# Web API's and NLP

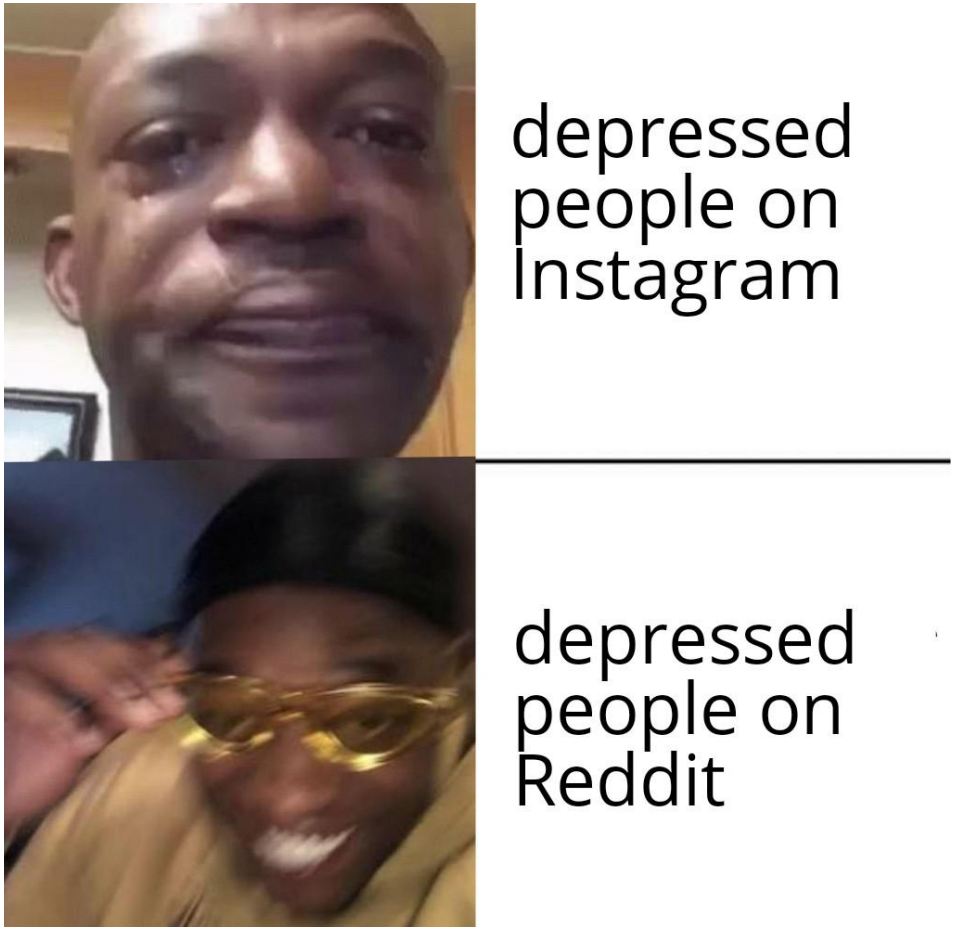
By: Irene Anibogwu
Data Scientist

# Problem Statement:

- We've been tasked with using NLP to train 2 classifier models that can accurately predict which sub reddit a given post is from.
- Subreddits: r/relationship_advice, r/confession
- Chosen classifier models: Logistic Regression and Naive Bayes
- Draw conclusions of which model is better
- Tie in information on how our models can be used for marketing purposes

