NEW FEATURE IMPLEMENTATION

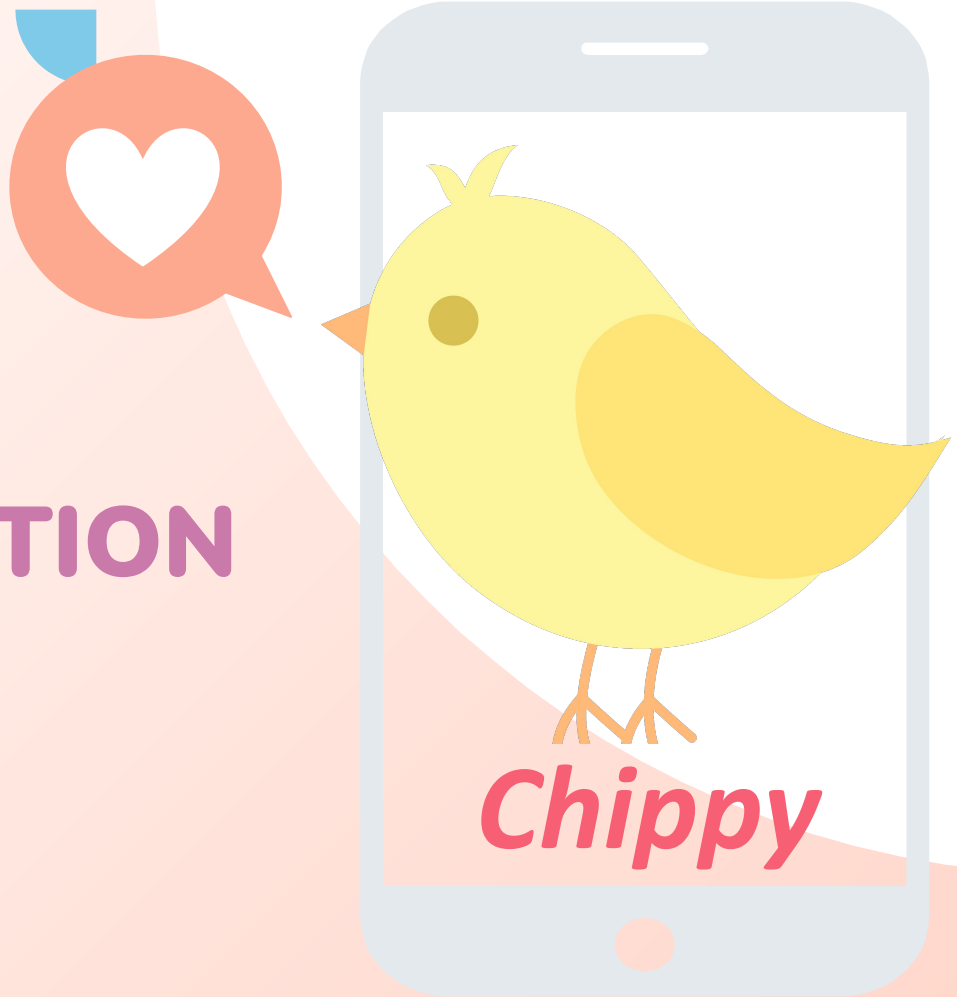


TABLE OF CONTENTS

01 BACKGROUND

02 PROBLEM STATEMENT

03 WORKFLOW

04 DATA CLEANING & EDA

05 MODEL & EVALUATION

06 CONCLUSION & RECOMMENDATIONS

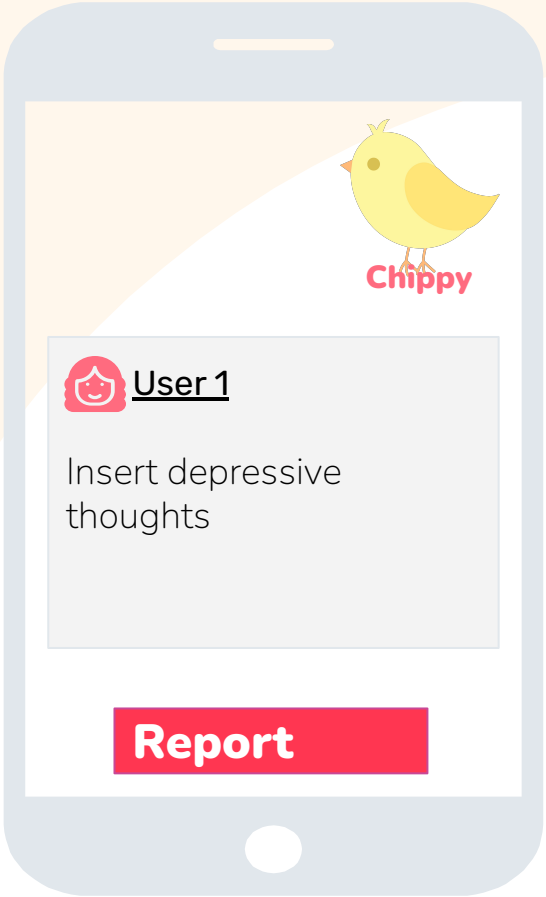
01
Corporate Social Responsibility

According to JAMA Psychiatry,
studies show that ***adolescents*** who **use social media**
> three hours /day may be at ***heightened risk*** *of*
mental health problems

