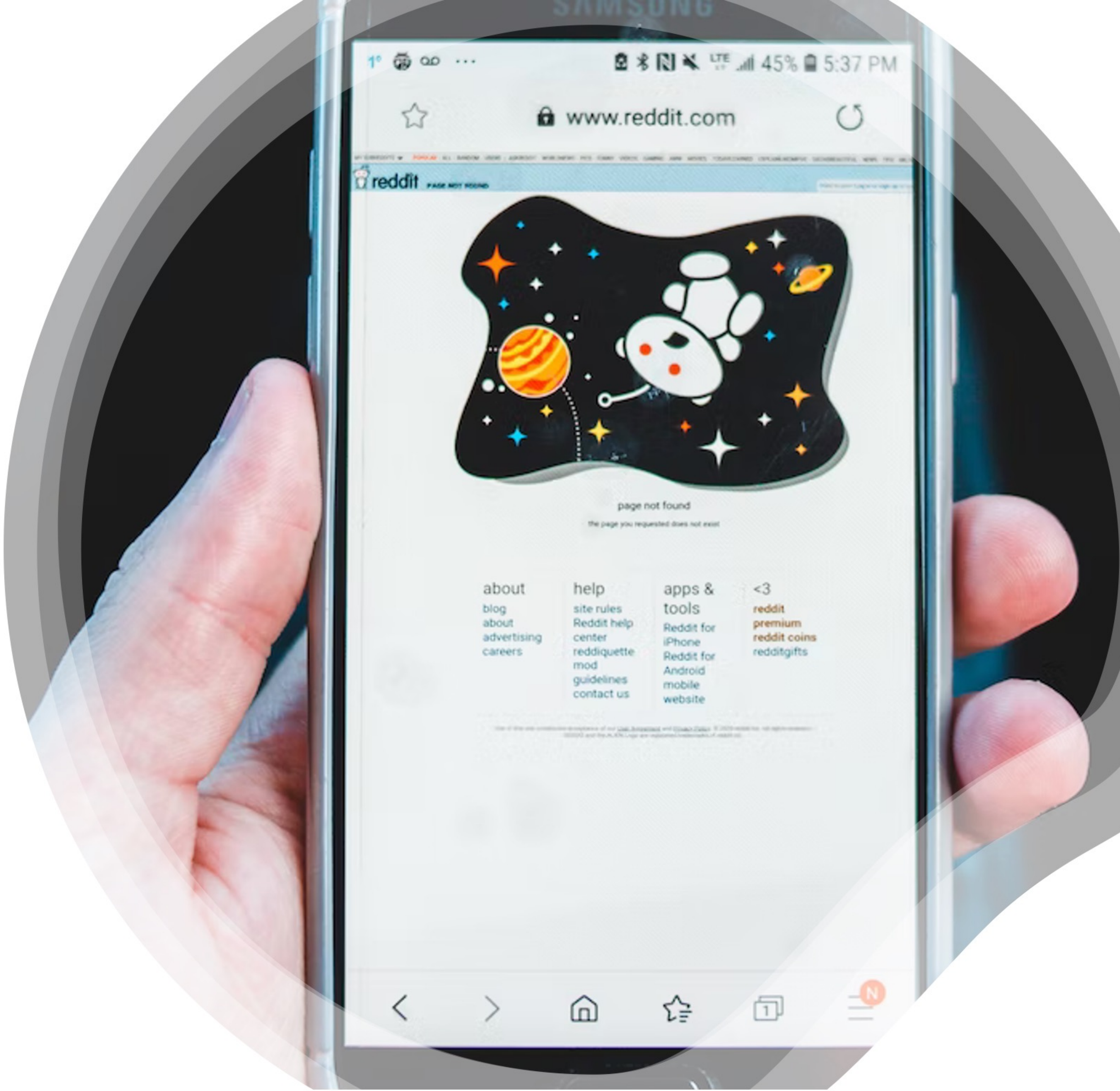
WARNING

The contents of this presentation is based on mental health/illnesses. This includes depression, anxiety, panic attacks and dark thoughts that people face in the world.

Natural Language Processing Binary Classification of Subreddits (r/anxiety and r/depression)

Presented by:
Nicholas Khoo



Presentation Outline

- Background
- Data Science Approach

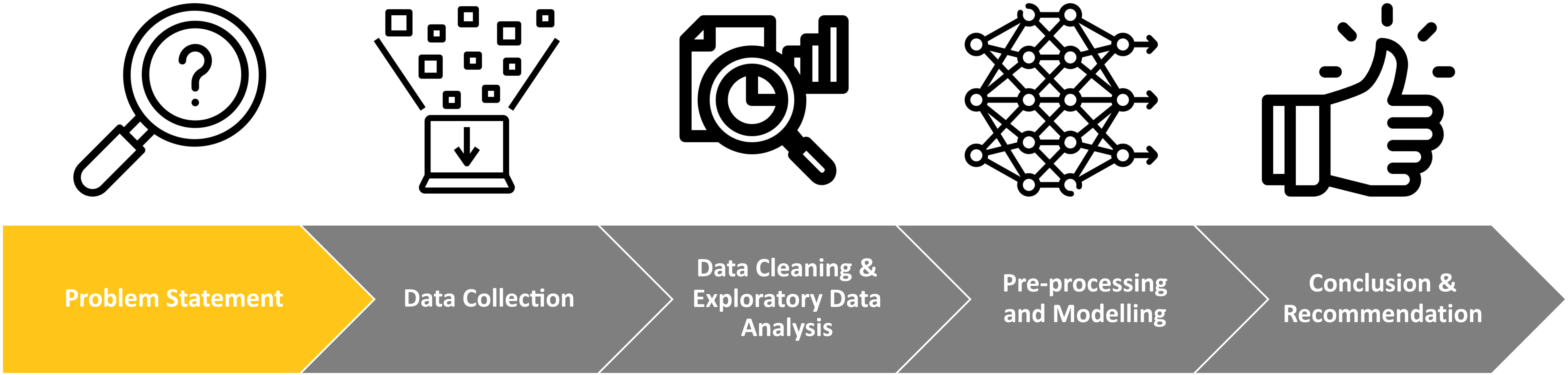
Background

- Reddit is a social networking site with subreddits, including r/depression and r/anxiety, focused on mental health support.
- Proper subreddit categorization is crucial to ensure that users receive appropriate support and guidance.
- Moderators play a crucial role in maintaining a safe and helpful community.

Presentation Outline

- Background
- Data Science Approach

Data Science Approach



Problem Statement

Data Collection

Data Cleaning & Exploratory Data Analysis

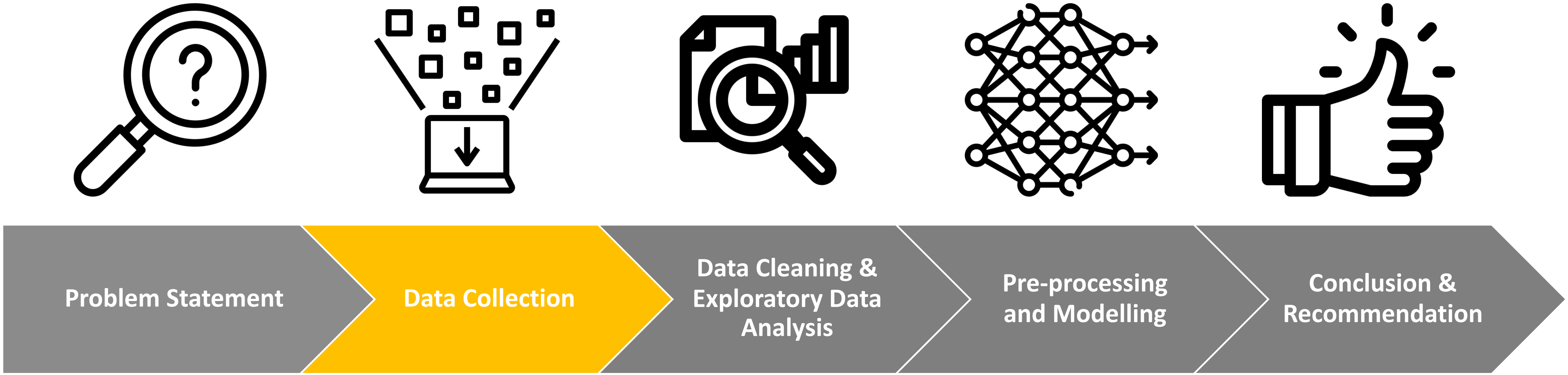
Pre-processing and Modelling

Conclusion & Recommendation

Problem Statement

How can the moderators of r/depression and r/anxiety improve the classification of users' posts to ensure their communities remain a safe and supportive space?

Data Science Approach



Problem Statement

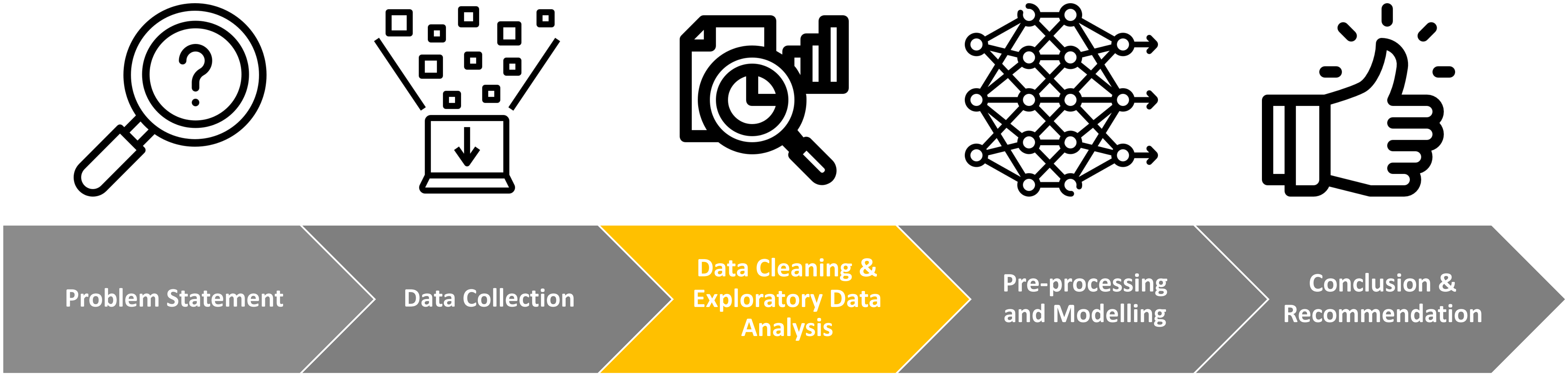
Data Collection

Data Cleaning & Exploratory Data Analysis

Pre-processing and Modelling

Conclusion & Recommendation

Data Science Approach



Problem Statement

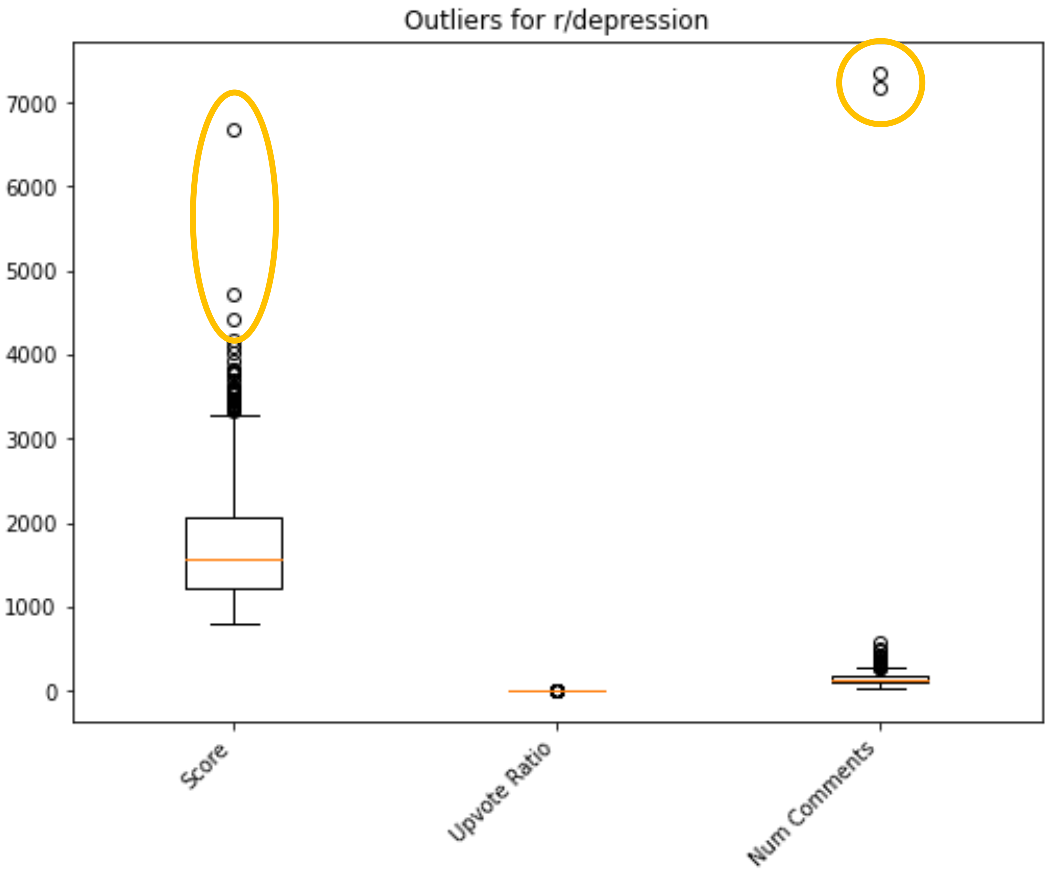
Data Collection

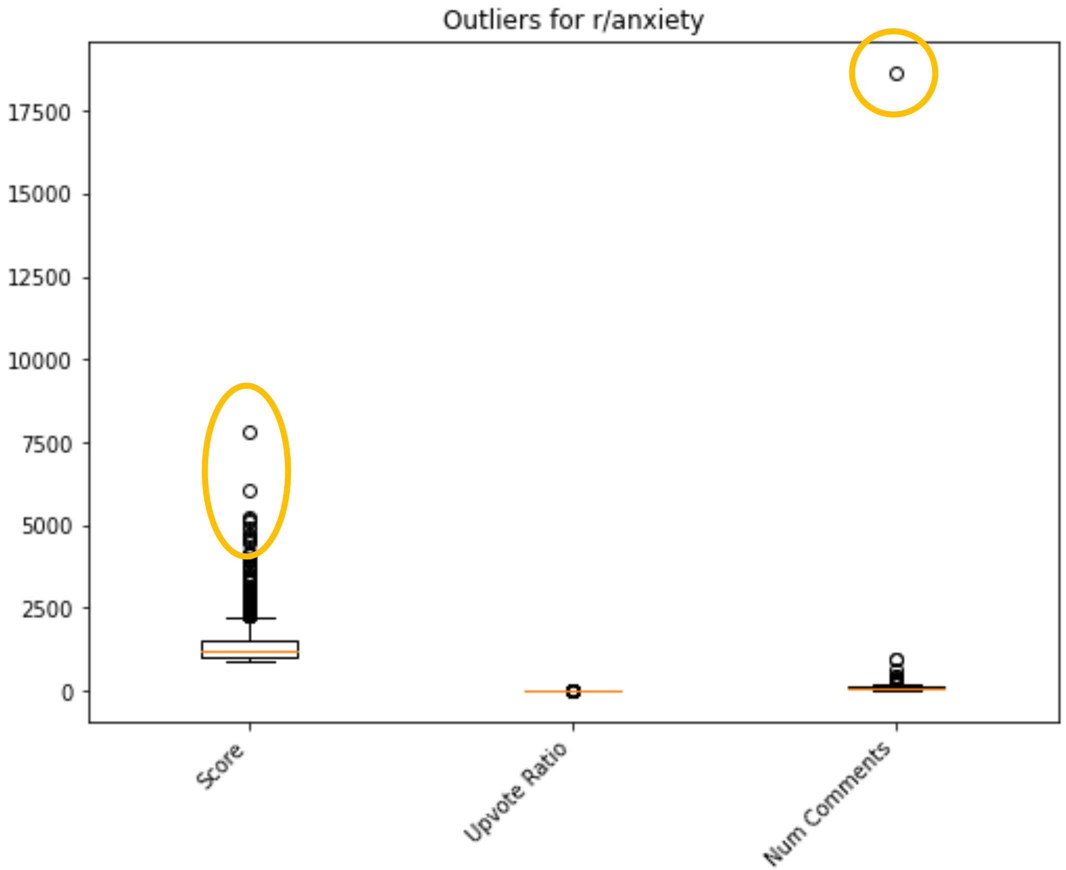
Data Cleaning & Exploratory Data Analysis

Pre-processing and Modelling

Conclusion & Recommendation

Outliers


Outliers for r/depression
7000
6000
5000
4000
3000
2000
1000
0
Score
Upvote Ratio
Num Comments


Outliers for r/anxiety
17500
15000
12500
10000
7500
5000
2500
0
Score
Upvote Ratio
Num Comments

Outliers (Score)

r/depression

Title: Shout out to the particular hell that is functional depression.

Content: This is me. Don’t get me wrong, it’s better than don’t-leave-my-bed-for-a-week depression. I am grateful I can be an independent person. But there is something uniquely horrible about being able to go to work every day, occasionally clean up after yourself, pay your bills, generally put yourself together enough to look like a human being... but that’s it. Nothing else. No social life. No hobbies. Constantly battling your mind. And being absolutely fucking exhausted all the time.

r/anxiety

Title: Despite the anxiety, despite the depression, despite all my self criticism and imperfections - I was a beautiful bride this Saturday!

Outliers (Num Comments)

r/depression

Title: Regular Check-In Post

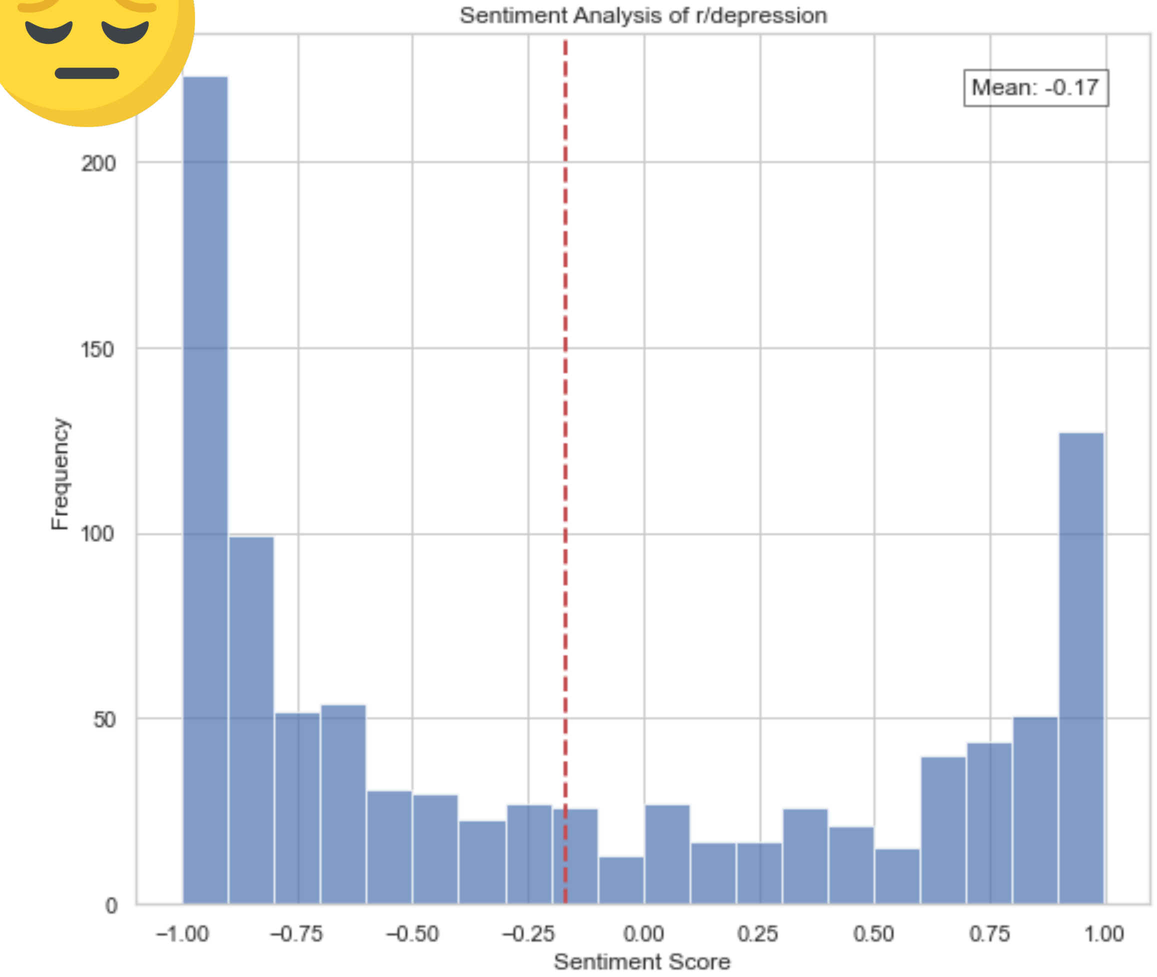
Content: Welcome to /r/depression's check-in post - a place to take a moment and share what is going on and how you are doing. If you have an accomplishment you want to talk about (these shouldn't be standalone posts in the sub as they violate the "role model" rule, but are welcome here), or are having a tough time but prefer not to make your own post, this is a place you can share. We try our best to keep this space as safe and supportive as possible on reddit's wide-open anonymity-friendly platform …

r/anxiety

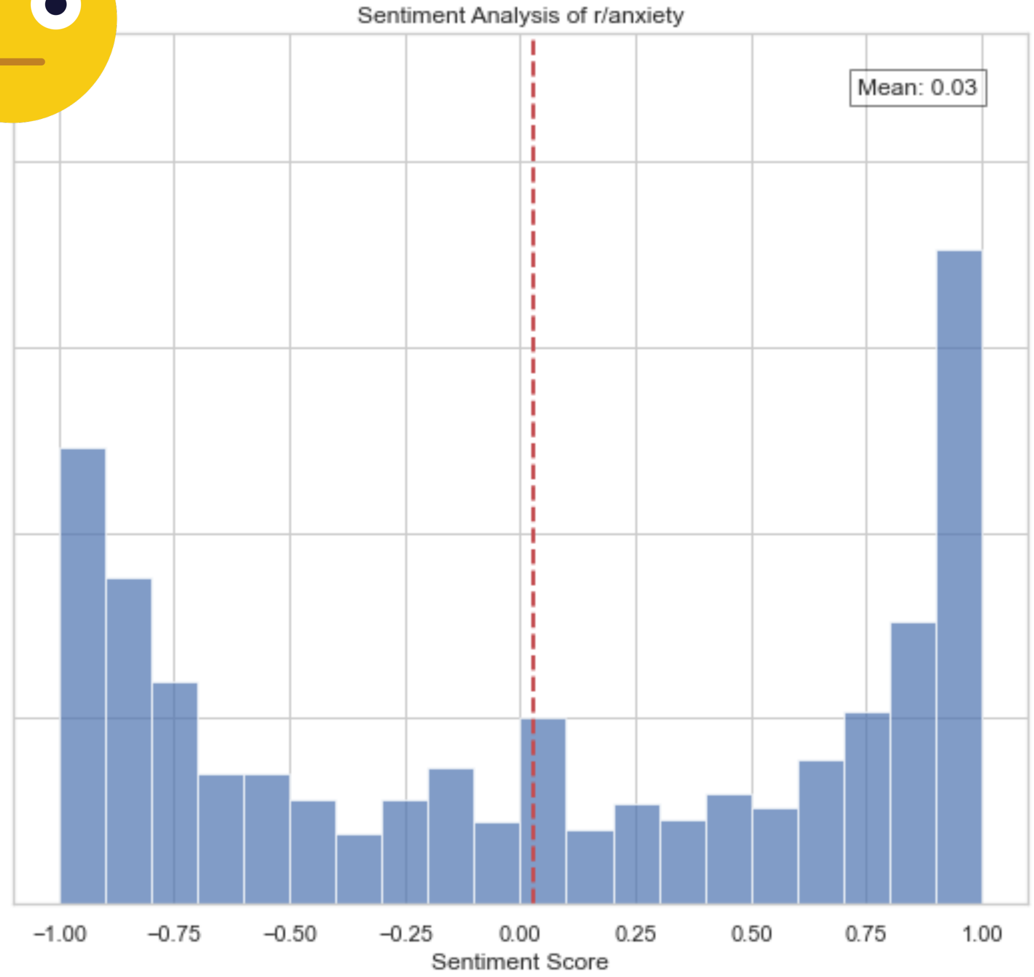
Title: Let's post good news on the coronavirus here.

Content: A place where only good news is posted, please keep this a positive thread. a place we can go for some reassurance that everything will be okay. We WILL get through this. edit: the link for this thread will be posted in the main thread, I will keep updating so save this thread to keep checking ❤️ …

Sentiment Analysis

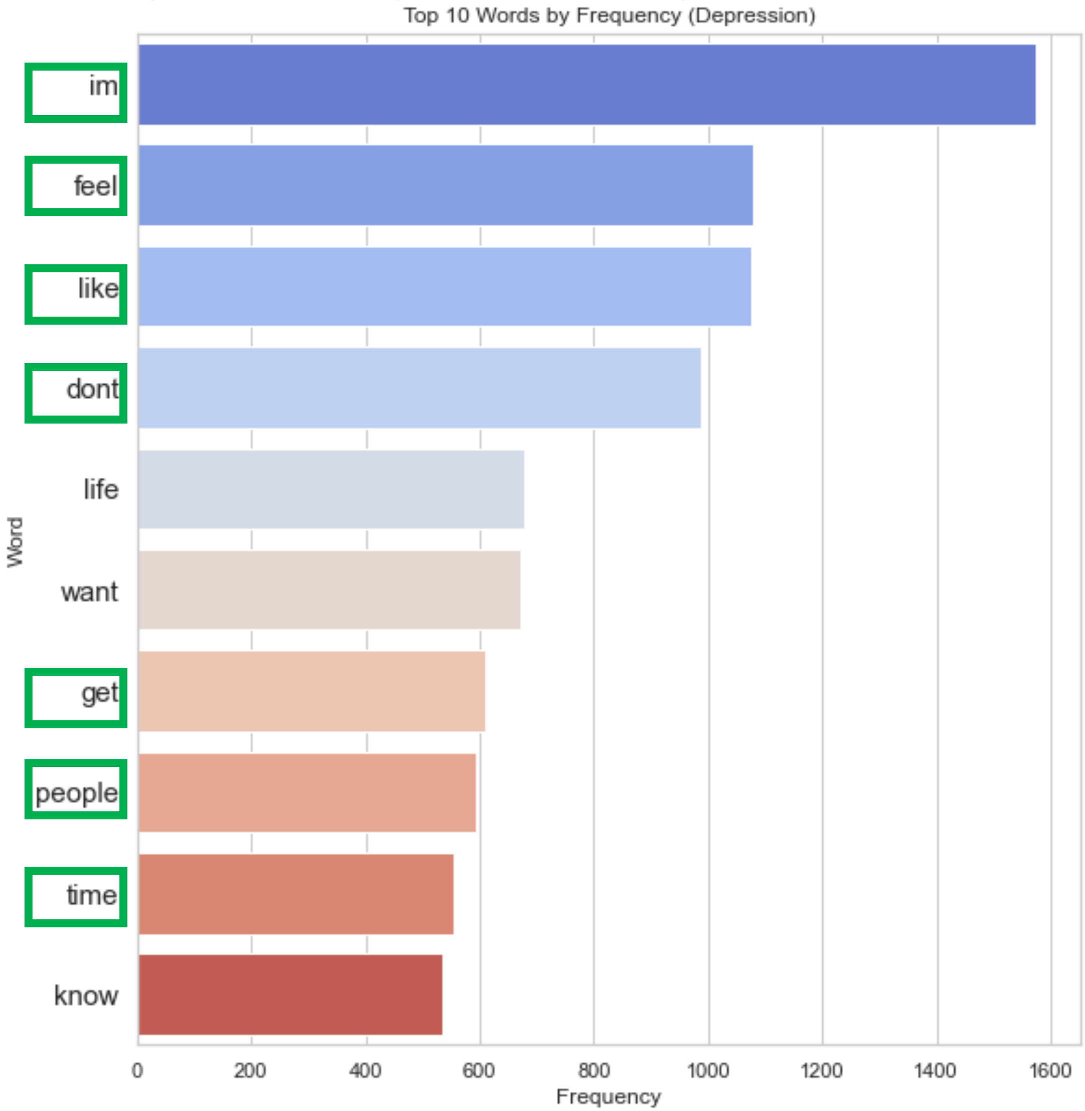


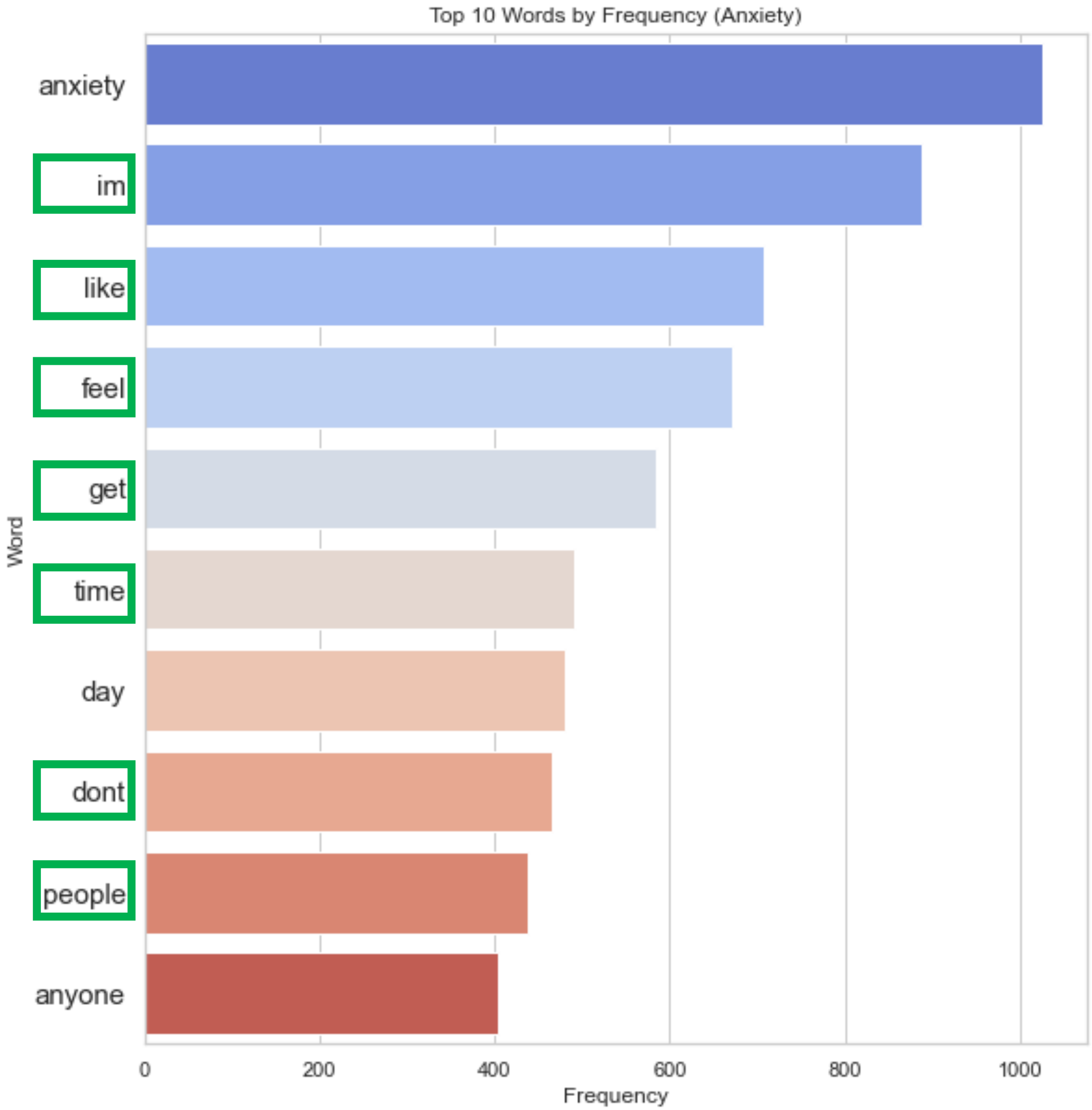
Sentiment Analysis of r/depression



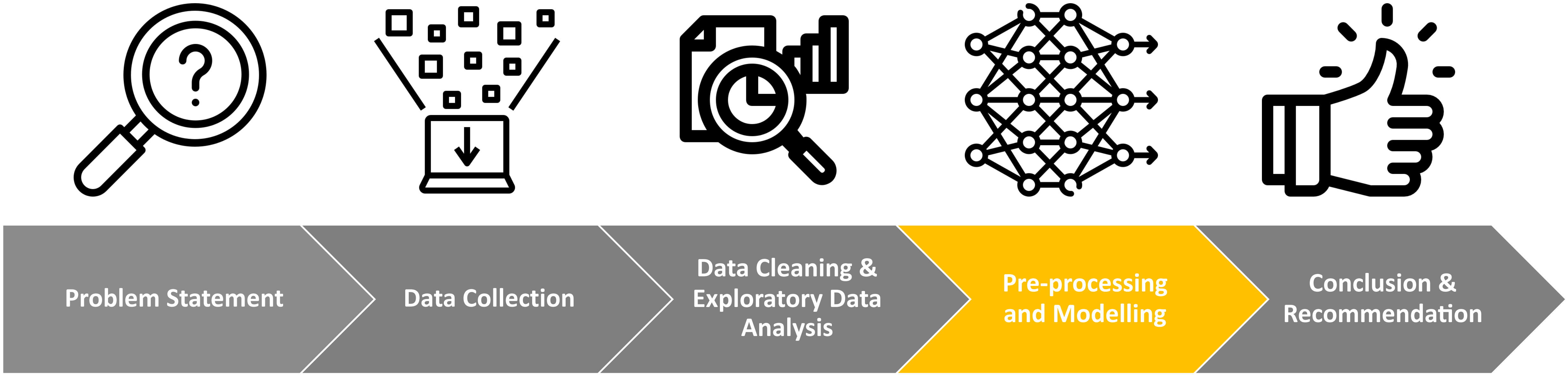
Sentiment Analysis of r/anxiety

Frequency of Top Words





Data Science Approach



Problem Statement

Data Collection

Data Cleaning & Exploratory Data Analysis

Pre-processing and Modelling

Conclusion & Recommendation

Model Training and Comparison

Models Evaluated:

- Naïve Bayes Model
- Logistic Regression

Performance Benchmark

(Predicts majority class in dataset)

50.39%

Score Comparison – Which to prioritise?

Precision $TP/(TP + FP)$

Ratio of true positives to the total of the true positives and false positives.

Recall $TP/(TP + FN)$

Ratio of true positives to the total of the true positives and false negatives.

F1 Score ✓

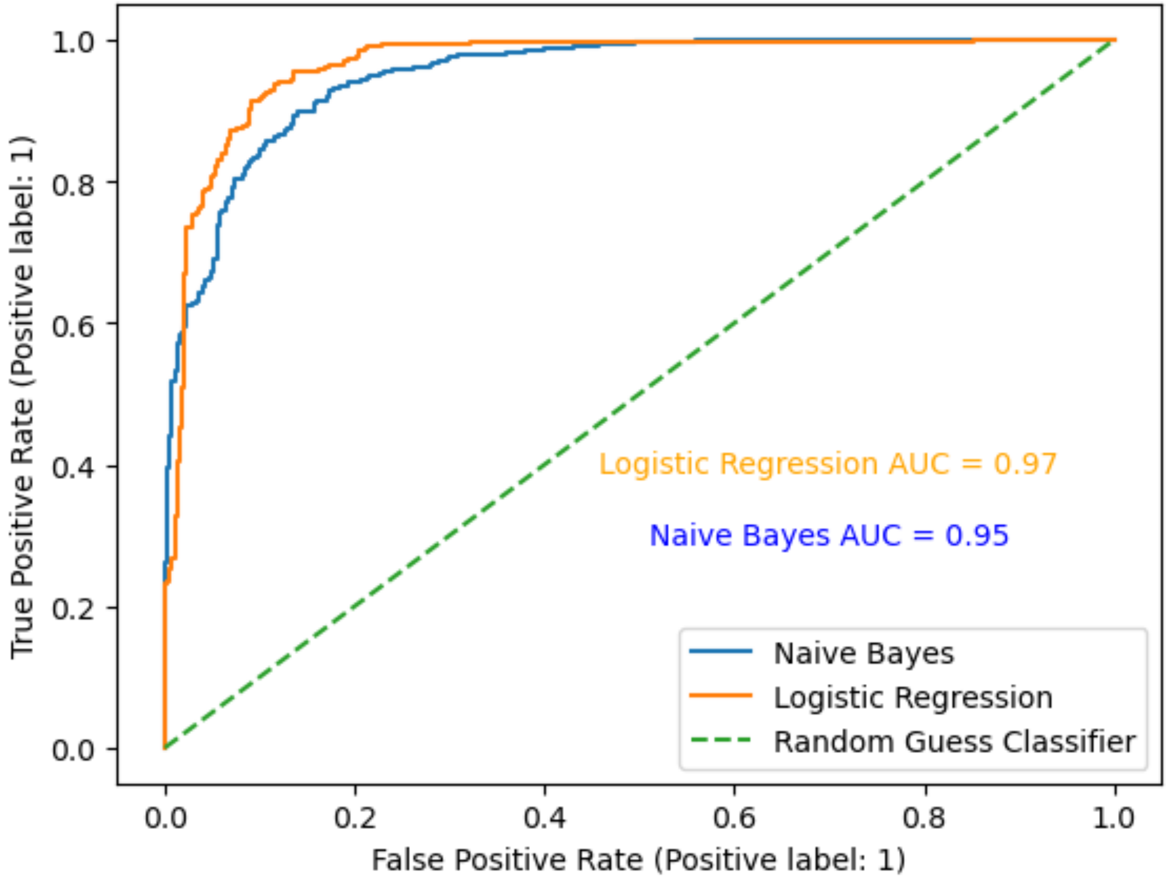
Balance of precision and recall.

Score Comparison

In the context of mental health classification between r/depression and r/anxiety with the goal of helping individuals receive appropriate help and support from the community, both **recall (sensitivity)** and **precision** are equally important.

Metrics*	Naïve Bayes	Logistic Regression
Mean Cross-validation score	0.8565560821484992	0.878041074249605
Accuracy	0.8775252525252525	0.9090909090909091
Precision	0.8416289592760181	0.9077306733167082
Recall	0.9323308270676691	0.9122807017543859
F1 Score	0.8846611177170036	0.9099999999999999

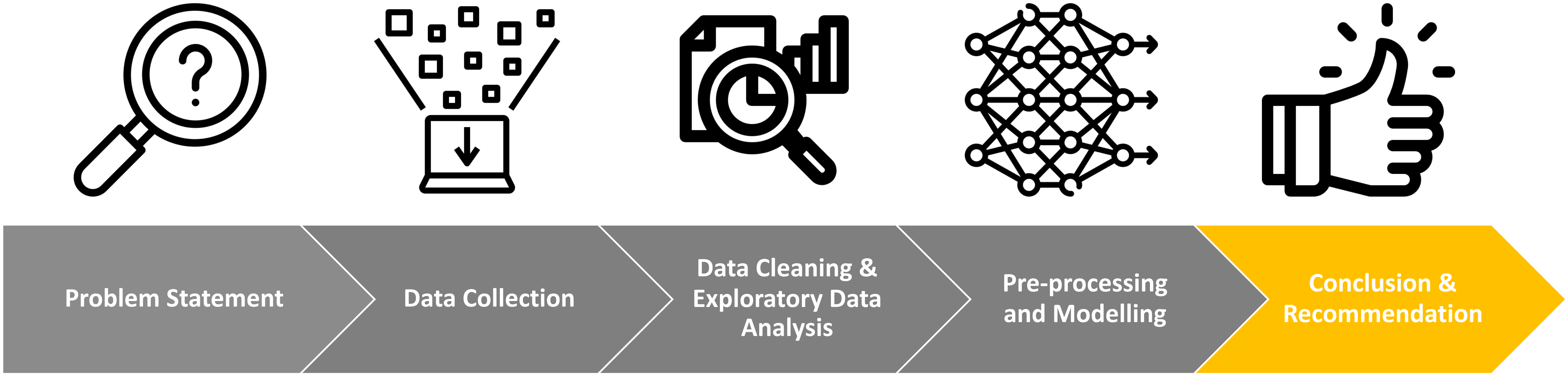
Receiving Operator Characteristic (ROC) Curve



Logistic Regression

90.9%

Data Science Approach



Problem Statement

Data Collection

Data Cleaning & Exploratory Data Analysis

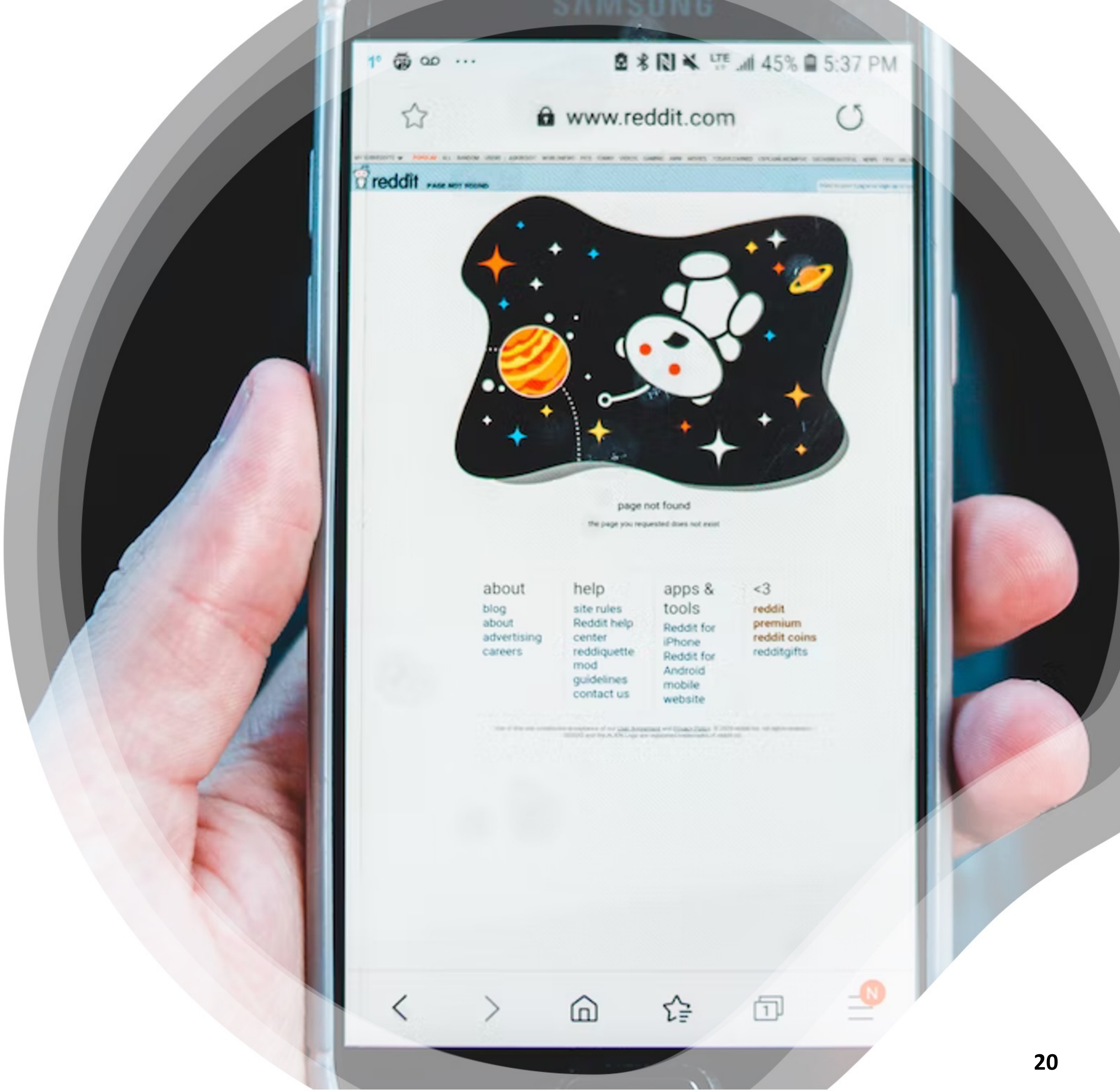
Pre-processing and Modelling

Conclusion & Recommendation

Conclusion and Recommendation

- Moderators of both subreddits can utilise the trained model to have a preliminary classification of users' posts.
- However, there is a need to scrutinise the content to have a second evaluation.

Thank You


www.reddit.com
reddit
page not found
about
blog
about
advertising
careers
help
site rules
Reddit help center
reddiquette
mod guidelines
contact us
apps & tools
Reddit for iPhone
Reddit for Android
mobile website
<3
reddit premium
reddit coins
redditgifts
45%
5:37 PM

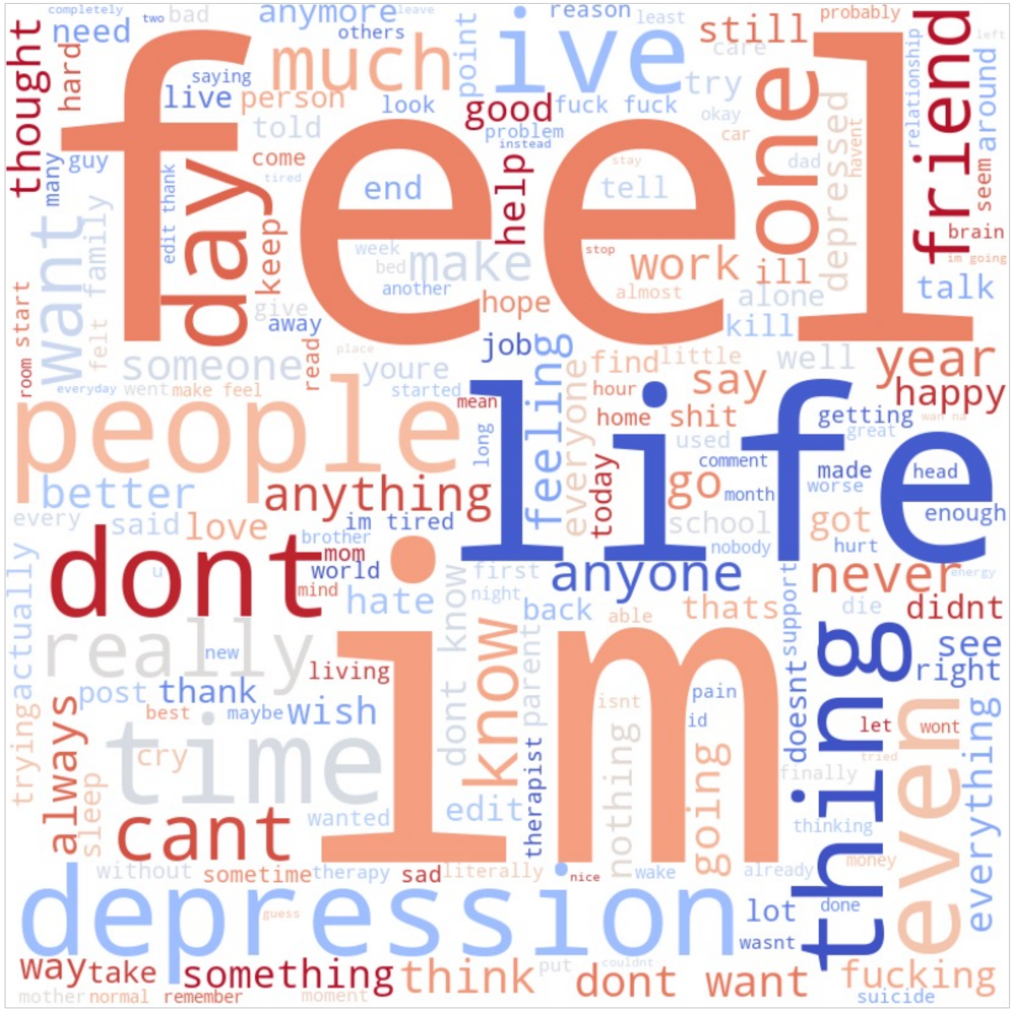
Annex

Data Collection

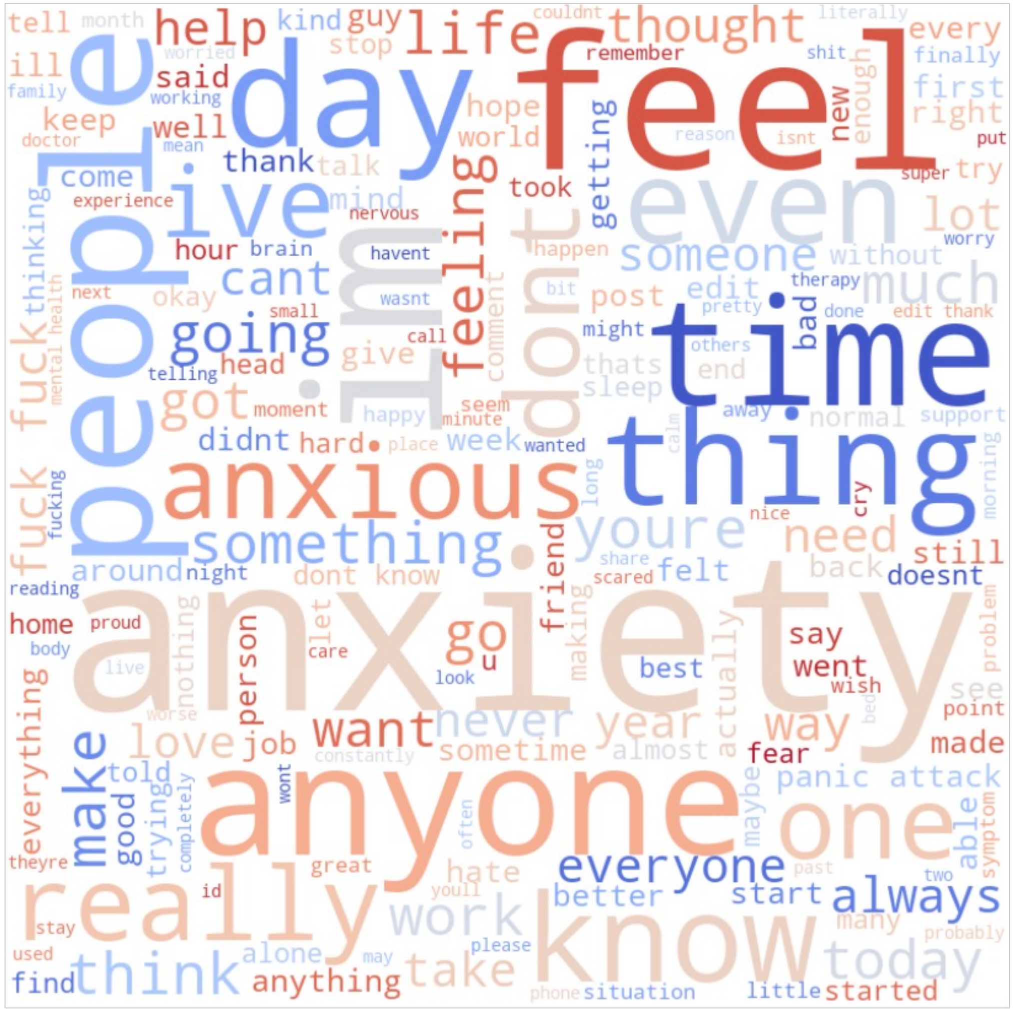
Feature	Description
id	Unique identifier of a particular post or comment within a subreddit
created_utc	The time the post or comment was created
title	The title of the post.
is_self	Indicates whether the post is a self-post or not.
selftext	The actual text content of a self-post
score	The upvotes minus the downvotes of the post.
upvote_ratio	The ratio of upvotes to total votes.
num_comments	The total number of comments on the post.
permalink	The permanent link to the post.
author	The username of the person who submitted the post.
distinguished	Indicates whether the post or comment has been distinguished by a moderator or admin.

Word Cloud

r/depression



r/anxiety



Misclassified Posts

Post 1:

Title: What do you do when you can’t focus on anything? Not even a shitty program or music. I don’t have adhd, just so you understand.

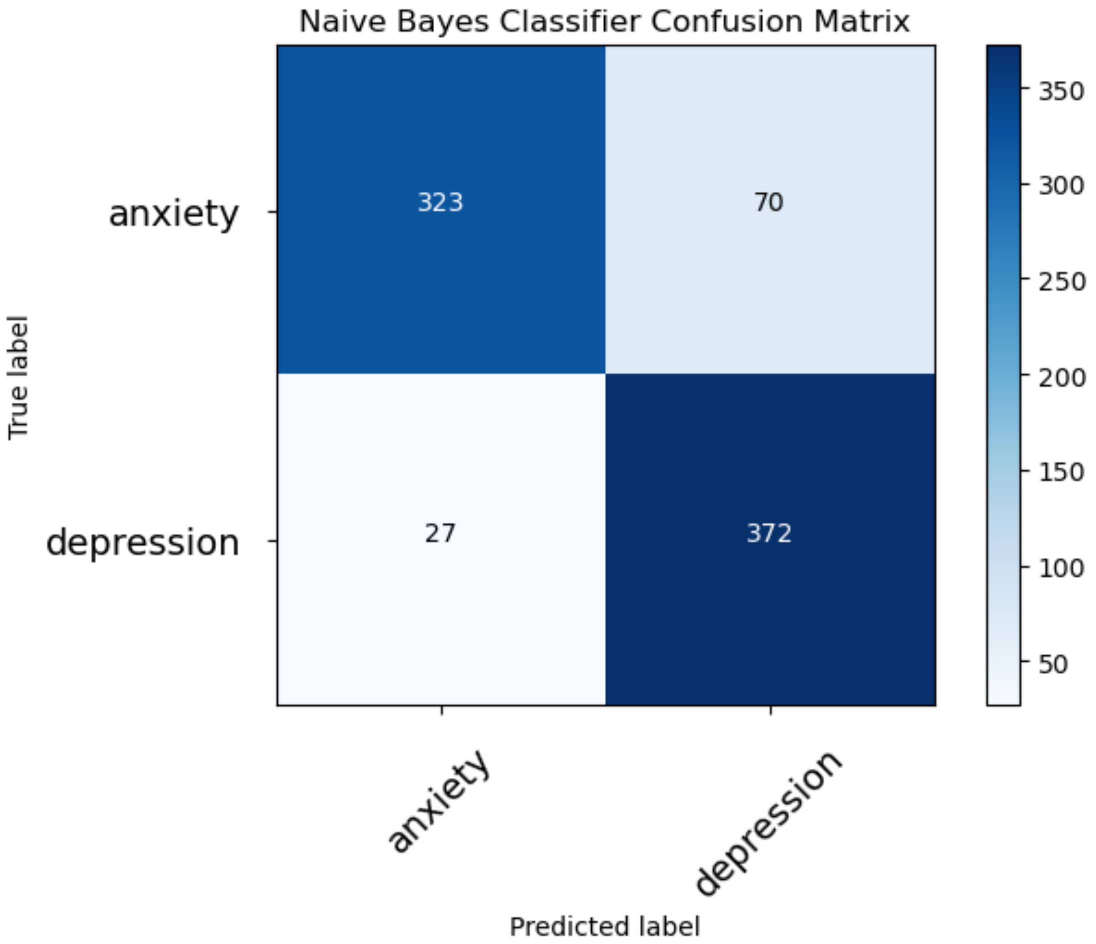
Post 2:

Title: Sorry... I couldn't help myself. I haven't felt this happy since I was a kid.

Post 3:

Title: Feeling unattractive again.
Selftext: Being a bbw I constantly struggle to feel beautiful. I honestly hate my body…

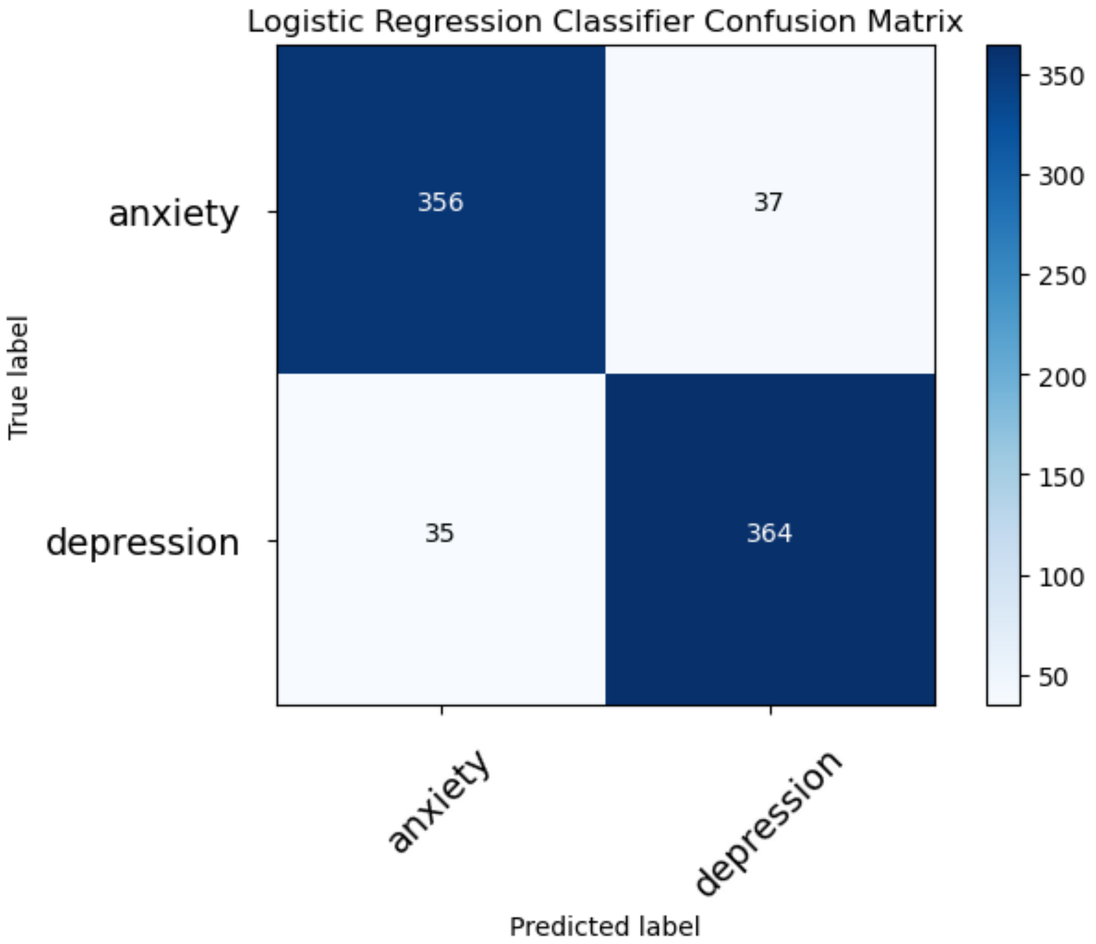
Naïve Bayes Model Confusion Matrix



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Logistic Regression Model Confusion Matrix



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Feature Importance

Top 10 Features	Importance
anxiety	-10.2146
depression	6.3353
removed	5.0625
anxious	-4.6802
depressed	4.1259
life	3.7131
panic	-2.7902
attack	-2.7340
kill	2.4345
tired	2.2143