Mental Health of Today's Youth

Prepared by: Deenie
Jacob
Jamin
Nicholas

Table of contents

01 Problem Statement

02 Data Selection & EDA

03 Modeling

04 Conclusion & Future Works

01

Problem Statement

Background

Major concerns of Online toxicity and different forms of abuse experienced by youths.

- Offensive name calling (41%)
- Physical threats (17%)
- Sexual harassment (16%)

Common venues of abuse experienced

- Social platforms
- Streaming services

Problem Statement

Ministry of Health (MoH)'s initiative, MindSG, to promote cyber wellness targeted at youth.

Curative Measures

- Helpline
- School Counselling Services
- Self-help Resources

Preventive Measures

Data Scientist in MOH to:

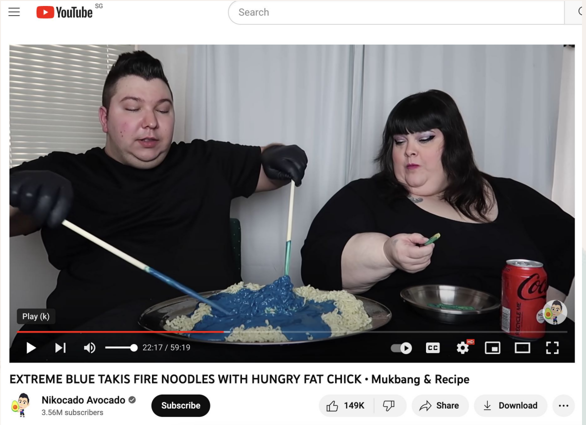
Classifier to detect hateful comments

Data Collection, Cleaning & EDA

02

Data Collection

1. Video Selection



Controversial Youtuber that focus on "Mukbang" (similar to binge-eating)

3.56m followers, 14m views (for this video)

2. Data Extraction

Comments extracted using Youtube Data API v3

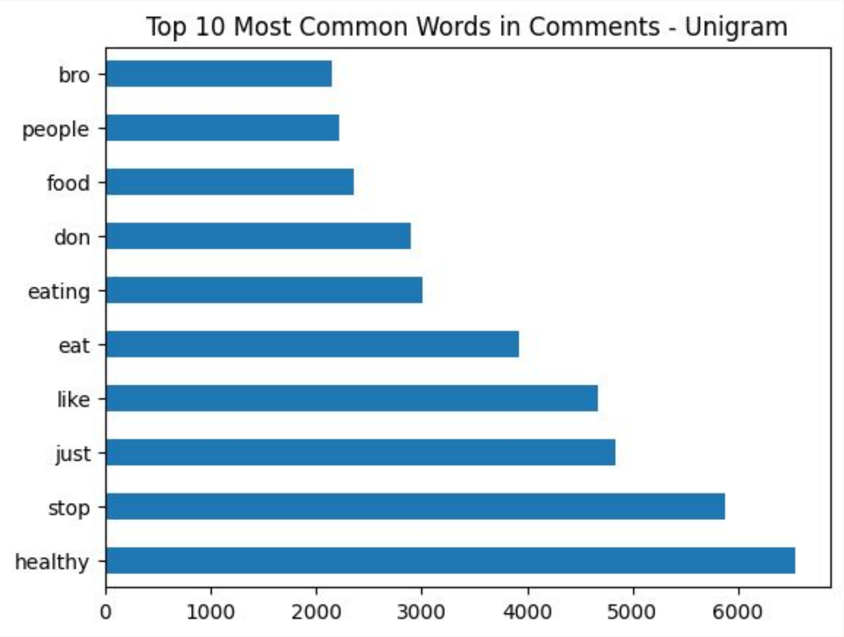
Total of 67487 comments with 14 features was collected through custom functions and automation.