According to Bachmann S. Epidemiology of Suicide and the Psychiatric Perspective,
most suicides are related **to**
psychiatric diseases and depression

REPORT FEATURE

- Enables users to flag other users if they suspect a user to be suffering from mental health issues based on their posts



Benefits of Corporate Social Responsibility

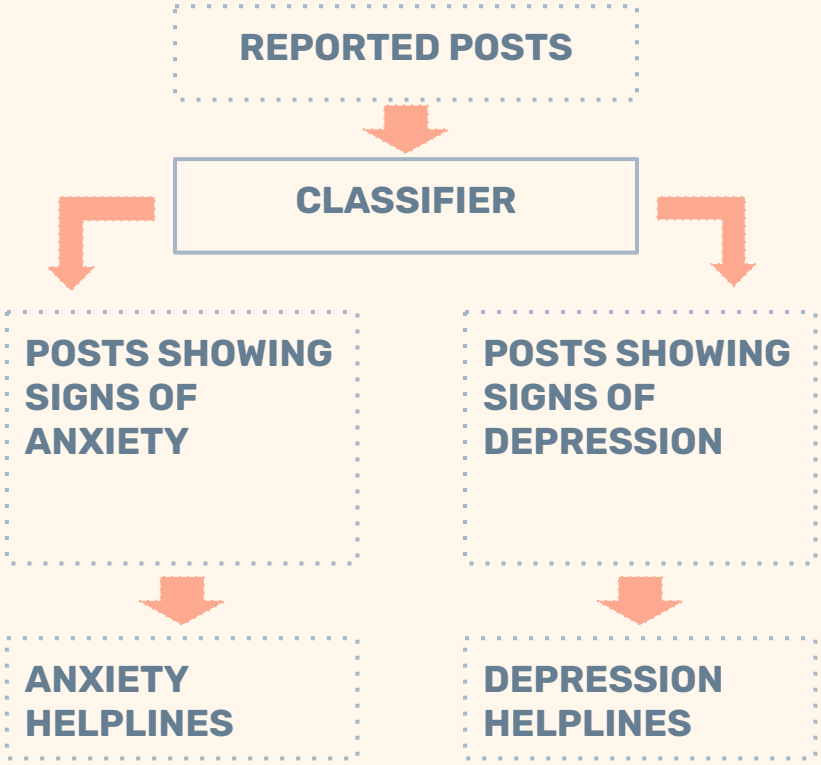
BETTER BRAND RECOGNITION

POSITIVE BUSINESS REPUTATION

GAIN BUSINESS ADVANTAGE

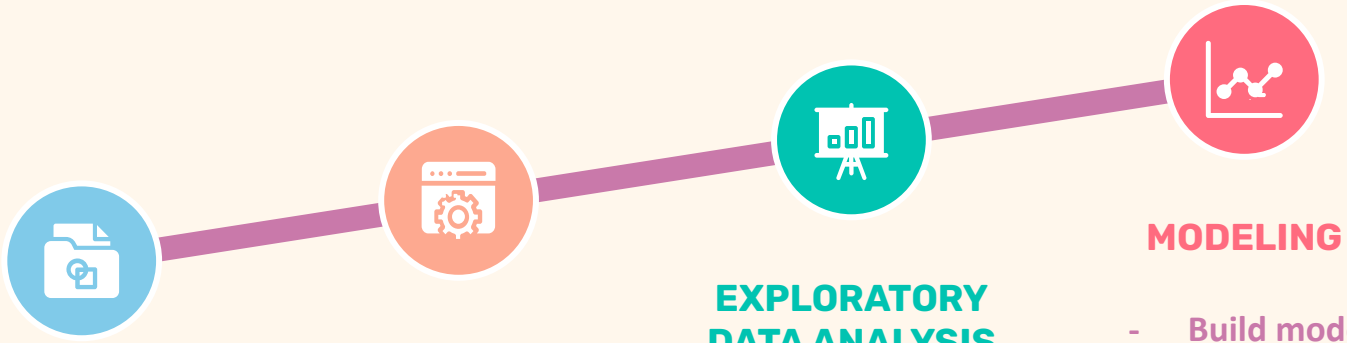
02
Problem Statement

Our Task


REPORTED POSTS
CLASSIFIER
POSTS SHOWING SIGNS OF ANXIETY
POSTS SHOWING SIGNS OF DEPRESSION
ANXIETY HELPLINES
DEPRESSION HELPLINES