# What is Reddit?

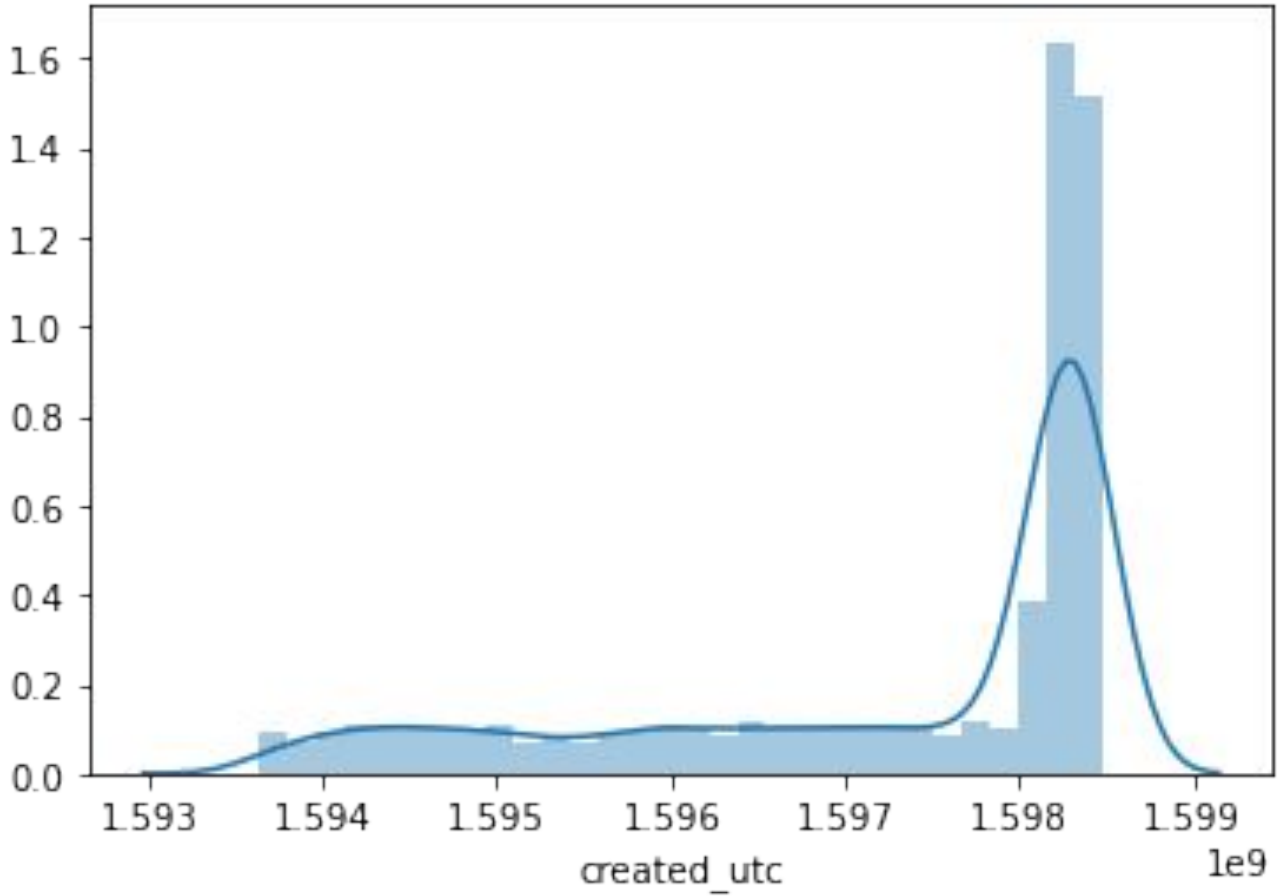
- Social news aggregation, web content rating and discussion site
- This online user community has over 300 million active users
- Relationship advice: 3.9 million members
- Confession: 2.2 million members