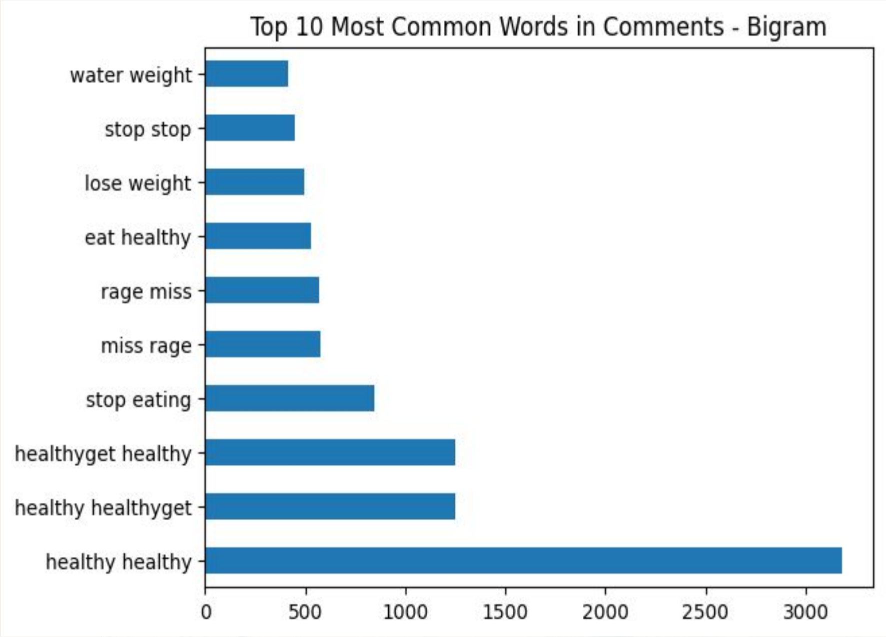
Data Cleaning & EDA

Missing Values, Duplicates, data type checks, column renaming, etc.

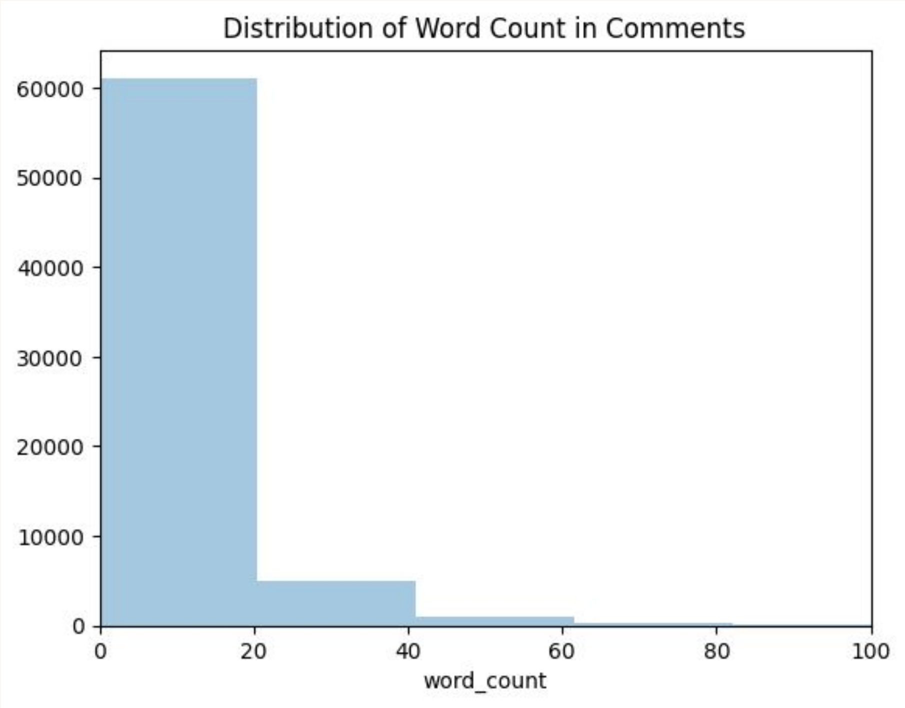
Removed emoji with demoji and URLs and special characters with Regex