02
WORKFLOW

WORKFLOW



WEB-SCRAPING

- r/depression
- r/Anxiety

DATA CLEANING

- Clean text
- Tokenize text

EXPLORATORY DATA ANALYSIS

- Unearth keywords

MODELING

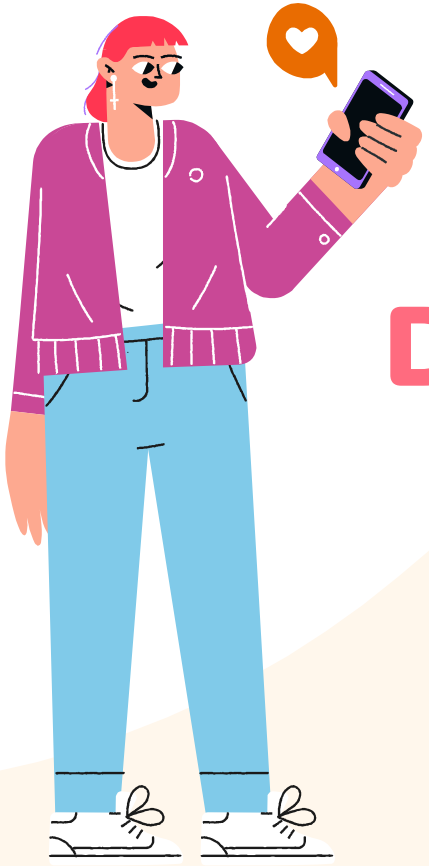
- Build model
- Predict topic in new text

USING PROXY DATA FROM REDDIT

r/Depression

r/Anxiety

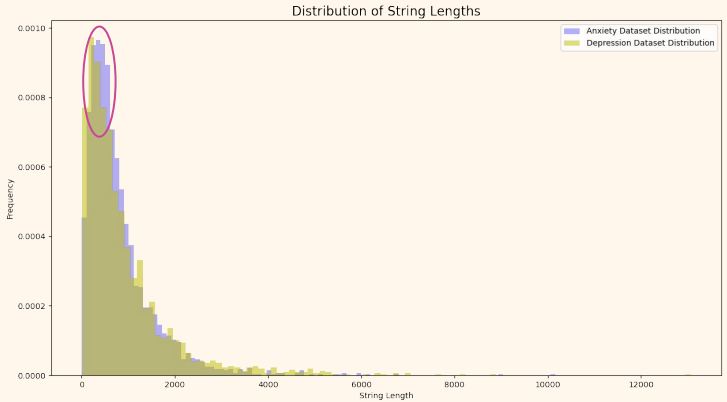
Tota **6000** posts scraped from
both subreddits

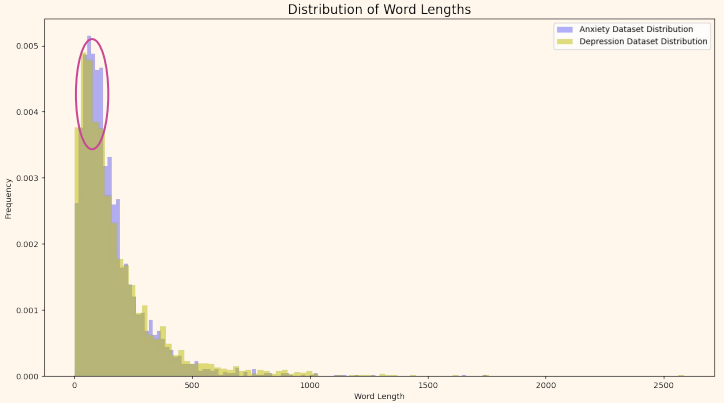


03
DATA CLEANING & EDA

EDA : SIMILAR DISTRIBUTION IN BOTH DATASETS

- REQUIRES NLP FOR IN DEPTH ANALYSIS





What is NLP?

Natural Language Processing

Machine learning technology to interpret, manipulate and comprehend human language

Use Cases

- Spam Filter in Email Account
- Autocomplete in Google Search
- Autocorrect in Word Document

Applications

Text Classification

'What topic does a text instance belong to?'

Preprocessing text data

Remove Punctuations

.,?!

