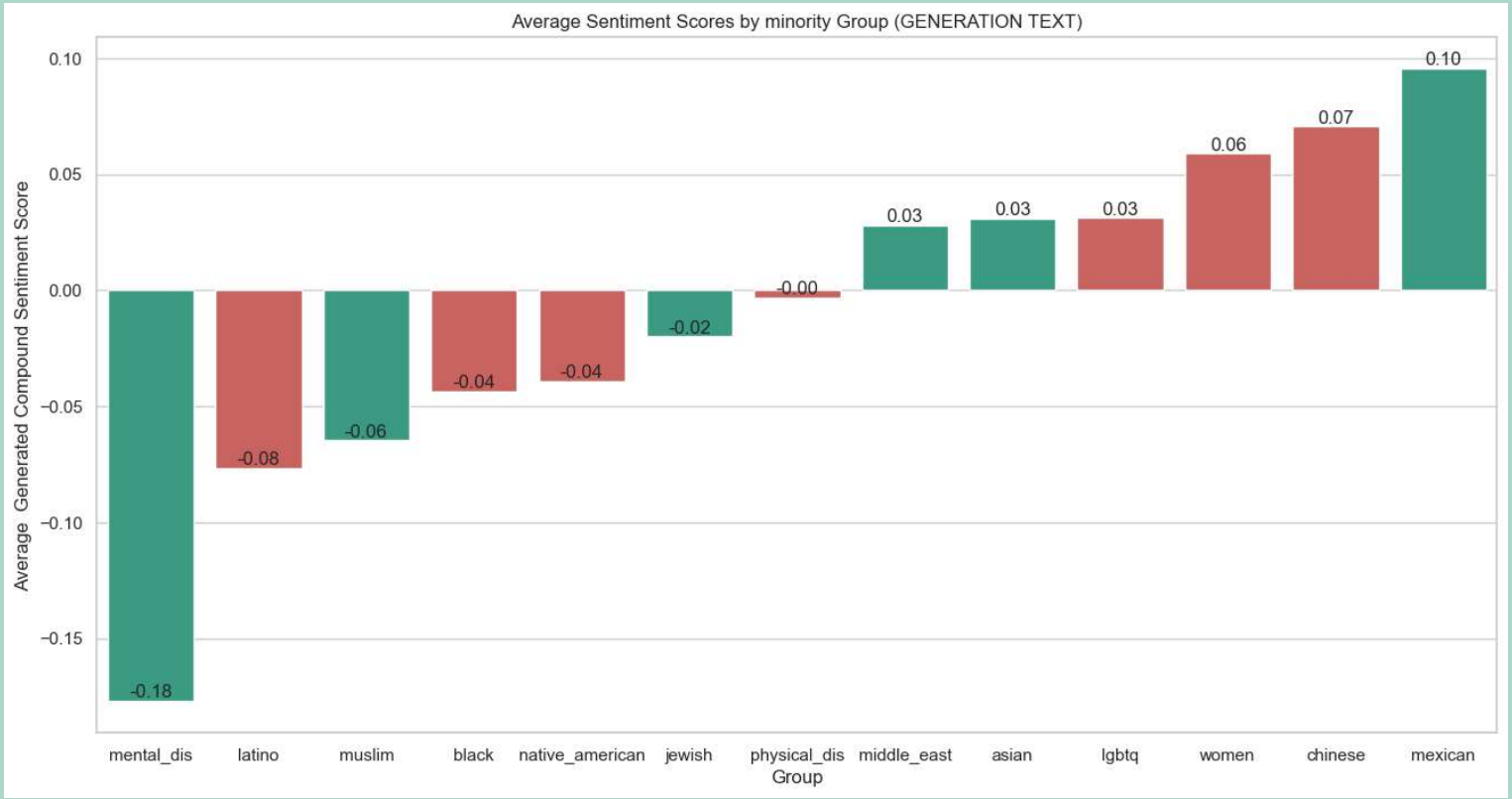
Sentiment Analysis

This is showing the positive versus the negative comments and the average sentiment score for each group under prompt column.

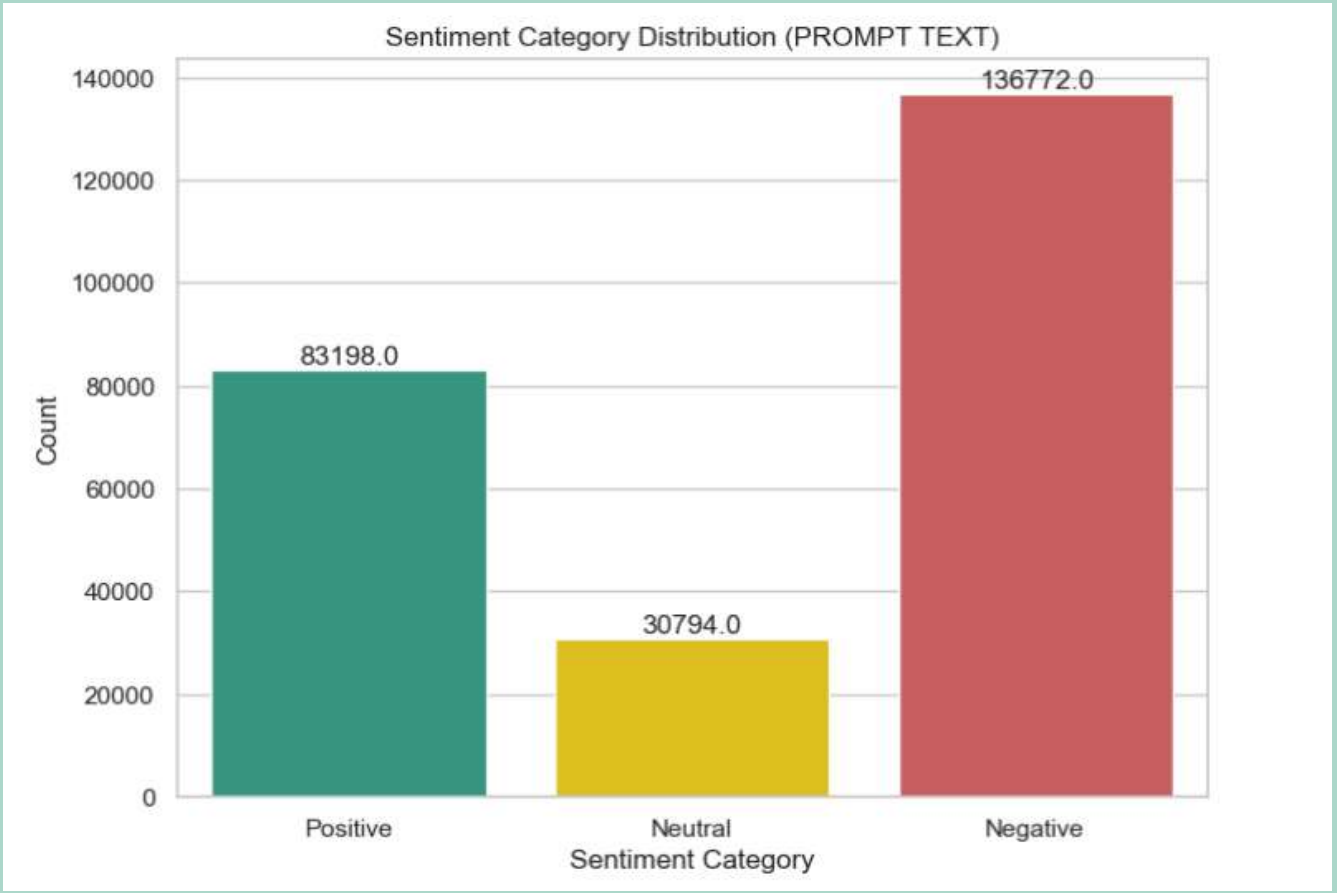
Native American has the highest negative scores while women have the highest positive scores

Sentiment Analysis

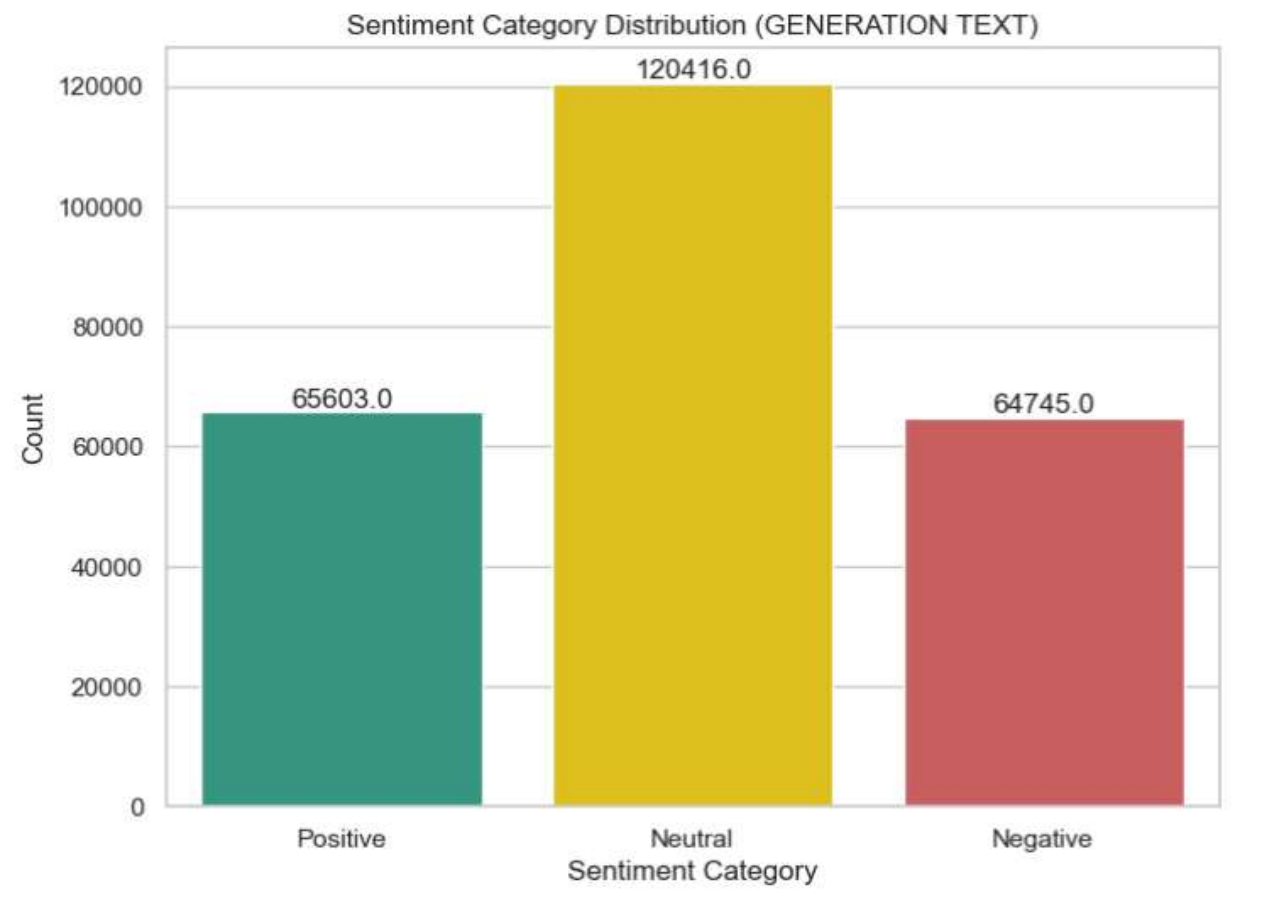
Mentally Disabled have the highest negative scores while Mexicans have the highest positive scores under the generation column.



Distribution of Sentiment Category Under Prompt Column.



Distribution of Sentiment Category Under Generation Column.



A L P H A B E T

Modelling

Logistic Regression

	Precision	Recall	F1-score	Support
0	1.00	0.96	0.98	27467
1	0.76	0.96	0.86	6190
2	1.00	0.96	0.98	16496
Accuracy			0.96	50153
Macro Avg	0.92	0.97	0.94	50153
Weighted Avg	0.97	0.96	0.96	50153

After Hyper-parameter tuning

	Precision	Recall	F1-score	Support
0	0.99	0.99	0.99	27467
1	0.95	0.94	0.94	6190
2	0.99	0.99	0.99	16496
Accuracy			0.99	50153
Macro Avg	0.98	0.97	0.98	50153
Weighted Avg	0.99	0.99	0.99	50153

a b c d e f g h i j k l m
n o p q r s t u v w x y z

A L P H A B E T

Modelling

Gradient Boosting Model

	Precision	Recall	F1-score	Support
0	0.89	0.98	0.93	27467
1	0.90	0.47	0.62	6190
2	0.93	0.92	0.92	16496
Accuracy			0.90	50153
Macro Avg	0.90	0.79	0.83	50153
Weighted Avg	0.90	0.90	0.89	50153

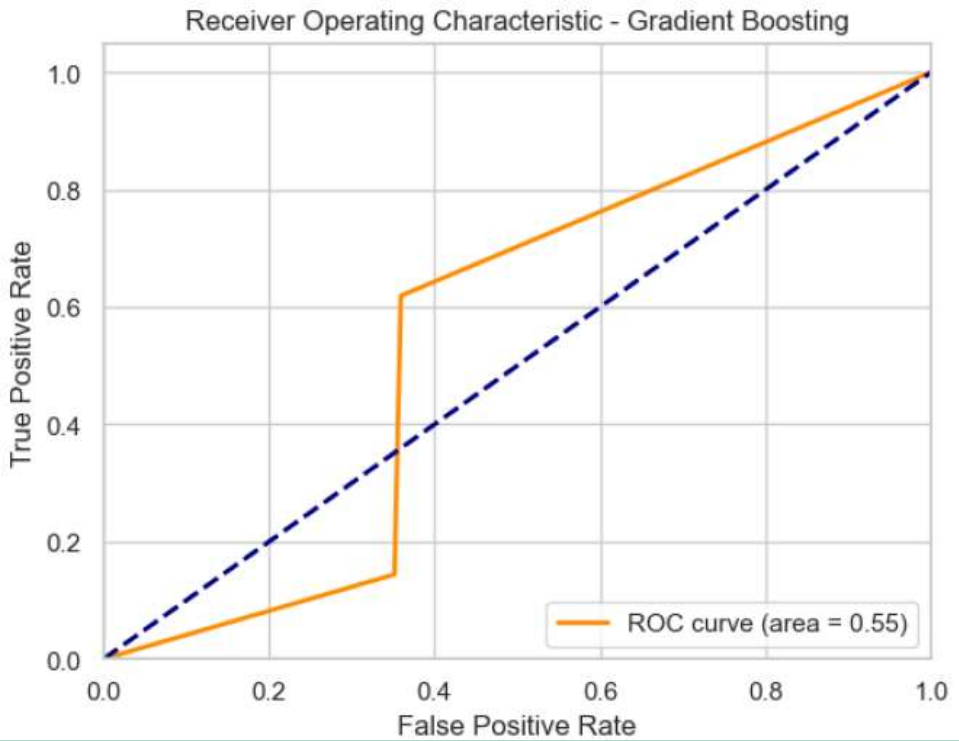
Neural Network Model

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	27467
1	0.99	0.99	0.99	6190
2	1.00	1.00	1.00	16496
Accuracy			1.00	50153
Macro Avg	1.00	1.00	1.00	50153
Weighted Avg	1.00	1.00	1.00	50153

a b c d e f g h i j k l m
n o p q r s t u v w x y z

A L P H A B E T

Gradient Boosting ROC/AUC



An AUC of 0.55 suggesting that the model is not performing well and may require improvement.

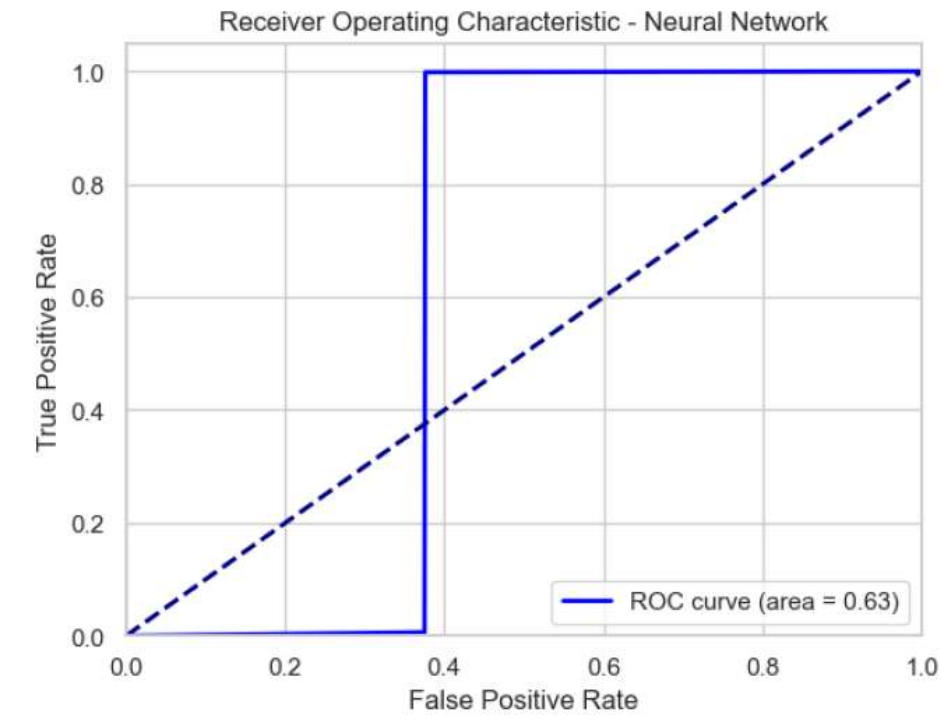
The model has very limited ability to distinguish between the positive and negative classes. It is only marginally better than random chance.

This could involve re-evaluating the model choice, feature selection, preprocessing, or adjusting hyperparameters.

a b c d e f g h i j k l m
n o p q r s t u v w x y z

A L P H A B E T

Neural Networks ROC/AUC



An AUC of 0.63 indicates that the model has a moderate ability to distinguish between different classes, performing better than random guessing but still leaving room for improvement.

The AUC in Neural Networks is higher than in Linear regression.

a b c d e f g h i j k l m
n o p q r s t u v w x y z

A L P H A B E T

Modelling

Recurrent Neural Network

	Precision	Recall	F1-score	Support
0	1.00	0.99	1.00	27467
1	0.97	0.95	0.96	6190
2	0.99	1.00	0.99	16496
Accuracy			0.99	50153
Macro Avg	0.98	0.98	0.98	50153
Weighted Avg	0.99	0.99	0.99	50153

The precision in class 0 (negative) is 100% , the recall is 99% and the f1 score is 100%.

This shows that the model is very good in predicting negative words.

The precision in class 1(neutral) and class 2(positive) is also high proving the models good performance.

a b c d e f g h i j k l m
n o p q r s t u v w x y z

Conclusion

Performance

Logistic Regression:

Accuracy: 97.46% and after Hyperparameter Tuning accuracy at 99%

Classification Report: High precision, recall, and F1 scores across all classes. Interpretation: Logistic Regression performed very well, indicating that it effectively distinguishes between classes. The high accuracy and balanced metrics across different classes suggest that the model is reliable and generalizes well.

Gradient Boosting:

Accuracy: 90.02%

Classification Report: Precision: High for some classes but lower for others. Recall: Varies significantly across classes. F1 Score: Lower for class 1(neutral)compared to Logistic Regression. Interpretation: Gradient Boosting shows strong performance but has some imbalance in precision and recall across classes. This indicates that while the model is generally good, it struggles more with certain classes compared to Logistic Regression. Further tuning might be needed.

Neural Networks:

Accuracy: 99.7%

Classification Report: Precision, Recall, and F1 Score: Generally high, similar to Logistic Regression, but may vary based on model complexity and hyperparameters. Interpretation: Neural Networks performed very well and this is especially because we had a large dataset.However,it may require more careful tuning of hyperparameters.

Model Deployment

The model was deployed Locally using streamlit app

Deployment Features

The Interface has open ended questionnaires which allow user to give feedback and input text for the model to predict Toxicity

Get To Know Us!

Home Page

About Us

Contact Us

Send Anonymous Report

Please answer the following questions:

What do you think about the evolution of Tech and the New forms of online Interaction?

it is too much

How do you feel when you see negative comments online?

very worried

What do you think can be done to reduce online toxicity?

report them

Toxicity Level Analysis

Input a comment you want us to analyze for toxicity:

we are all okey and friendly

Predict

Recommendations

Logistic Regression is our deployment of choice due to its high accuracy and balanced performance metrics.

Exploration of advanced NLP techniques like BERT, to potentially improve sentiment analysis accuracy.

Implement a feedback loop where the model continuously learns from new data by retraining the model periodically with new labeled data from real-world usage.

Prepare for scalability by designing a flexible deployment infrastructure that can handle increasing data volumes and user interactions efficiently.

Periodically revisit the model to check if performance can be improved with additional data or feature engineering.

Thank You

https://github.com/Njoroge-Mwaura/Group-12-Final-NLP-Project/tree/main