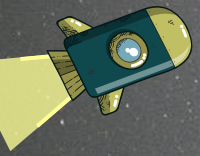
Classification Modelling and NLP of subreddits

Andrea Wong

TABLE OF CONTENTS

Objectives



1) Collect posts from two subreddits (r/aliens & r/conspiracytheories)

2) Use NLP to train a classification models to predict whether a post originated from a specific subreddit

3) The credibility of aliens' existence with conspiracy theories revolving around aliens

r/aliens

A community dedicated to discussion of the possibility of extraterrestrial life.

r/conspiracytheories

A community to discuss every aspect of conspiracy theorism, from theories and current events to debunkings and popular culture.

Data Aquisition

Aimed to collect 10,000 posts using Pushshift Reddit API

r/aliens

9996 posts

r/conspiracytheories

9999 posts

Cleaning and preprocessing data collected

Stopwords

Remove stopwords, Added additional stopwords

Links

Remove HTML links

Emojis

Convert emojis to text

Lowercase

Convert words to lowercase

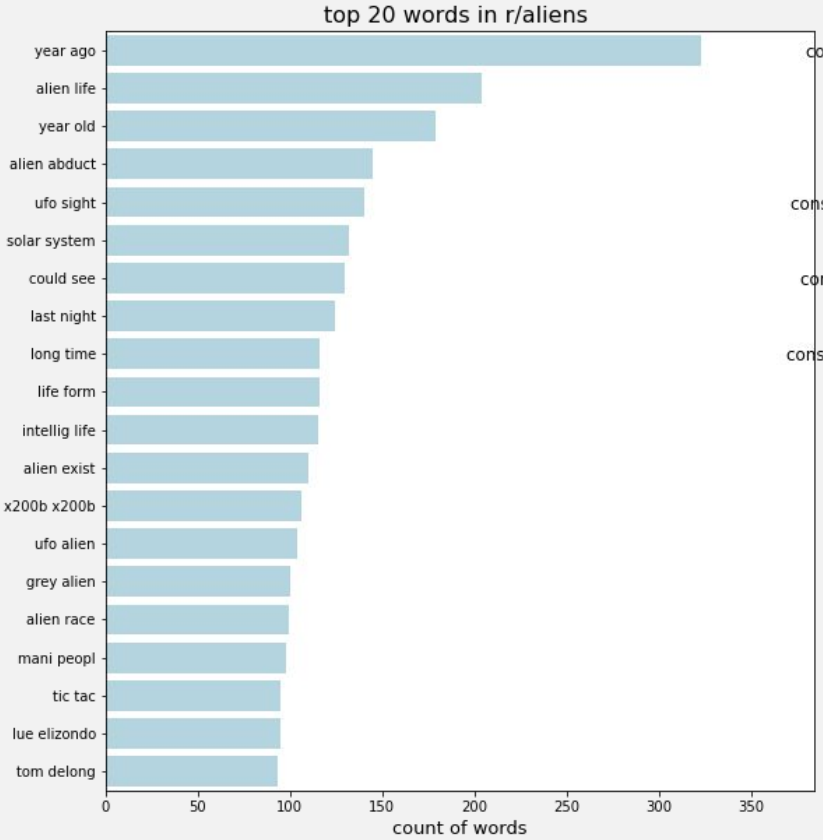
Stemming

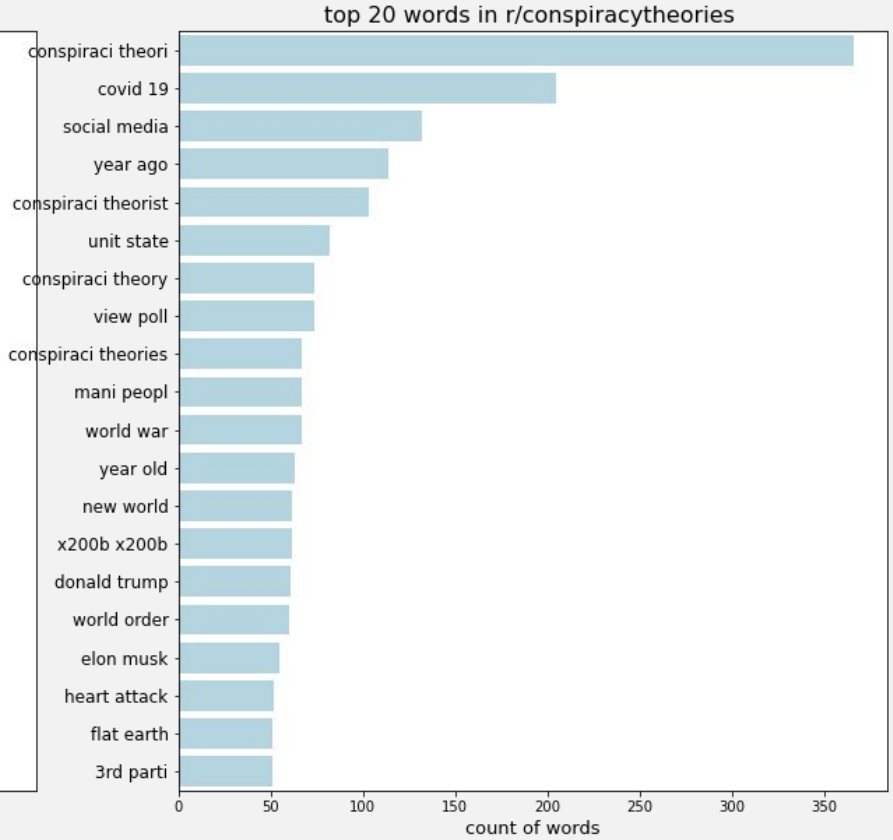
Stemming words to root form

Merged col

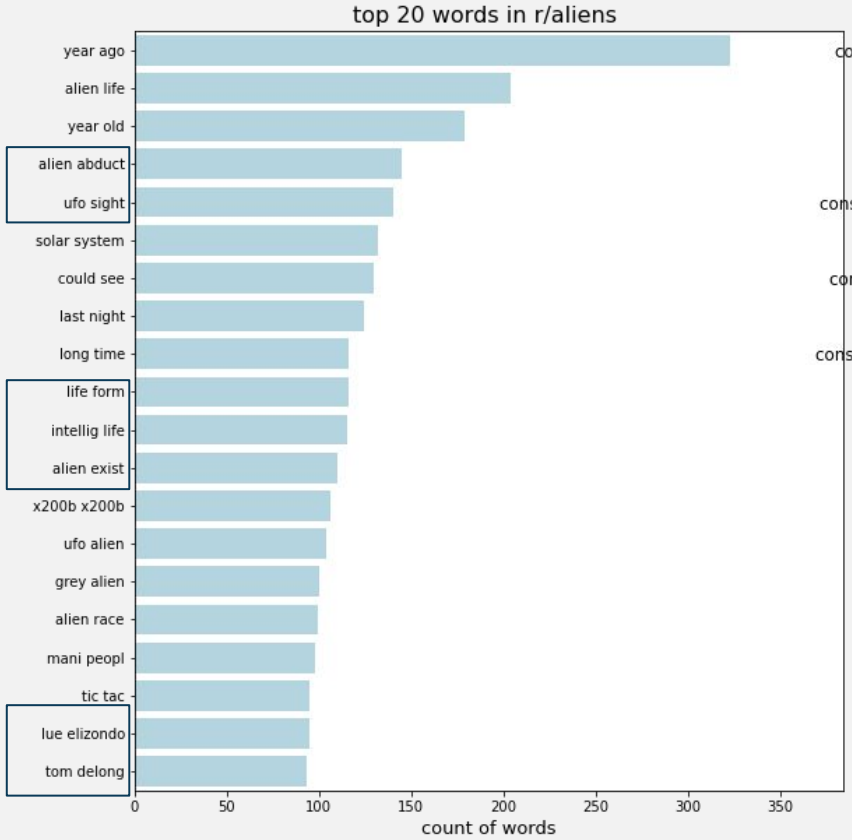
Merged title with selftext columns

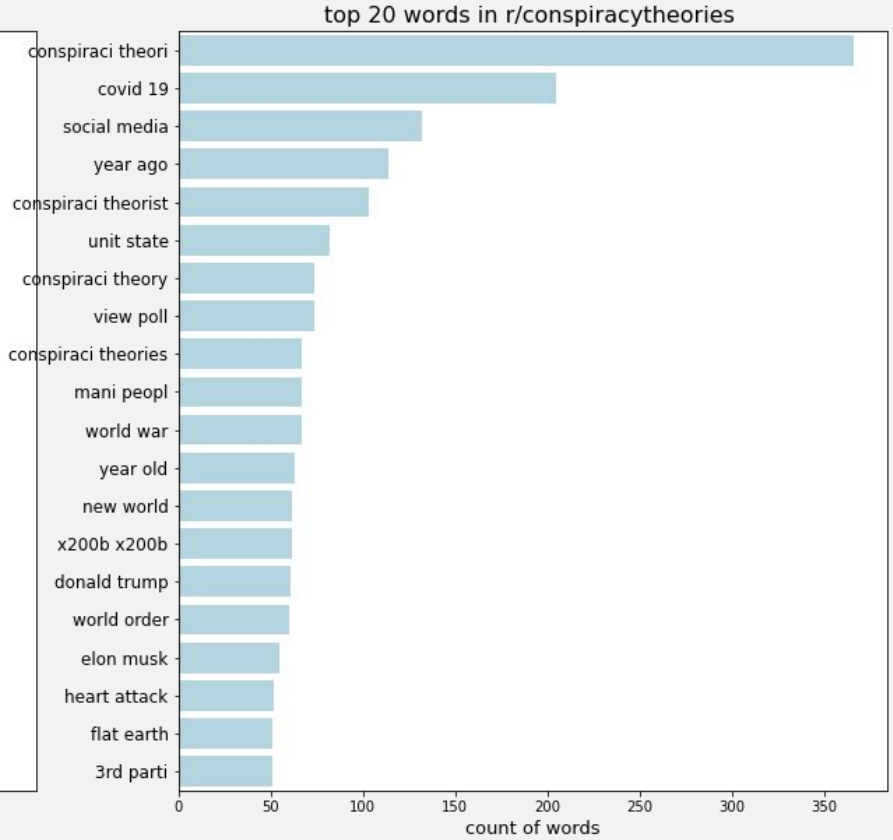
Top 20 bigrams in both subreddits



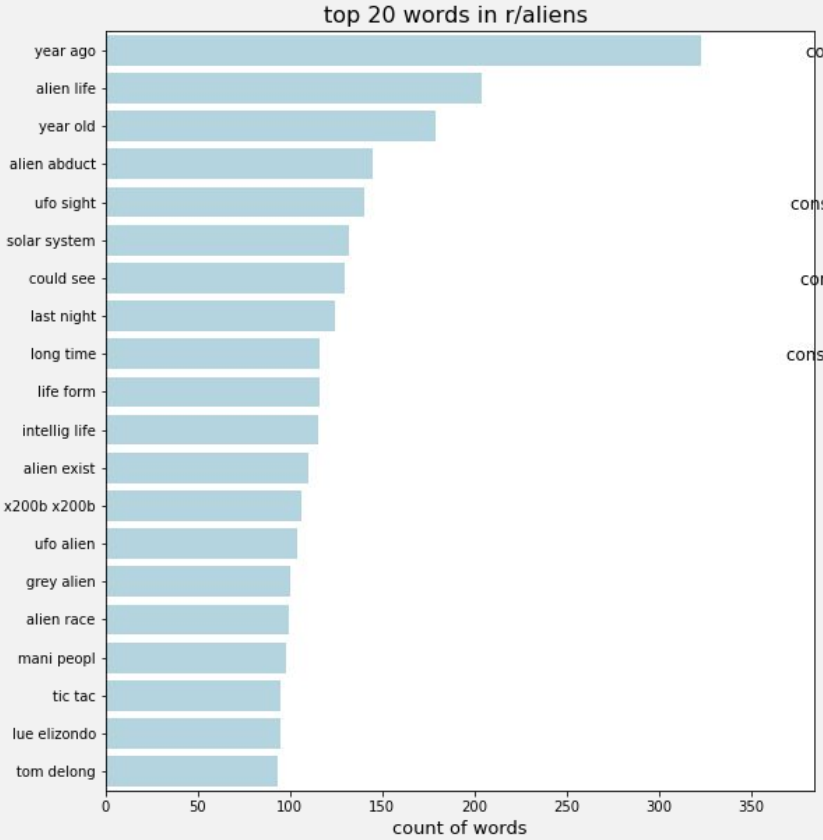


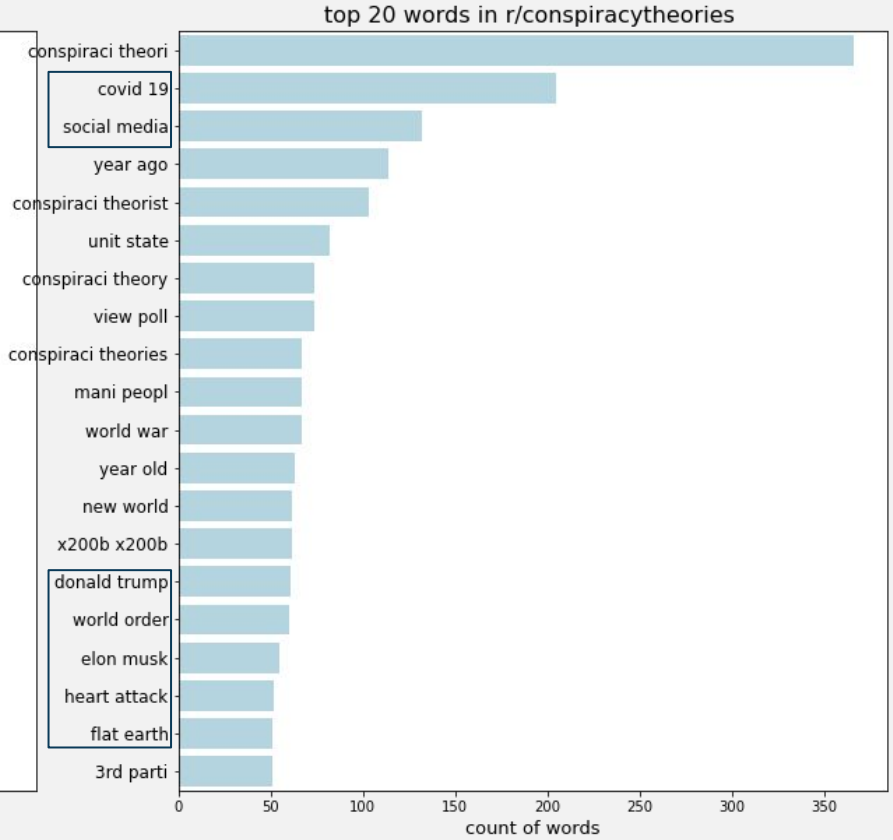
Top 20 bigrams in both subreddits



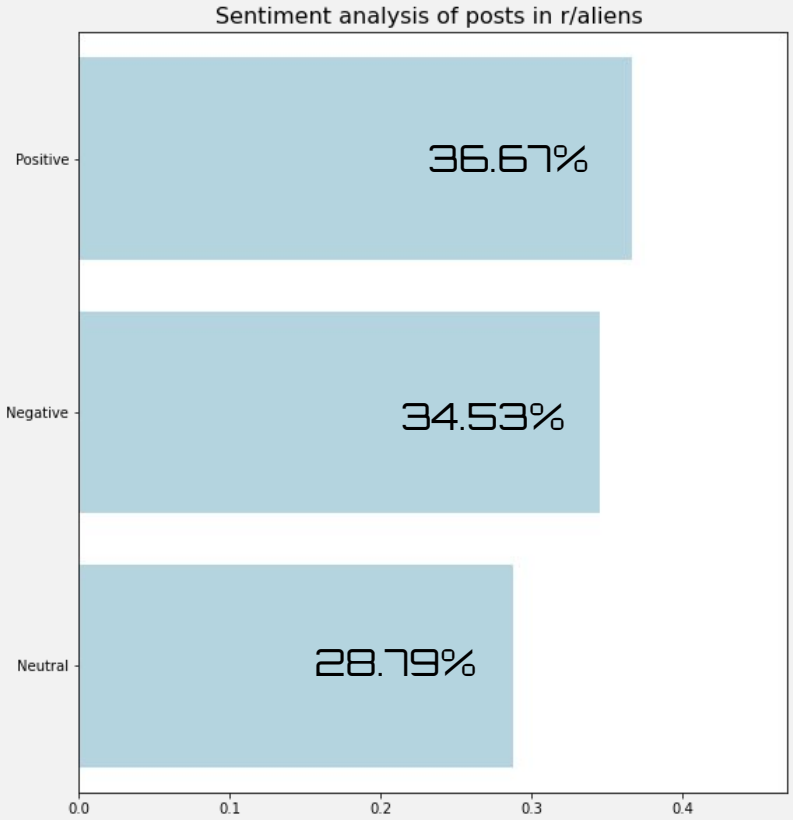


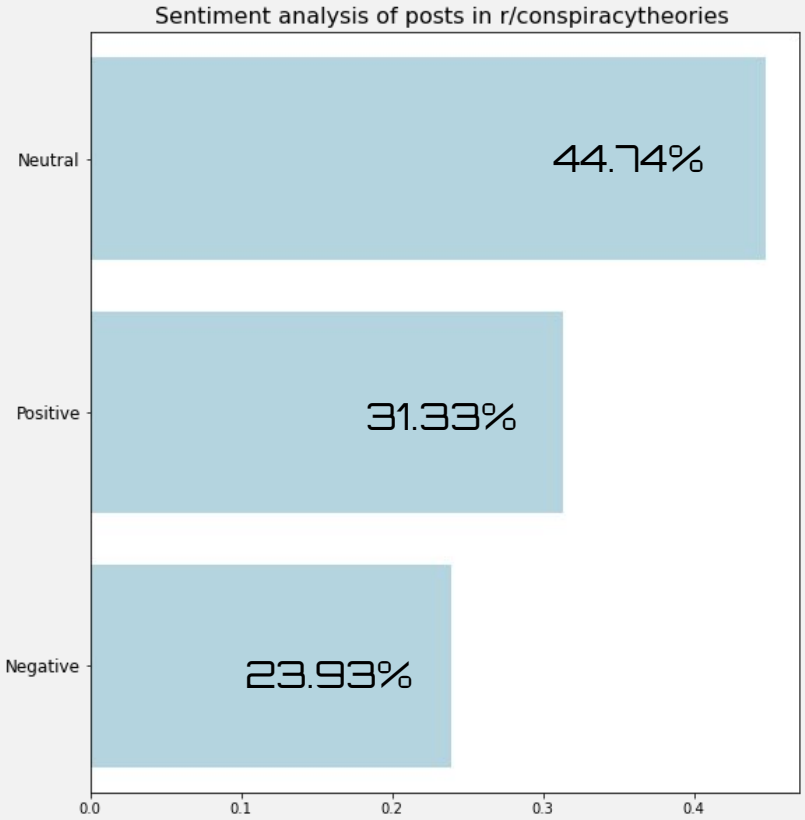
Top 20 bigrams in both subreddits





Sentiment Analysis





Modelling

Train-Test-Split

Test-size = 0.2

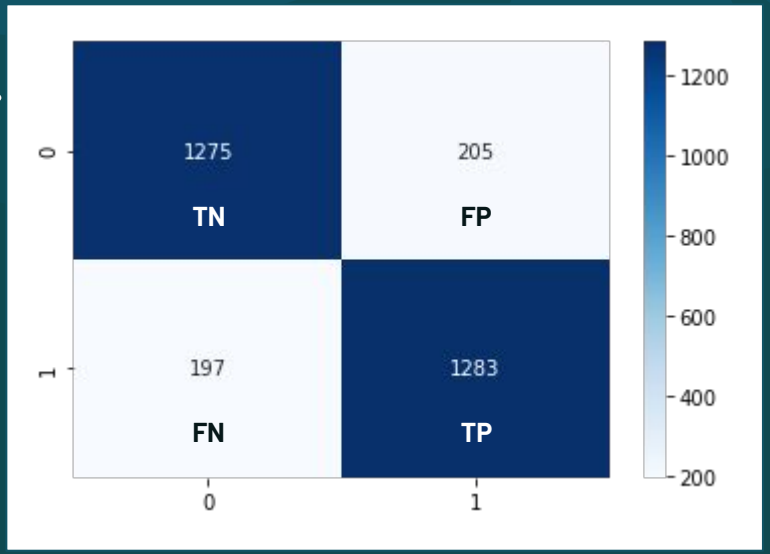
Classification Models

Logistic Regression
Naive Bayes
Random Forest Classifier

Vectorizers

Count Vectorizer
TF -IFD Vectorizer

Best Performing Model



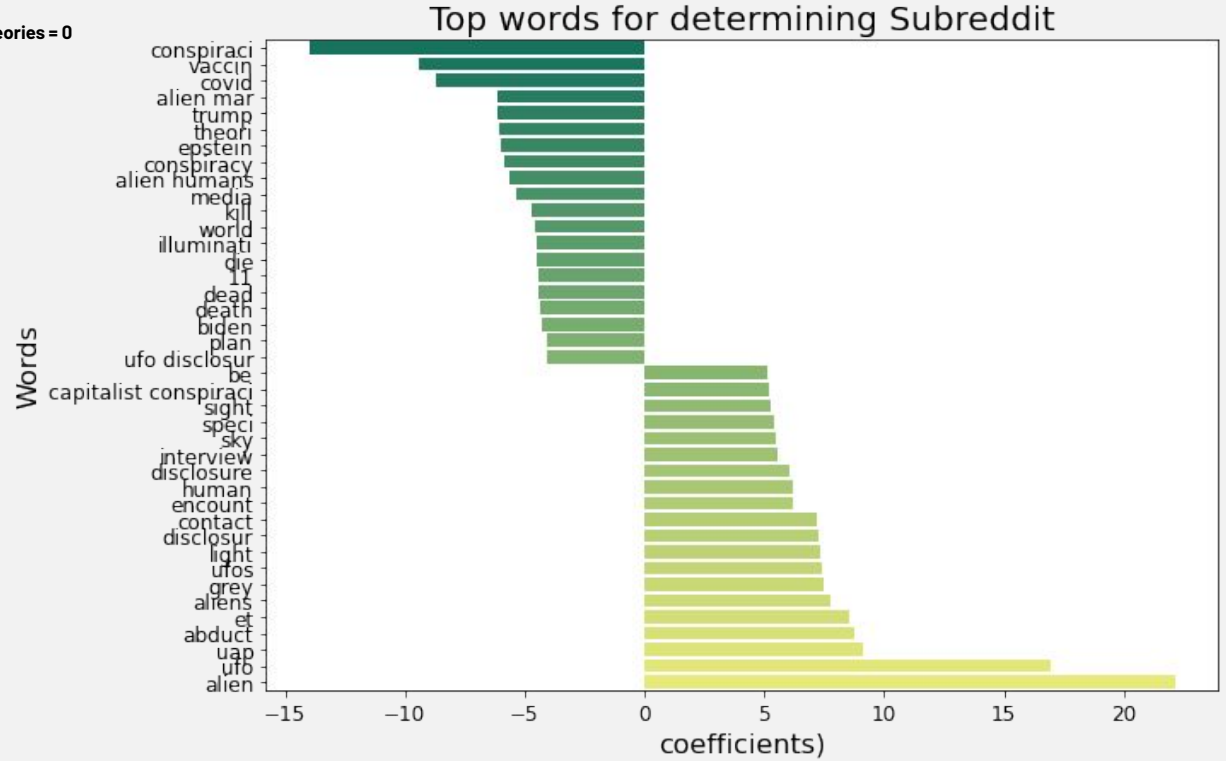
Logistic Regression

TF - IDF Vectorizer

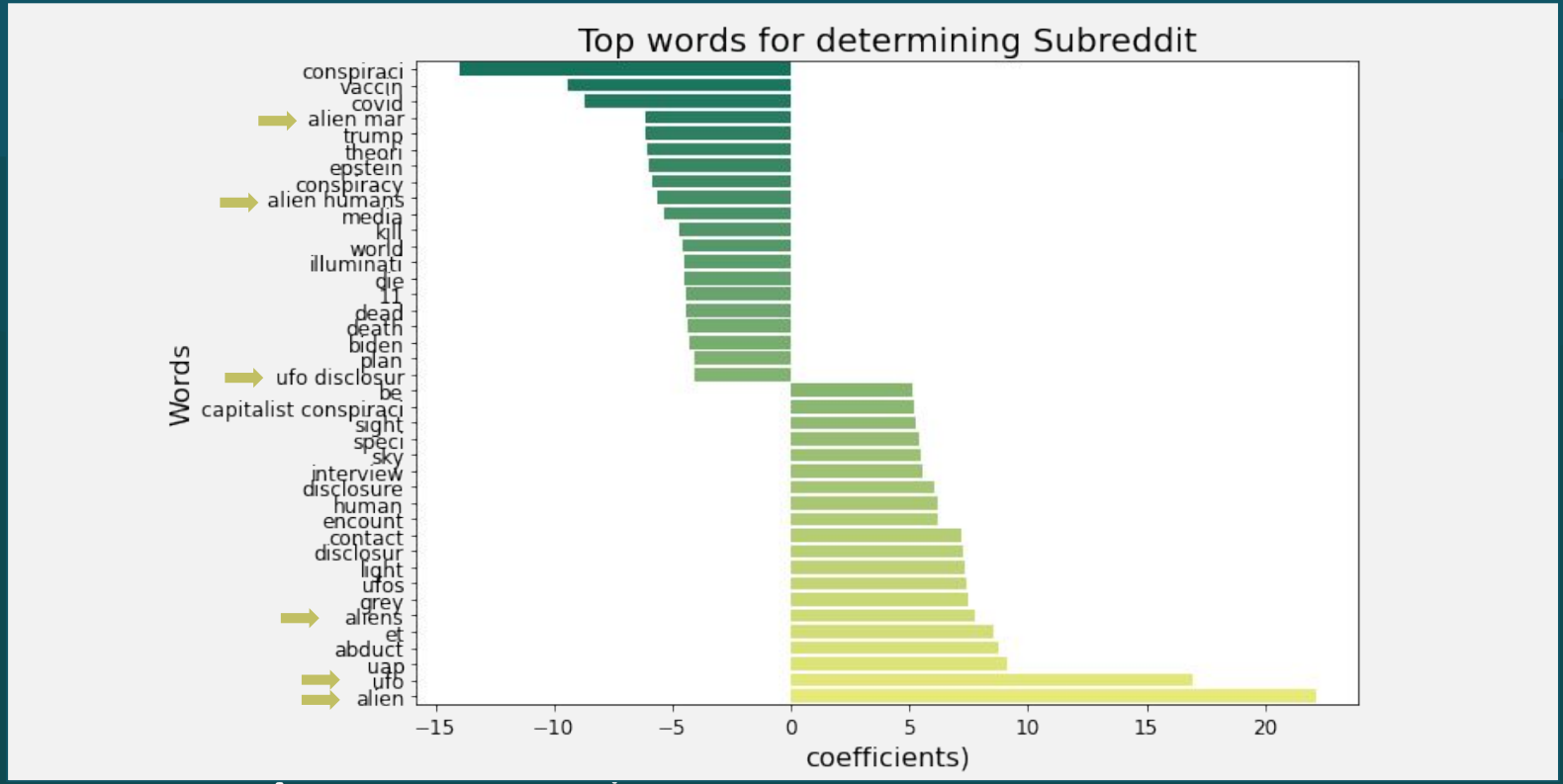
Accuracy - 0.864
F1 Score - 0.865

Model Evaluation, Findings and Analysis

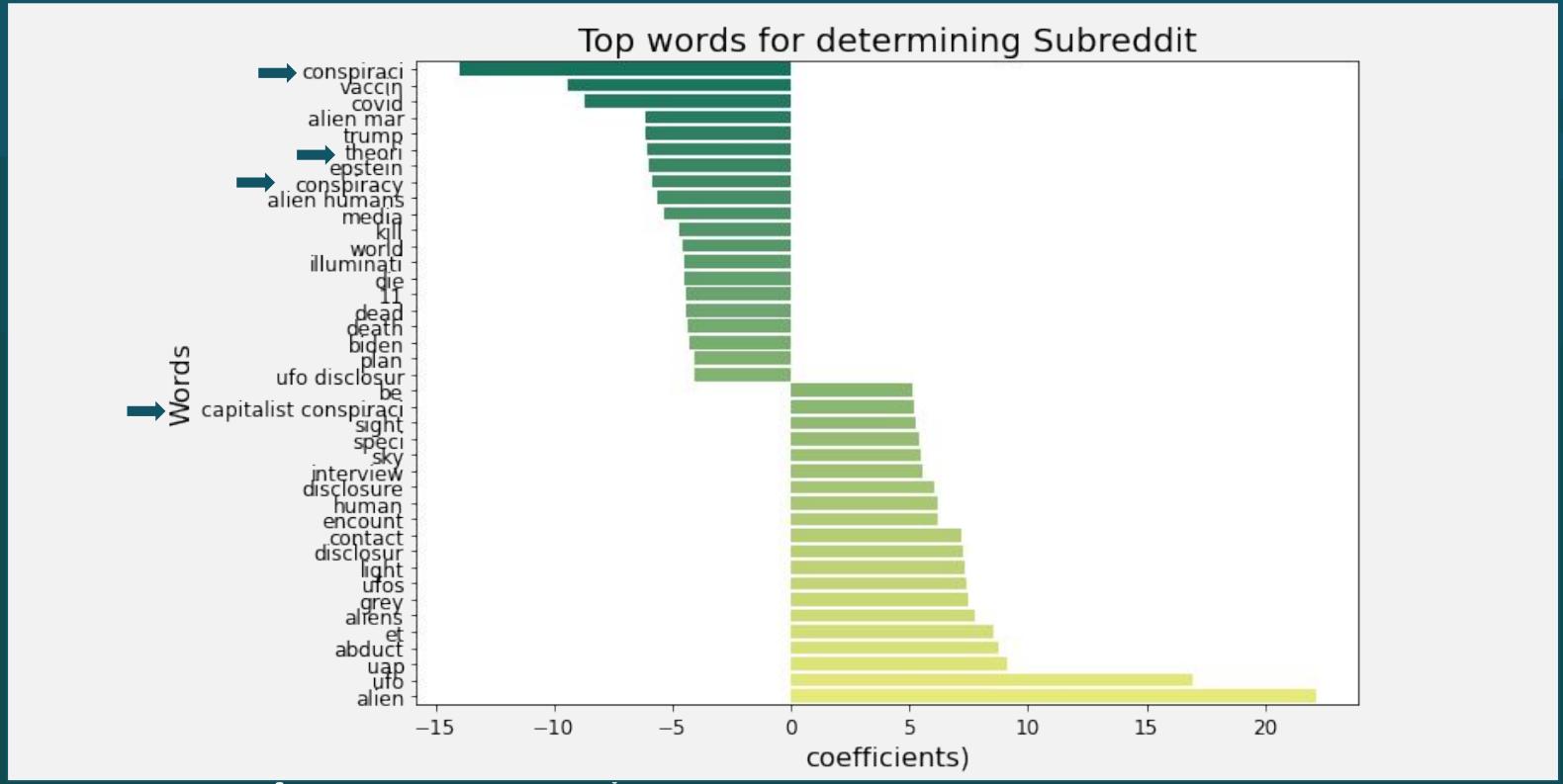
r/aliens = 1
r/conspiracytheories = 0



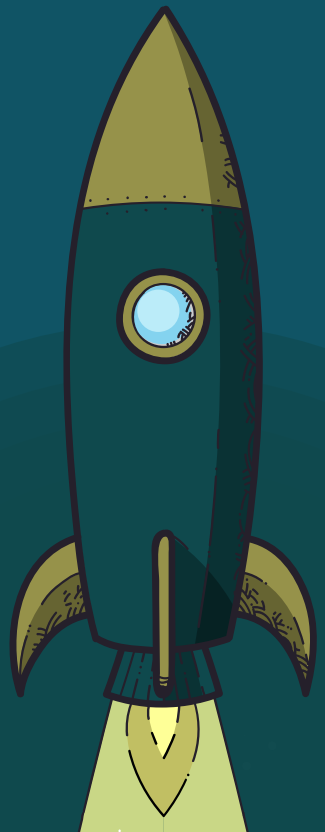
Model Evaluation, Findings and Analysis



Model Evaluation, Findings and Analysis



Conclusion & Recommendations



- Relook at preprocessing methods (add more words in the list of stopwords) and other modelling methods to improve the score further