Tokenization

Break up long strings & sentences into lists of words

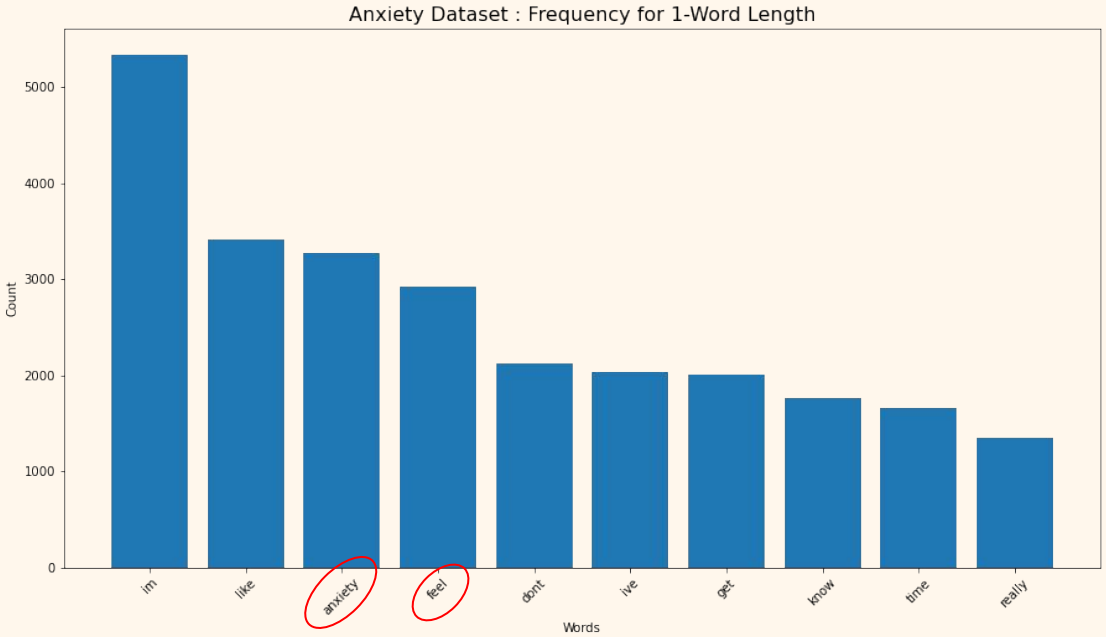
Remove Stopwords

‘but’, ‘if’, ‘the’

Lemmatization

Reduce words to base form
e.g. bats → bat,
feet → foot, better → good

Top 10 words: Anxiety dataset



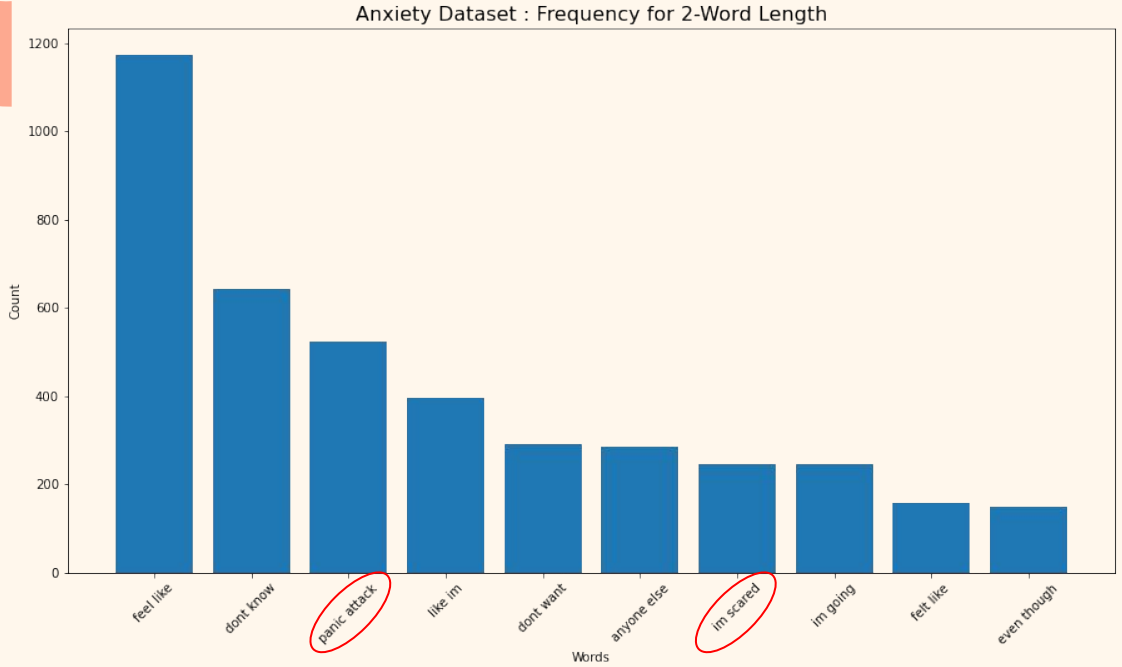
1.2

‘anxiety’ per post

1.1

‘feel’ per post

Top 10 bigrams: Anxiety dataset



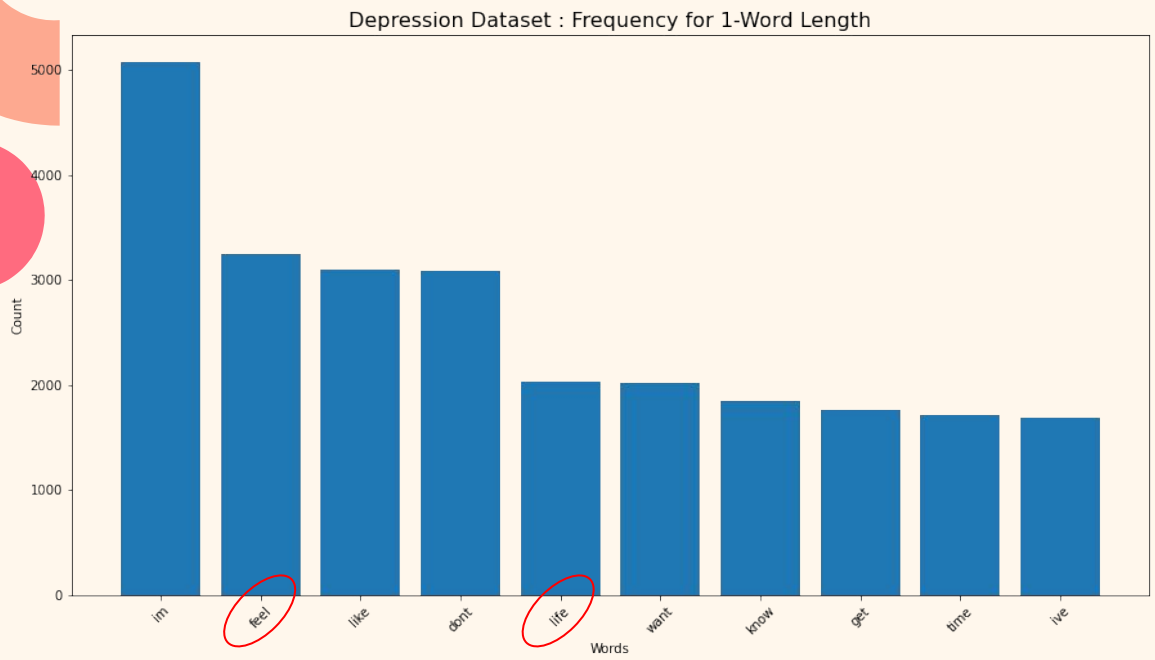
20%

of posts contains *'panic attack'*

10%

of posts contains *'im scared'*

Top 10 words: Depression dataset

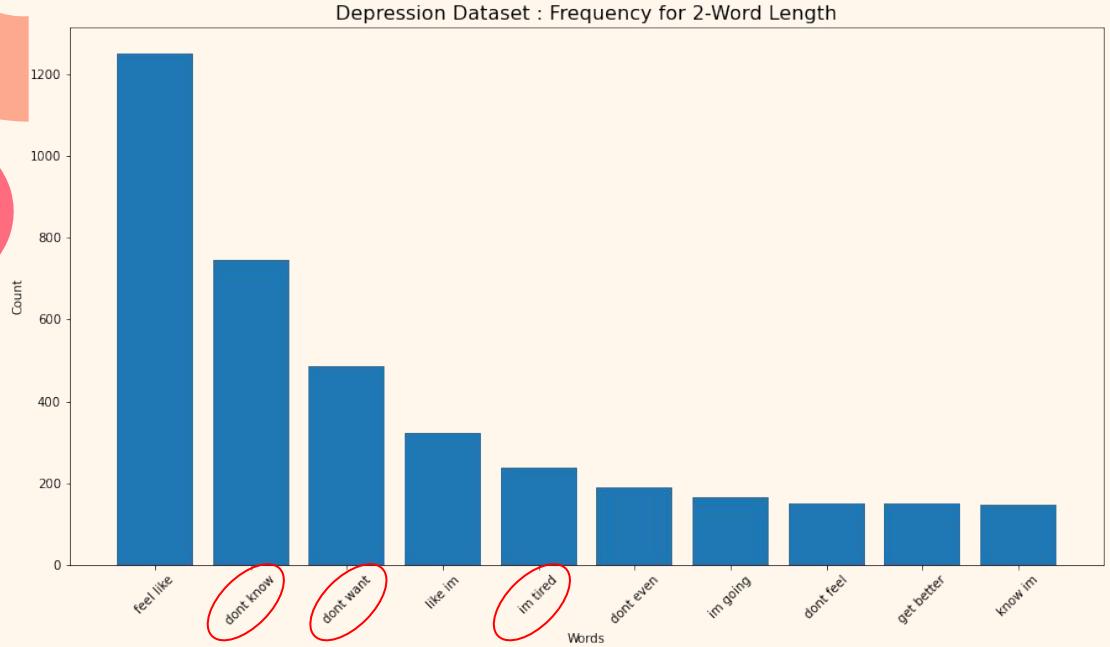


1.4
'feel' per post

0.9
'life' per post

Not in top 10:
'depressed'

Top 10 bigrams: Depression dataset



50%

of posts contain *'don't know' & 'don't want'*

10%

of posts contain *'i'm tired'*

04

MODEL & EVALUATION

Data Preparation for Modelling

N-Gram (1,2)	TF-IDF
Remove punctuation	
Remove words with numbers	
Remove stopwords	
Lemmatize	
CountVectorizer (max_features = 10,000)	TFIDFVectorizer (max_features = 10,000)

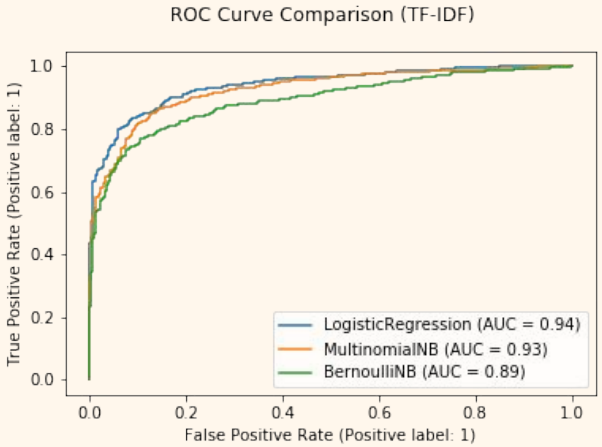
Logistic Regression

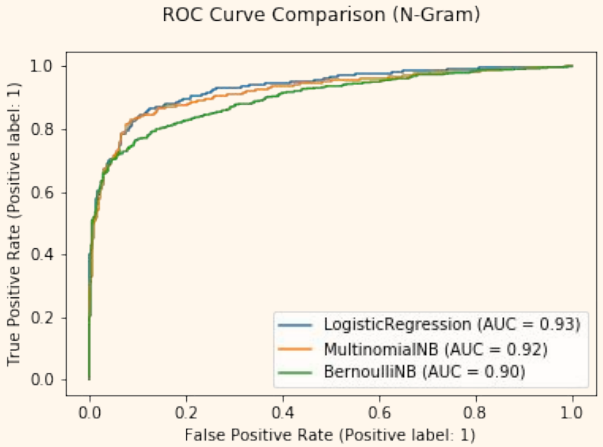
Multinomial Naive Bayes

Bernoulli Naive Bayes

Findings : ROC Curve

- TF-IDF vectorization performs better than Count vectorization
- Logistic regression perform better than Naive Bayes models





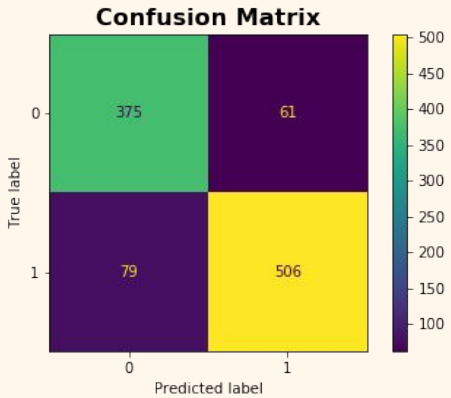
Model Results

- Logistic Regression with TF-IDF vectorization selected for production
- Optimized for accuracy

Model	Method	Train Score	Accuracy	Precision (Depression)	Precision (Anxiety)	AUC
Logistic Regression	TF-IDF	0.94	0.86	0.83	0.89	0.94
Multinomial Naive Bayes	TF-IDF	0.91	0.86	0.82	0.89	0.93
Bernoulli Naive Bayes	TF-IDF	0.84	0.80	0.81	0.80	0.89
Logistic Regression	N-Gram (1,2)	0.99	0.86	0.81	0.91	0.93
Multinomial Naive Bayes	N-Gram (1,2)	0.90	0.86	0.81	0.90	0.92
Bernoulli Naive Bayes	N-Gram (1,2)	0.85	0.80	0.79	0.80	0.90

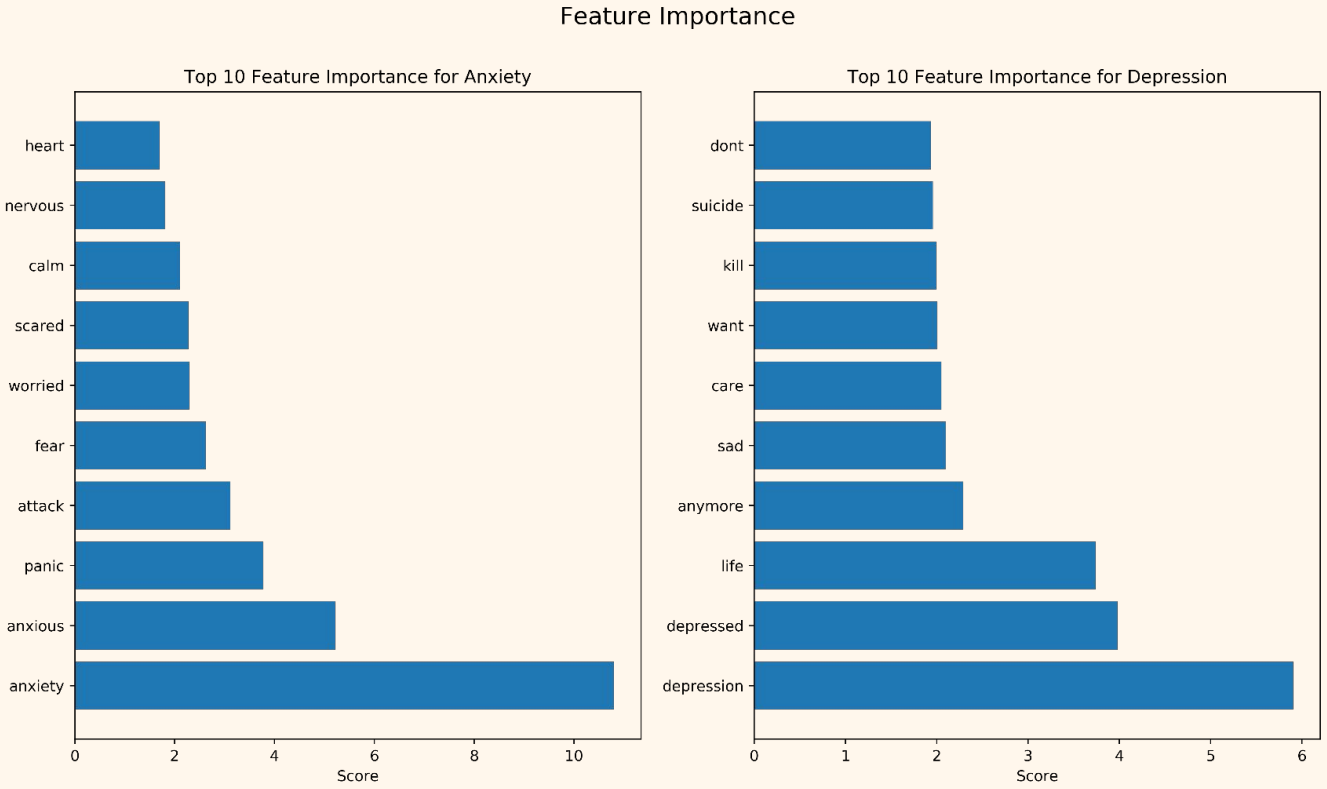
Logistic Regression Results

- Regression model is able to predict with 86% accuracy
- Train Score: 0.937
- Test Score: 0.863



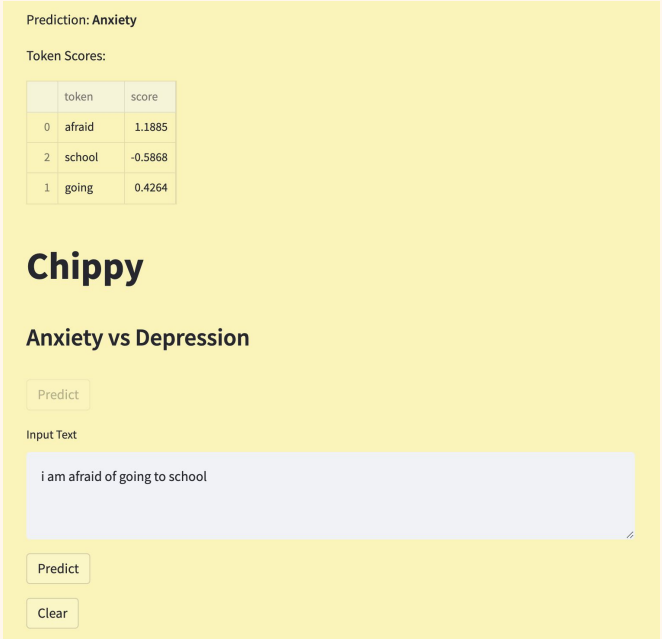
	precision	recall	f1-score	support
0 (depression)	0.83	0.86	0.84	436
1 (anxiety)	0.89	0.86	0.88	585
accuracy			0.86	1021
macro avg	0.86	0.86	0.86	1021
weighted avg	0.86	0.86	0.86	1021

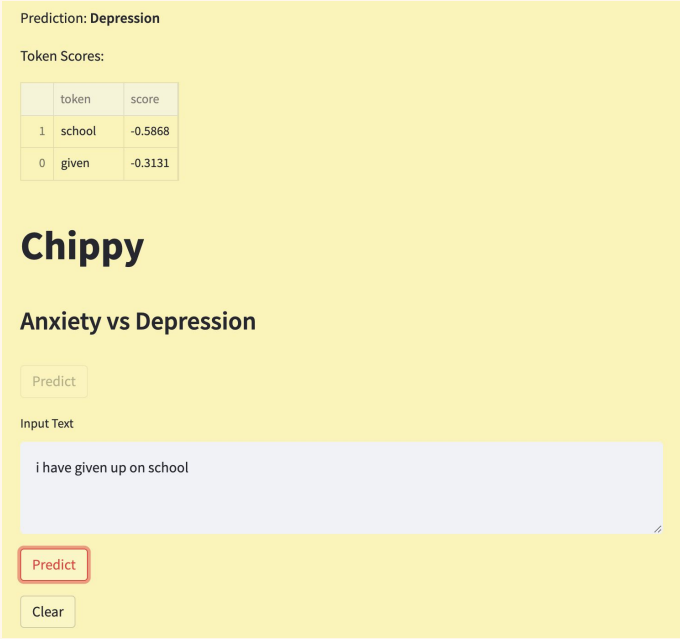
Top 10 Important Words That Classify Text



Production

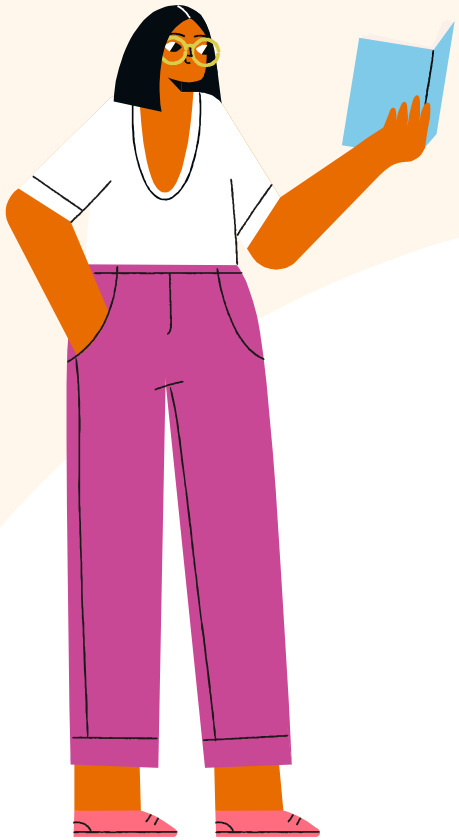
Our *classifier* uses logistic regression results to predict if users are likely to be suffering from Anxiety or Depression based on their posts





05

CONCLUSIONS

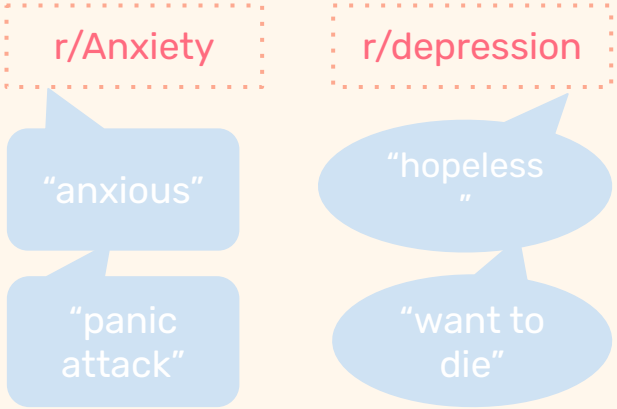


OUR RESULTS & FINDINGS

MISSION ACCOMPLISHED

Built a classifier model
>85% accurate

KEY INSIGHTS



LIMITATIONS

- **Co-morbidity**
  - Focus on predominant symptoms
- **Lack of internal data**
  - Reddit data as proxy

OUR RECOMMENDATIONS

EXPAND APPLICATION

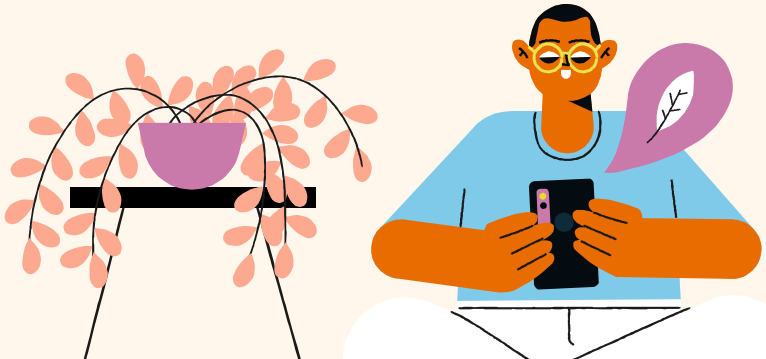
- Classify other pressing issues

DEVELOP FEATURE

- **Auto-flag** posts that contain experiences with poor mental health

COLLABORATE

- Local mental health helplines
- Provide support for caregivers



THANK YOU

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution