Reddit API and NLP Project

Created by Landry Houston

Project Objective

Develop a natural language processing model that minimizes both false positives and negatives to predict subreddit origin.

r/Anxiety

r/Depression

Subreddit API Web Scraping

Subreddit → Post → Dataframe → CSV

Task Scheduler

Scrapes subreddit automatically

↓

20 - 40 posts every hour

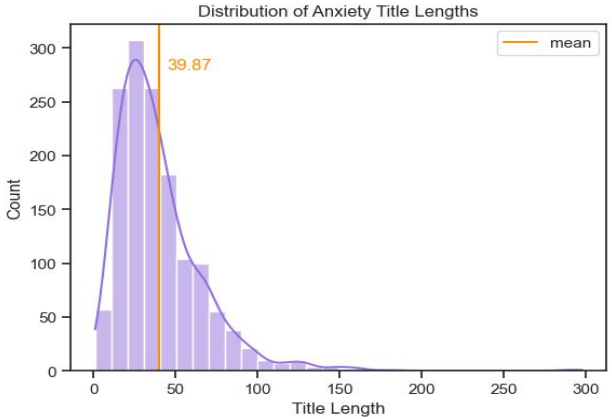
Data Cleaning

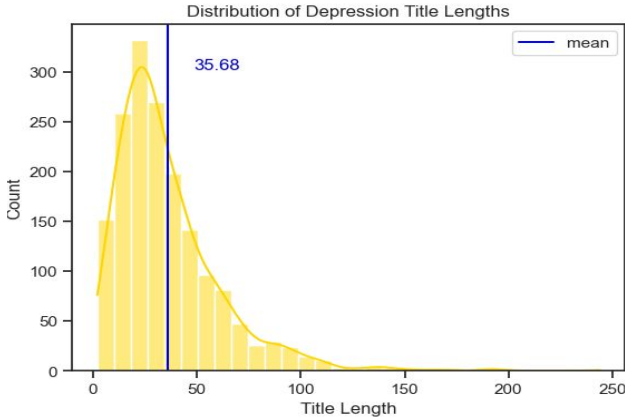
Data Dictionary

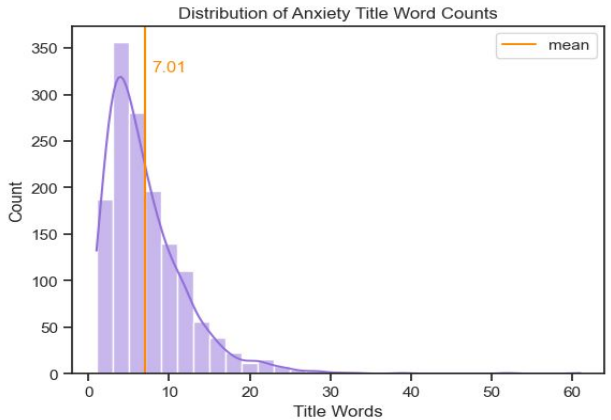
Feature	Type	Description
id	object	Post ID
subreddit	object	Reddit community
date	datetime	Date of post (yyyy-mm-dd)
title	object	Title of post
text	object	Text within a post

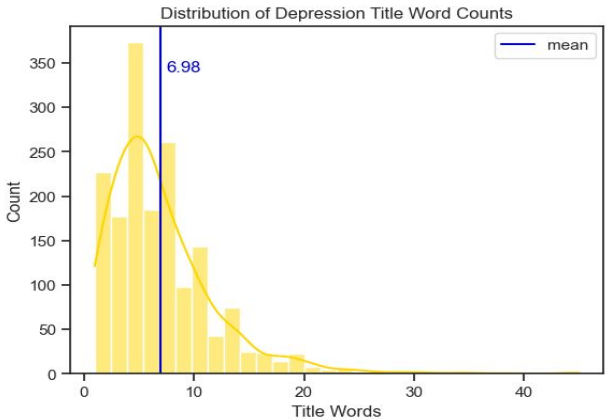
1. Remove rows with Null values
2. Convert UTC to datetime
3. Save as new CSV

Data Analysis

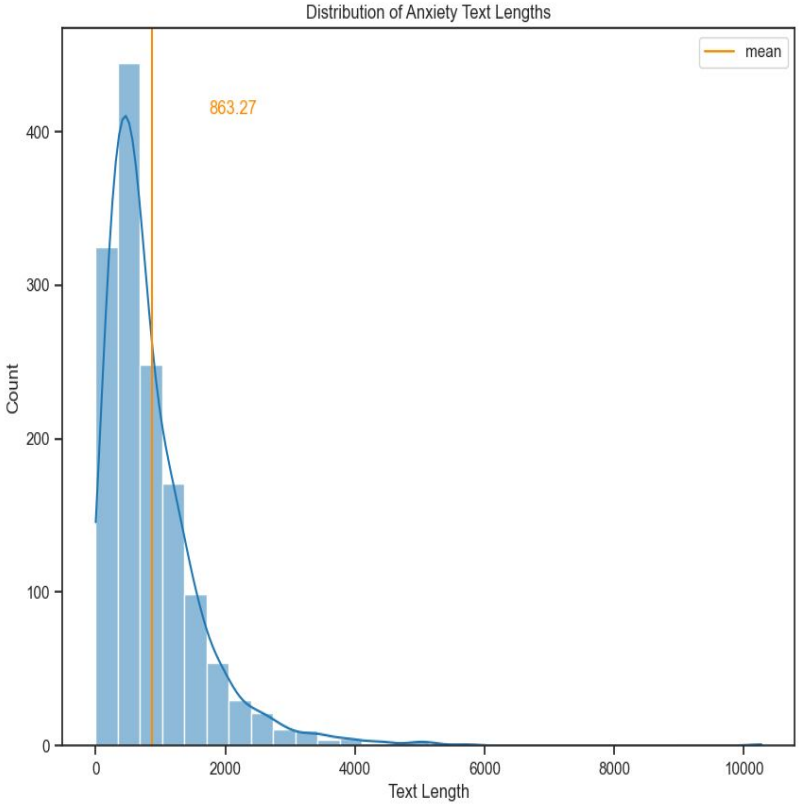


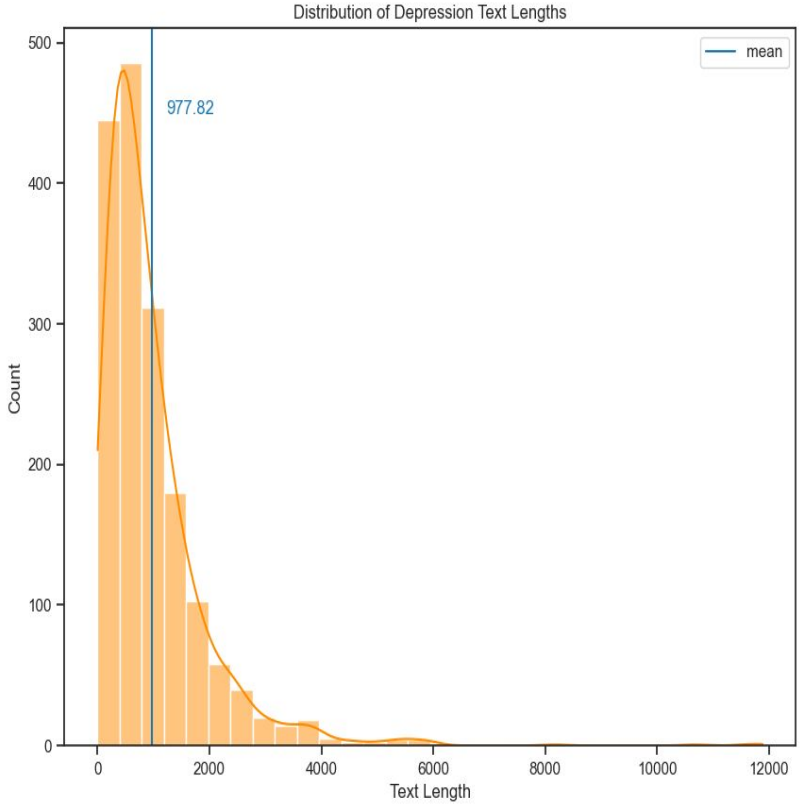




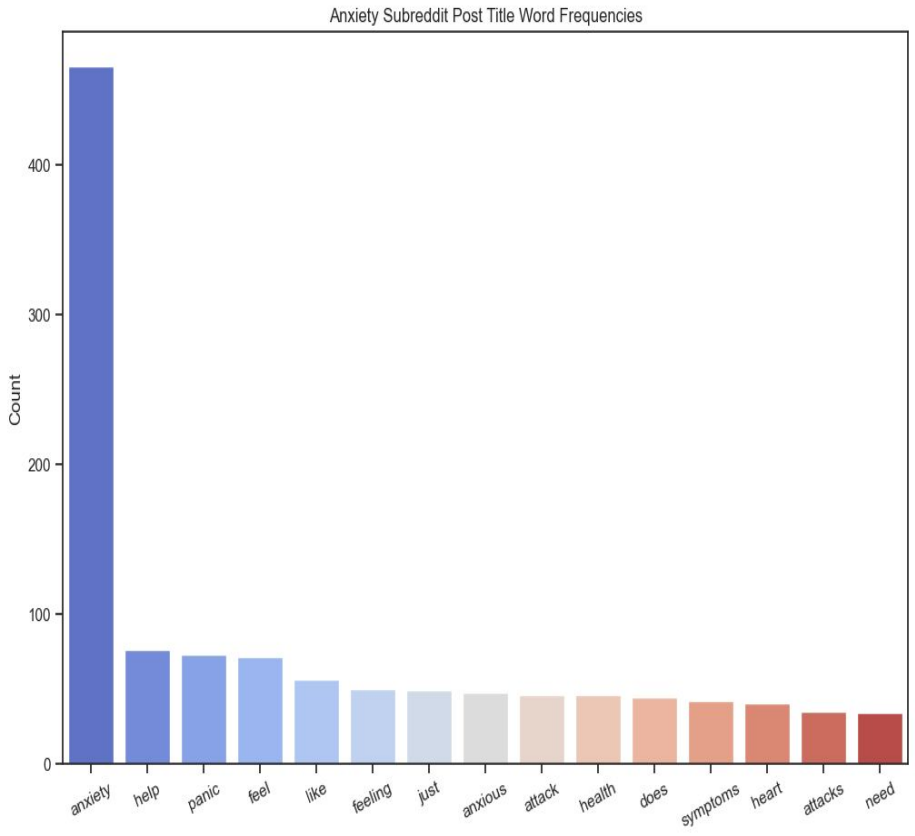


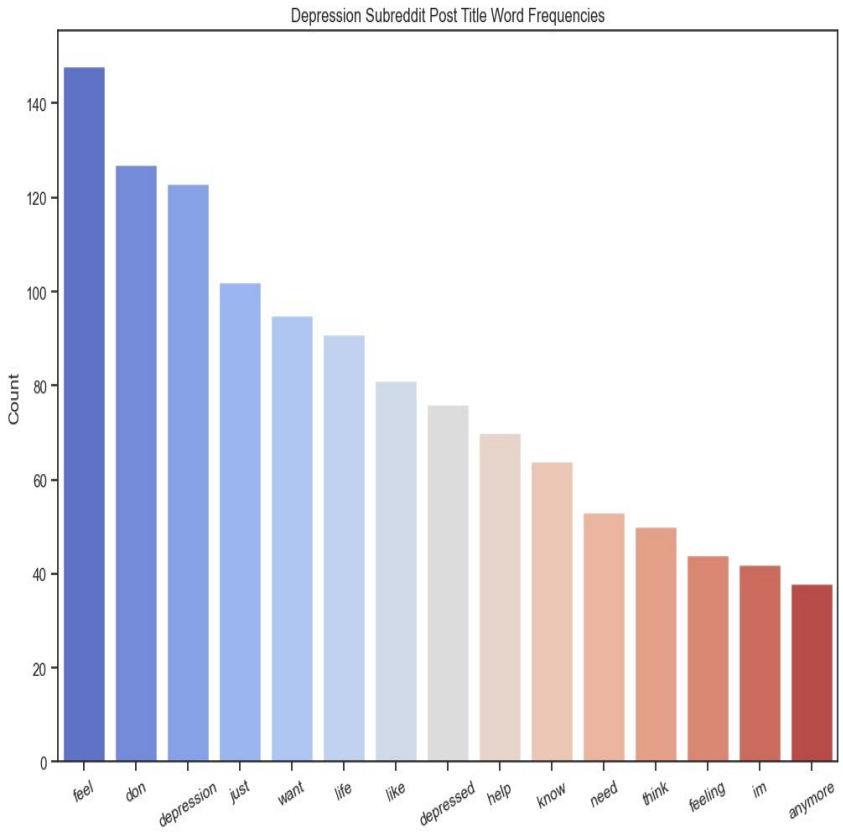
Data Analysis



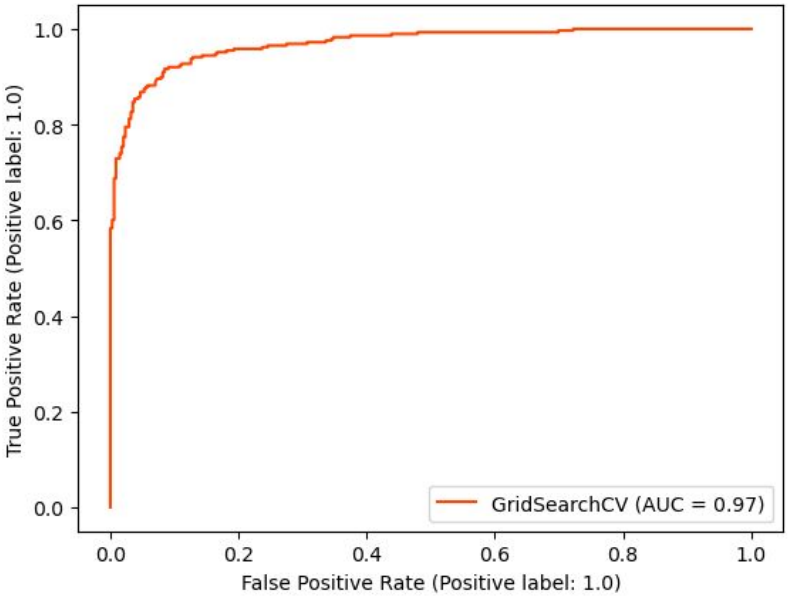


Data Analysis





Natural Language Processing Model



Best Model

TfidfVectorizer

LogisticRegression

Achieved > 90% Accuracy

I love a good model

Conclusion

- Successfully scraped Reddit API
- Scheduled script to run automatically
- Cleaned and analyzed data
- Created and compared NLP models
- Minimized false positives and negatives

Thank You!

Questions?