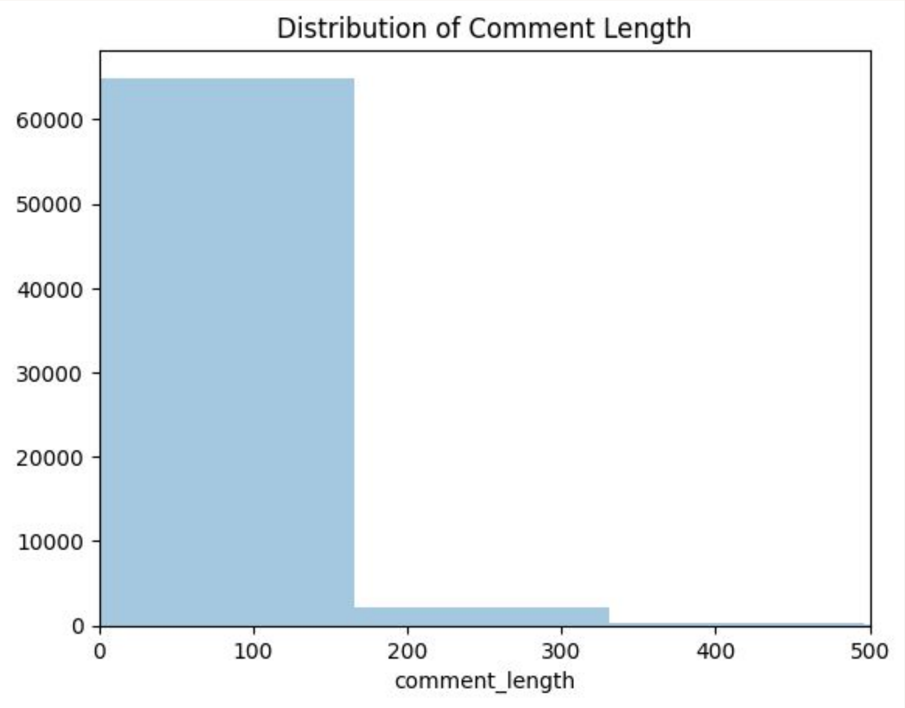
Removed stopwords, tokenized and lemmatized





EDA (Cont'd)


Distribution of Word Count in Comments
60000
50000
40000
30000
20000
10000
0
0
20
40
60
80
100
word_count


Distribution of Comment Length
60000
50000
40000
30000
20000
10000
0
0
100
200
300
400
500
comment_length

Data Labeling

Perspective

Used Google's Perspective API to obtain toxicity scores of ˜5636 comments

Label comments as toxic (1) or not toxic (0) based on toxicity scores.

Used as 'y values' for modeling

Modelling

03

Both Precision and Recall are important

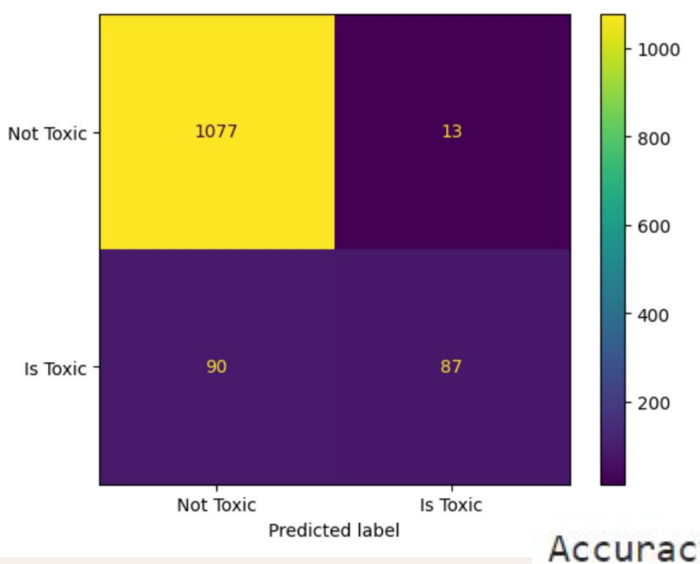
Recall is important as it would measure how many toxic comments are flagged

Precision is import as it would measure how many comments are unnecessarily flagged

Chosen Model: Naive Bayes/Count Vectorize max_features = 500

It has the best F1 score as F1 score is still able to relay true model performance when the dataset is imbalanced.

(15 / 85 Split on Toxic and Non-toxic comments)

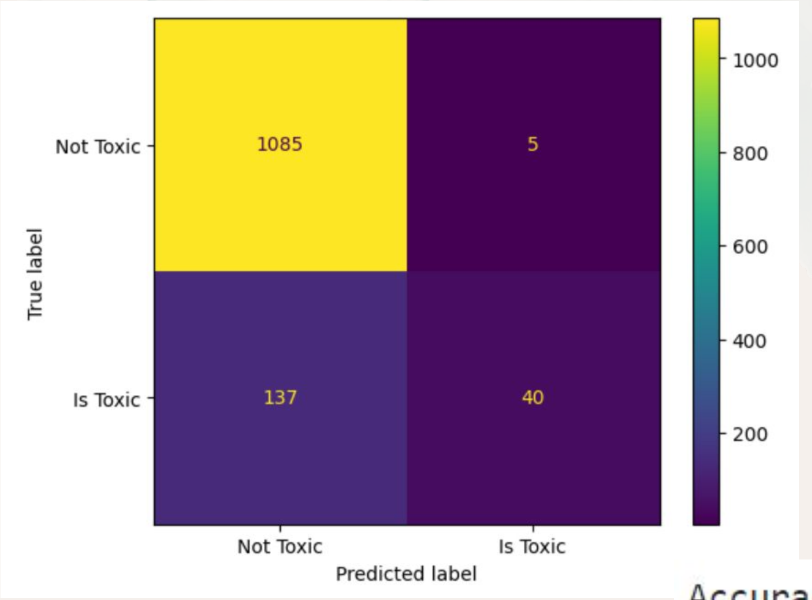


Train Score: 0.927
Test Score: 0.912

Accuracy: 0.9187
Precision: 0.8700
Recall: 0.4915
F1-score: 0.6282

Improvements we tried to make: max_features = 100

Decreased the number of features to reduce overfitting. But the F1-score is decreased, resulting in a model that performs poorer

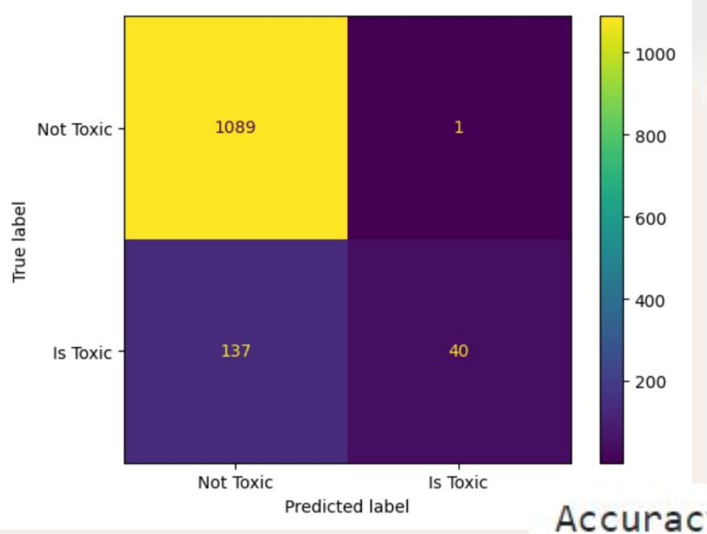


Train Score: 0.883
Test Score: 0.888

Accuracy: 0.8879
Precision: 0.8889
Recall: 0.2260
F1-score: 0.3604

Naive Bayes/Tfidf max_features = 500

Perhaps there are words that keep appearing, TFIDF penalizes those words.



Train Score: 0.892
Test Score: 0.891

Accuracy: 0.8911
Precision: 0.9756
Recall: 0.2260
F1-score: 0.3670

Models Used and Metrics

Classifier Parameters	Naive Bayes Count Vectorizer Max_features = 500	Naive Bayes Count Vectorizer Max_features = 100	Naive Bayes TfidfVectorizer Max_features = 500	Naive Bayes TfidfVectorizer Max_features = 100	Random Forests TfidfVectorizer Max_features=500	Random Forests TfidfVectorizer Max_features=100	GridSearchCV on RF TfidfVectorizer Max_features=100
Train Accuracy	0.927	0.883	0.892	0.871	0.986	0.943	0.943
Test Accuracy	0.9124	0.888	0.891	0.871	0.912	0.883	0.883
Test Precision	0.875	0.889	0.976	1.000	0.785	0.723	0.730
Test Recall	0.435	0.226	0.226	0.079	0.514	0.266	0.260
Test F1 score	0.5811	0.360	0.367	0.147	0.621	0.388	0.383

Conclusion

Our chosen model Naive-Bayes successfully identifies hateful speech related to mental health with a combination of high accuracy (91.2%) and F1 scores, allowing the Ministry of Health to monitor online communities for potentially harmful content.

By detecting and flagging such content early, the Ministry of Health can moderate hateful comments in the online community to mitigate the negative impact on mental health and well-being of individuals who may be targeted by such speech.

This model can serve as a useful tool for mental health advocacy groups and other organizations working to create a safe and supportive online community for people dealing with mental health issues.

The model can be further improved by taking the following steps:

1. more diverse training data and incorporating user-specific data, such as age, gender, or location, to better understand how different demographics are affected by hate speech targeting mental health.
2. Increase the size of the training dataset to improve the model's ability to generalise and lead to better performance on the test data.
3. Feature engineering: Adding new features to the model such as sentiment analysis, word embeddings, or part-of-speech tagging can help the model better understand the context of the comments and improve its accuracy.

Thank you

Remember to take a mental wellness break <3