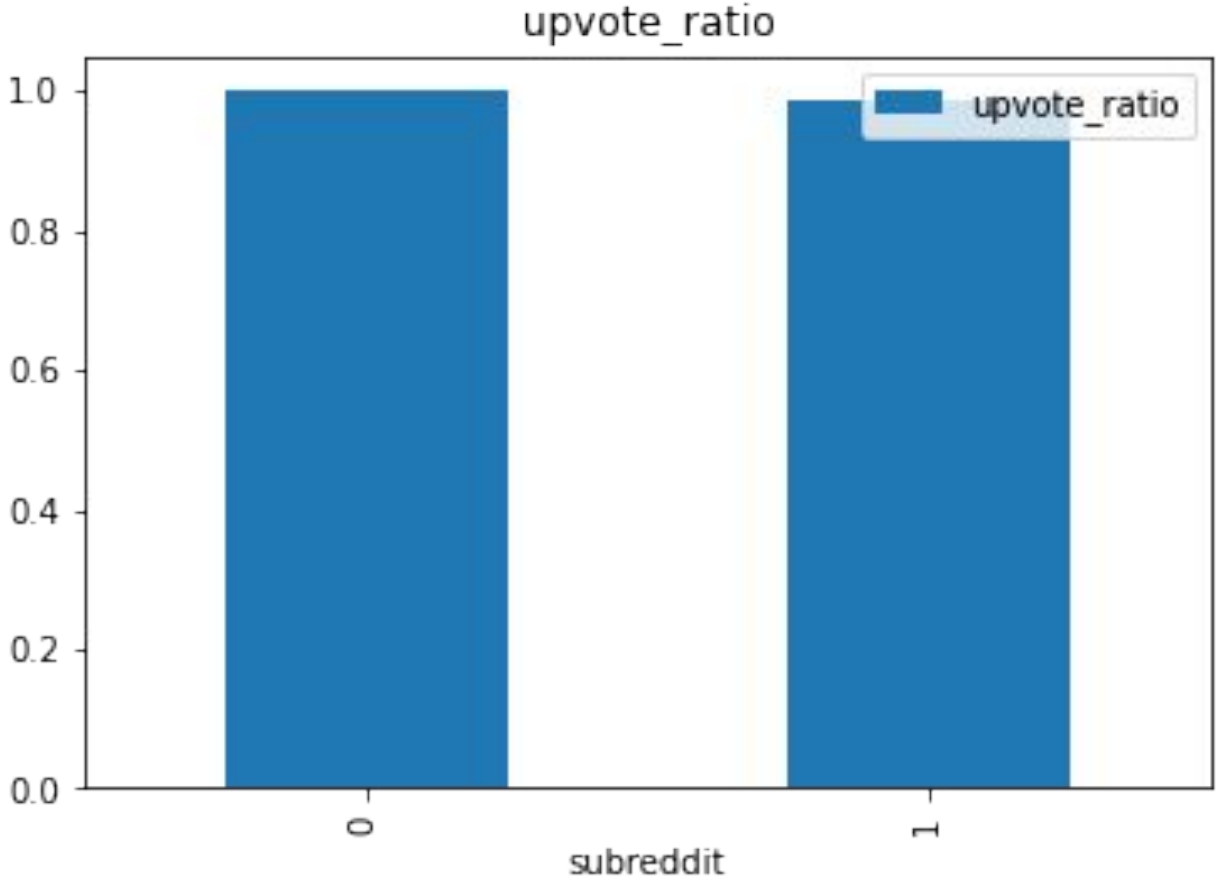


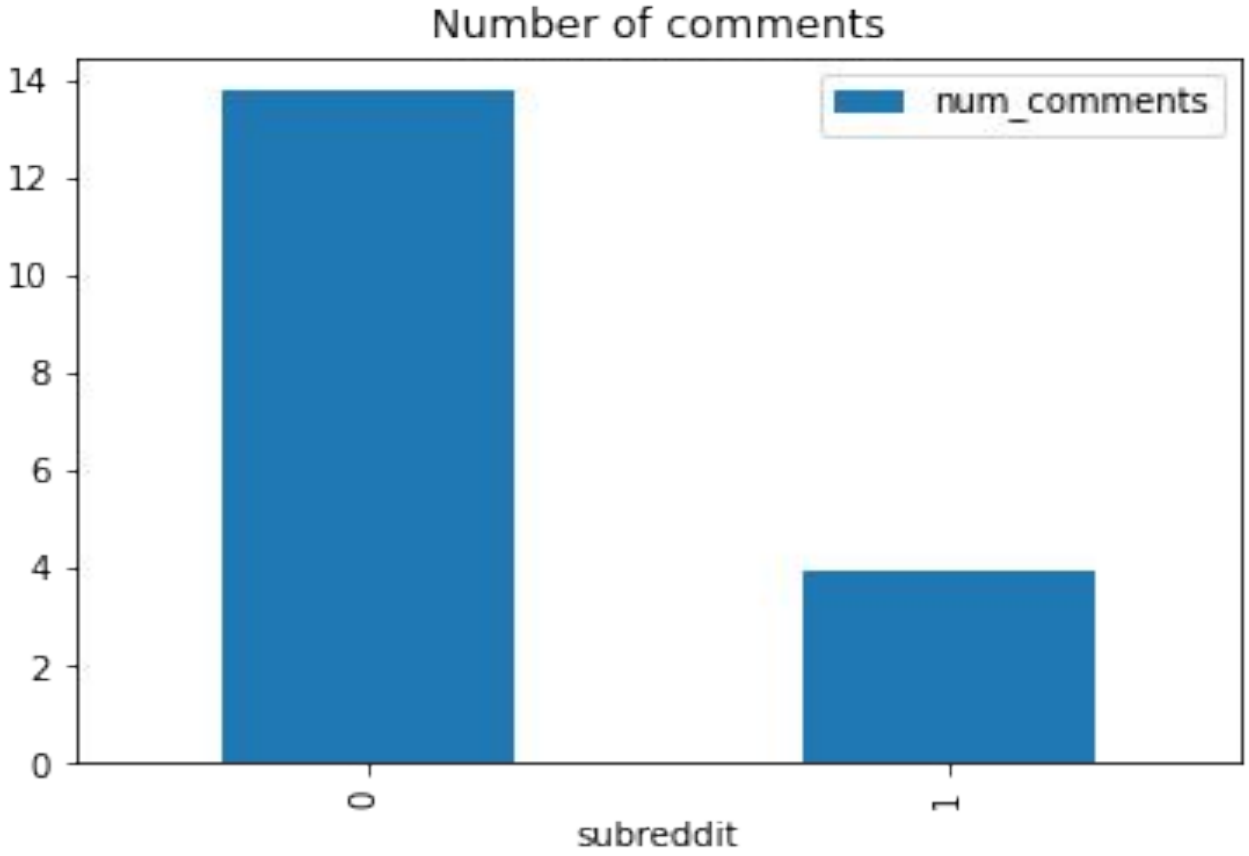
# Initial Data Cleaning & EDA

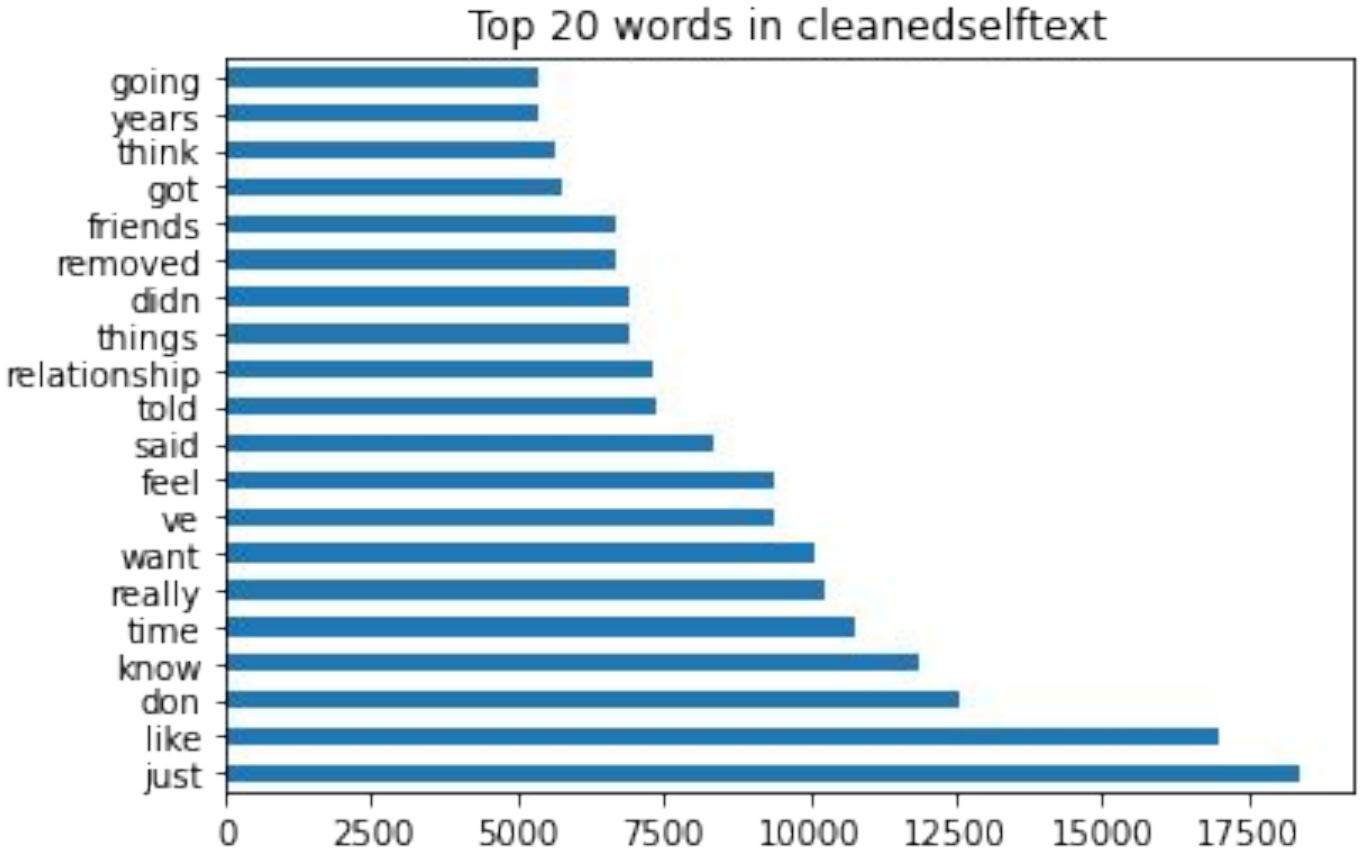
- Used Push shift API to pull data from Subreddits of our choice
- About 16,206 rows of data (8500 from each subreddit )
- Converted subreddits {Relationship Advice: 0 Confessions: 1}
- Removed stopwords, tokenized and dropped columns

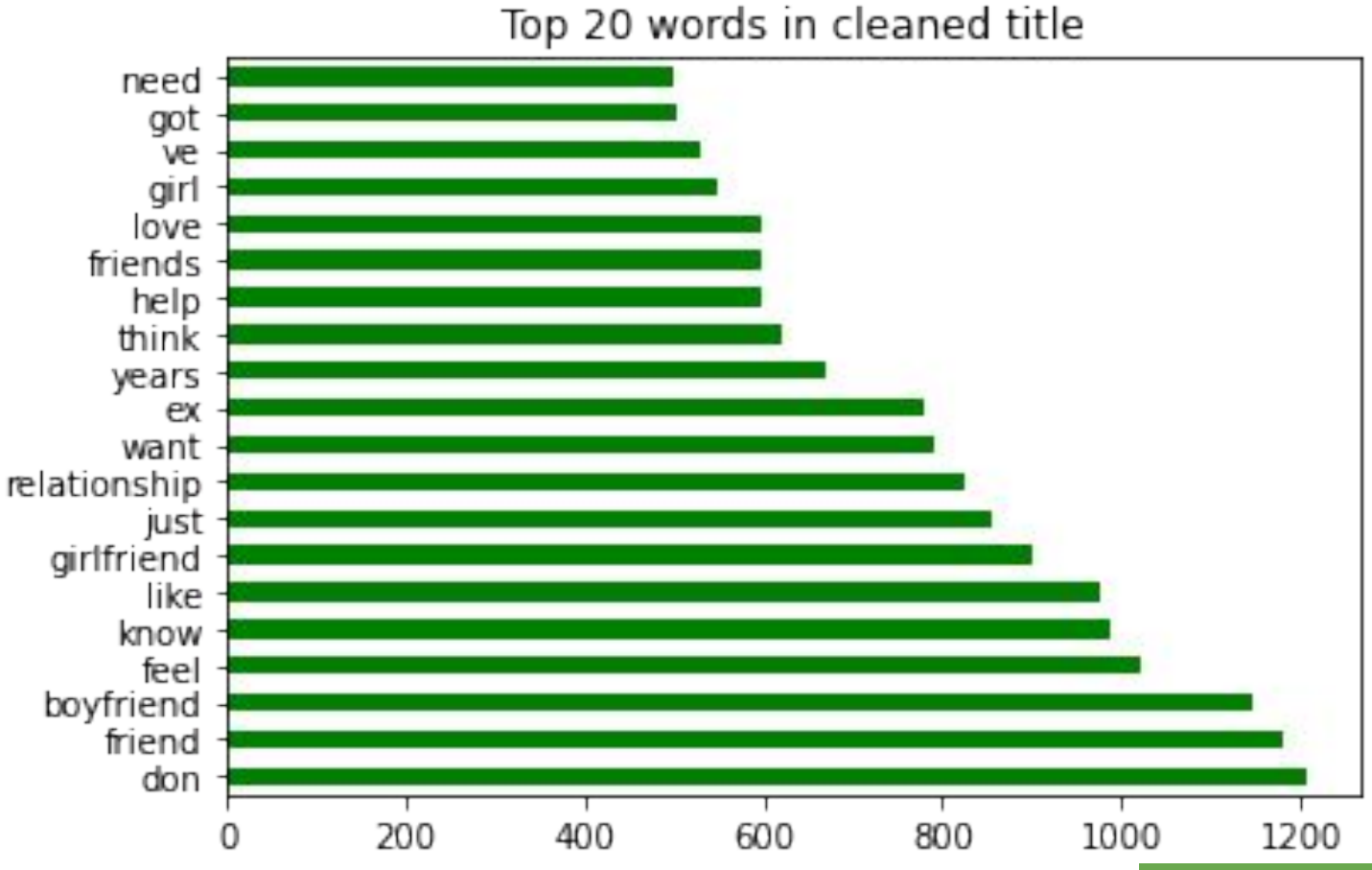
# August 1st 2020 - August 26th 2020




upvote_ratio
upvote_ratio
1.0
0.8
0.6
0.4
0.2
0.0
0
1
subreddit


Number of comments
num_comments
14
12
10
8
6
4
2
0
0
1
subreddit


Top 20 words in cleanedselftext
going
years
think
got
friends
removed
didn
things
relationship
told
said
feel
ve
want
really
time
know
don
like
just
0
2500
5000
7500
10000
12500
15000
17500




** Column used for models

# So which is better?

# Log regression confusion matrix

**WINNER!**

Accuracy ~ 0.852

Sensitivity ~ 0.856

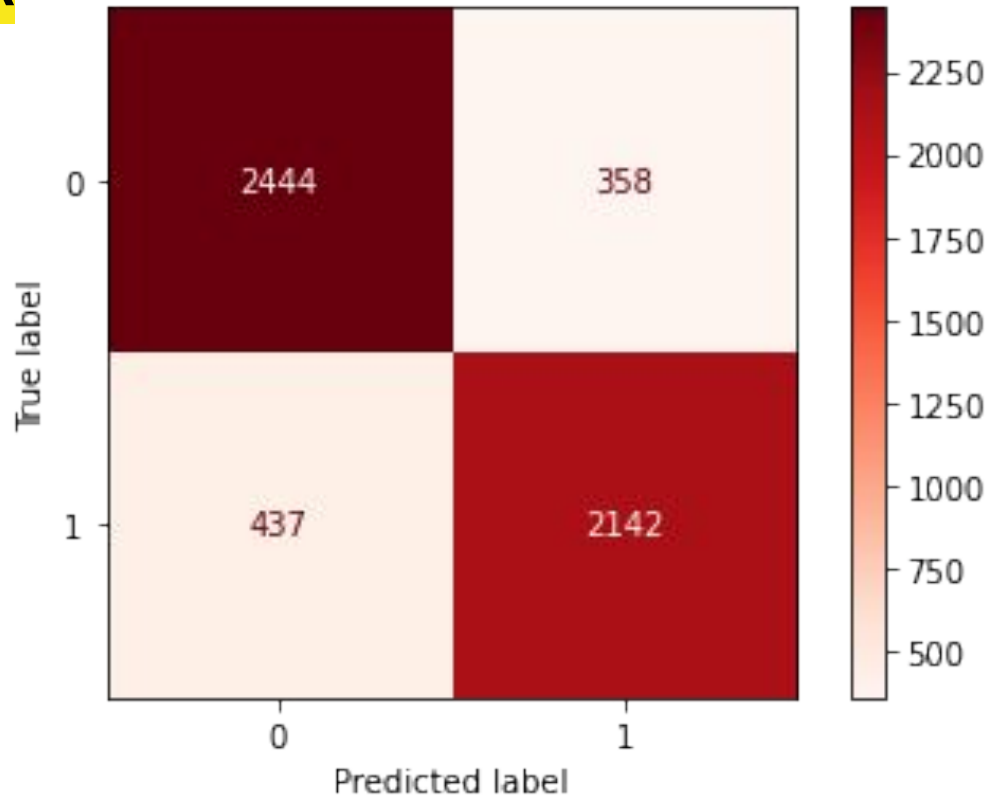
Misclassification ~0.147

Specificity ~ 0.872

Baseline:

0 ~ 0.52

1 ~ 0.47



# Naives Bayes GridSearch Confusion Matrix

Accuracy ~ 0.849

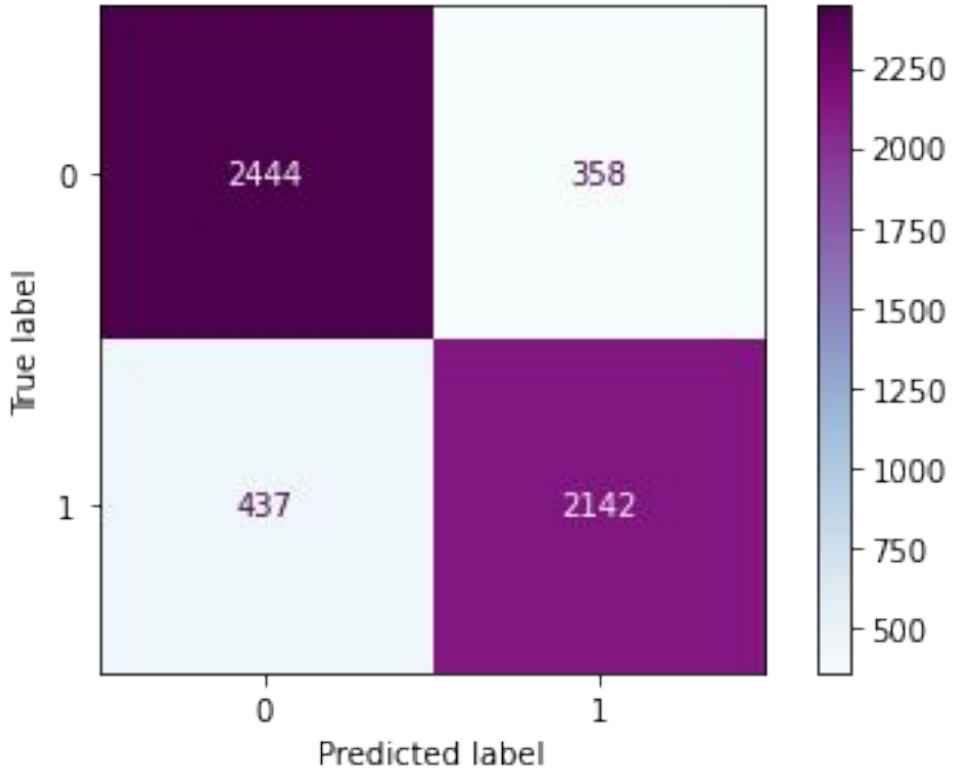
Sensitivity ~ 0.845

Misclassification ~ 0.15

Specificity ~ 0.858

Baseline~ 0- 0.52

1- 0.47



# Positives/Negatives of  Log models

Advantage:

- Highly interpretable, requires little computational resources
- Doesn’t need to be scaled/easy to regularize
- Trained easily and also easy to implement

Disadvantage:

- Not easy to solve non-linear problems with this approach
- Sensitive to outliers

# Positives/Negatives of Naive Bayes(w/Grid Search) Classifier

Advantages:

- Easy to implement
- Makes dealing with missing values easier
- Can be used for both binary and multi classification problems
- Not sensitive to irrelevant features

Disadvantages:

- Assumption of independent predictor features

# How can this model be used?

- Models outperformed baseline
- Bloggers/professionals who work in the space of counseling, relationship coaching , etc
- Top predictor words show what they can add to their blogs outside of reddit
- Also be an active contributor to the channel

# Sources:


https://buffer.com/resources/reddit-marketing-strategies-for-those-who-dont-have-time-for-reddit-marketing/

https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/

https://www.reddit.com/r/relationship_advice/

https://www.reddit.com/r/confession/

# Questions?: