Project 3: The Reddit Classifier

r/Anger & r/Meditation - Vidhu
r/MentalHealth & r/Psychology - Clement

The Problem

- Reddit was down and the tag of the subreddit somehow did not get saved
- Data science team was tasked to reclassify the posts into their subreddits (/rAnger, r/Meditation and r/MentalHealth, r/Psychology)
- Not all heros wear capes!

Approach Used

- Web scrape reddit to collect posts for 2 subreddit categories

- Make a dataset of combined posts

- Label categories as
  - Category 1 - 1
  - Category 2 - 0
- Save (on the disk as csv files) the big dataset for train/test purpose during modeling

- Read dataset to do cleaning
  - do train/test split
  - consider only words by using regex
  - lower case the characters in posts
  - apply lemmatization

- Use Count Vectorizer using stop-words to remove these words from features
  - get the features list
  - do this for train, test, final test dataset

- Train model

- For predictions
  - check for train/test score
  - accuracy of validation set

- Compare and recommend the best model based on comparison

Subreddits used

Anger

Meditation

Approach Used

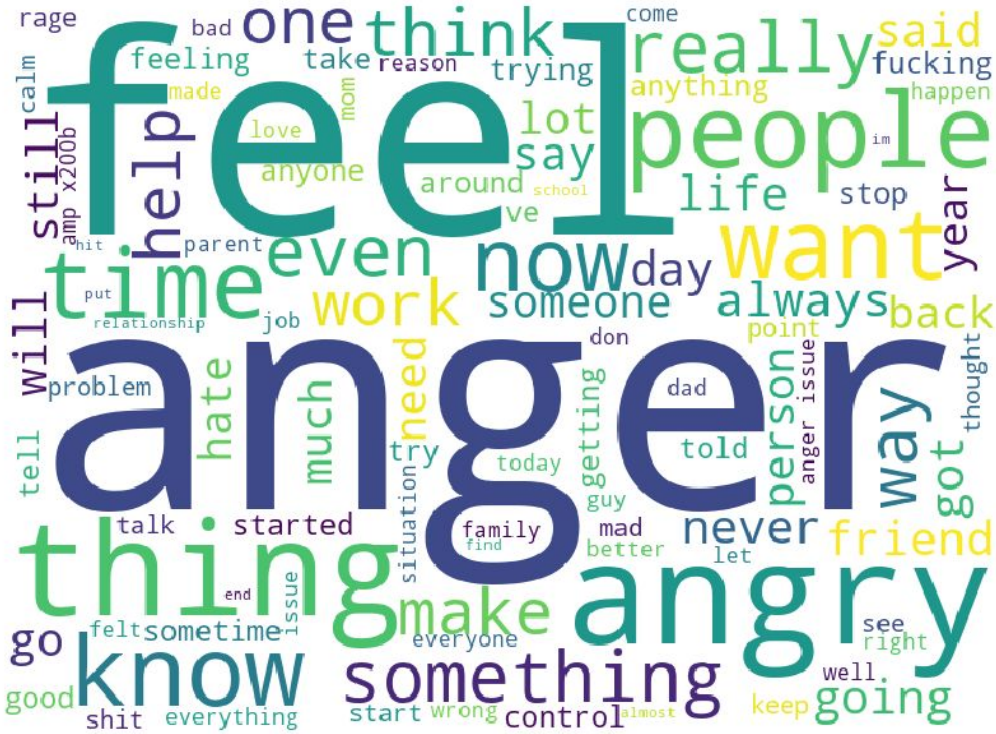
- Web scrape reddit to collect posts for category - **Anger, Meditation**
- Make a dataset of combined posts
- Label categories as
  - **Anger - 1**
  - **Meditation - 0**
- From this dataset, **randomly select 100 posts which will be used to test predictions** on various model that will be used for modeling
- Save (on the disk as csv files) the big dataset for train/test purpose during modeling and sliced dataset as final test on model
- Read both dataset on do cleaning
  - do train/test split
  - consider only words by using regex
  - lower case the characters in posts
  - apply lemmatization
- **Use Count Vectorizer using stop-words(including 'Anger' and 'Meditation' and synonym words to these categories) to remove these words from features**

```
0.9170274170274171
{'cvec__max_df': 0.9, 'cvec__max_features': 3500, 'cvec__min_df': 1, 'cvec__ngram_range': (1, 1)}
```

  - get the features list
  - do this for train, test, final test dataset
- Train model
- For predictions
  - **check for train/test score, final test score**
  - accuracy of validation set and final test set
- Compare and recommend the best model based on comparison

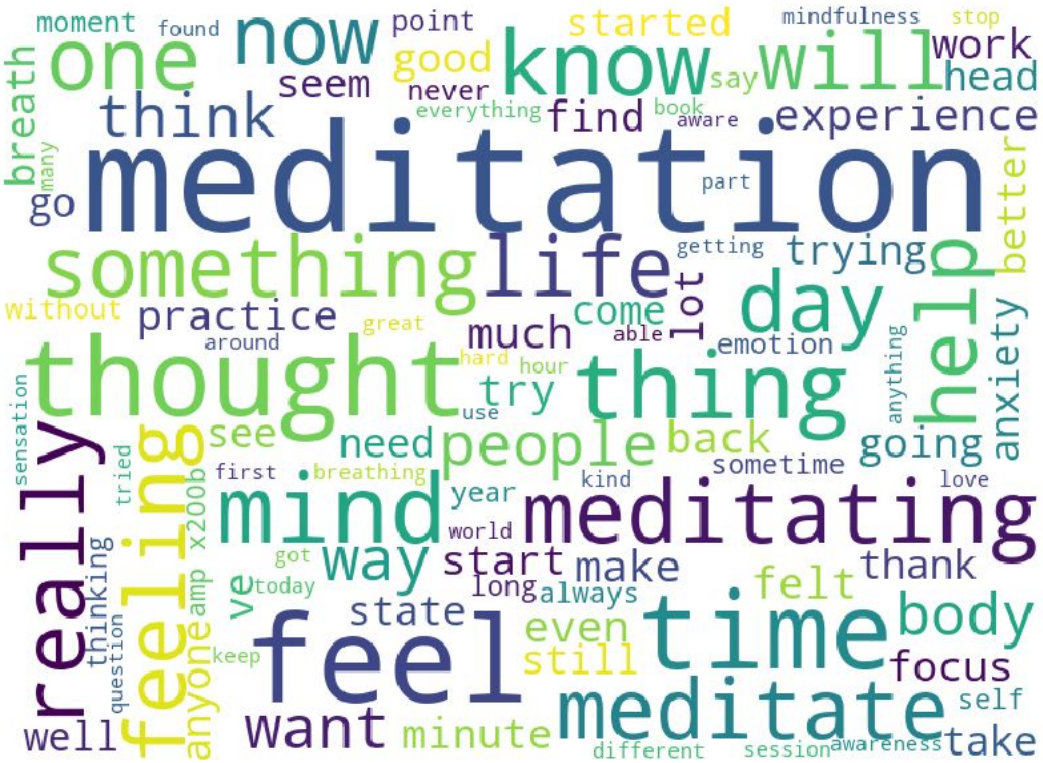
Word Cloud of category – Anger

- Word cloud before removing category related words from posts

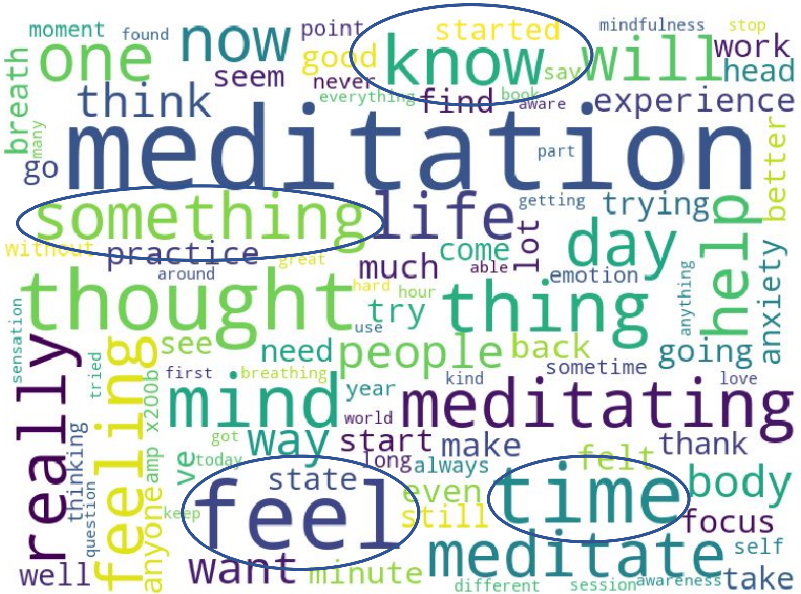


Word Cloud of category – Meditation

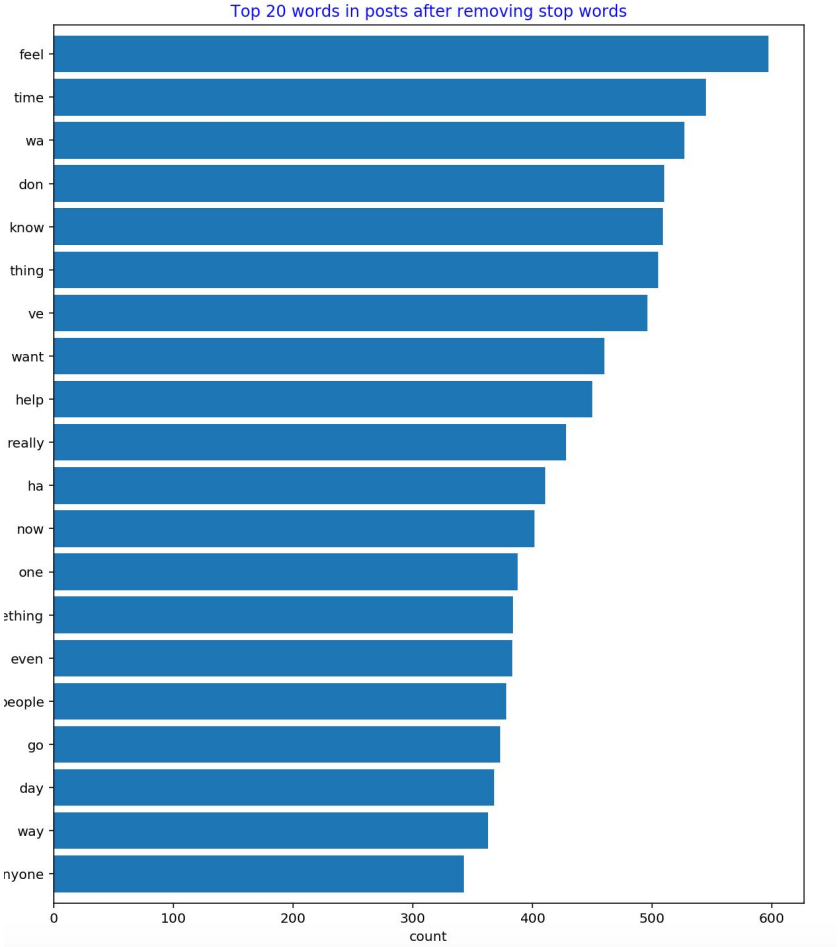
- Word cloud before removing category related words from posts



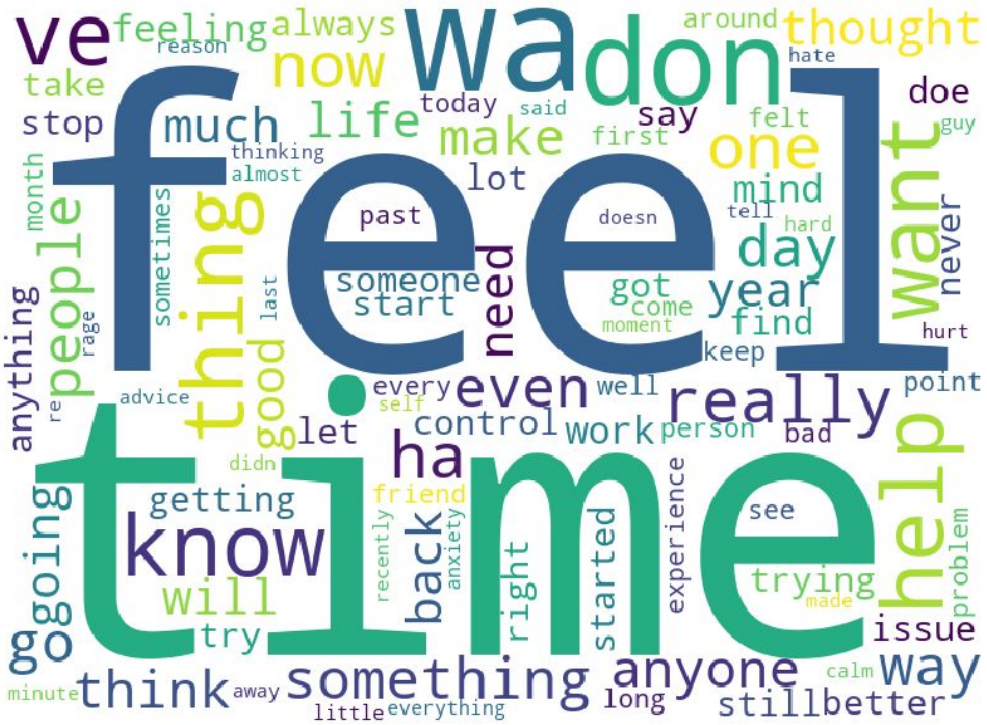
Common high frequency words between two categories



Top 20 words from the bag of words after removing stop words – including category related words



Word Cloud of common words after cleaning posts



Modeling

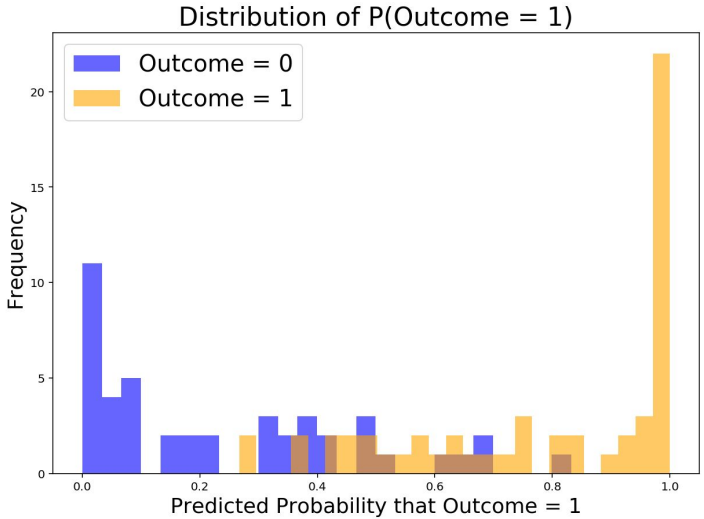
- Logistic Regression model using Lasso
- Logistic Regression model using Ridge
- Multinomial Naïve Bayes model
- Random Forest

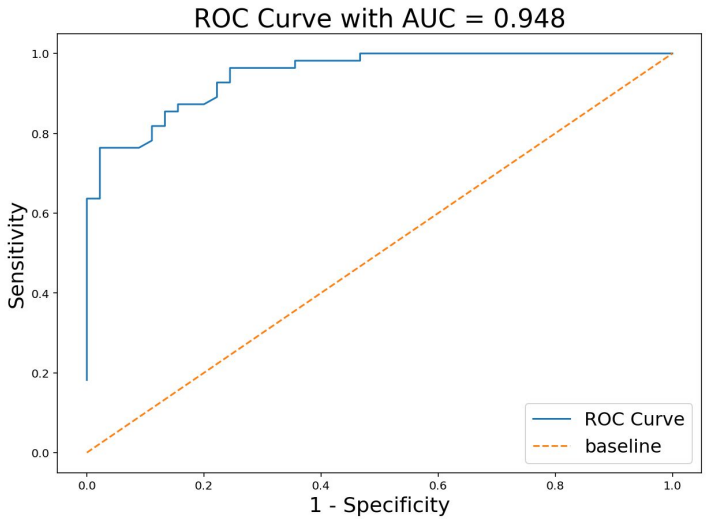
Comparison Matrix

	train_score	test_score	val_accrcy	final_test_accrcy	final_test_specificity	final_test_sensitivity	total_test_posts	fale_postive	fale_negative
Lasso	0.967893	0.840909	0.840909	0.81	0.844444	0.781818	100	7	12
Ridge	0.959235	0.864719	0.864719	0.84	0.866667	0.818182	100	6	10
MultiNomialNB	0.910173	0.884199	0.884199	0.84	0.8	0.872727	100	9	7
RandomForest	0.890693	0.805195	0.805195	0.79	0.822222	0.763636	100	8	13

Naive Bayes (Multinomial) Classification Model

Predicted Probability and ROC Curve with AUC on final Test Set





Conclusion and Recommendation

- Have fit 4 models on reddit posts with two categories - Anger and Meditation labeled as 1 and 0
  - Lasso Logistic Regression model
  - Ridge Logistic Regression model
  - Multinomial Naive Bayes model
  - Random Forest model
- From above metrics it can be seen that Multinomial NB and Ridge Logistic Regression have good validation and accuracy score with final Test accuracy of 0.84
- Need more posts to train the model so that model can give good test accuracy and can be used for best predictions
- From above metrics Multinomial NB can be considered as best fit

Project 3: Reddit Classifier

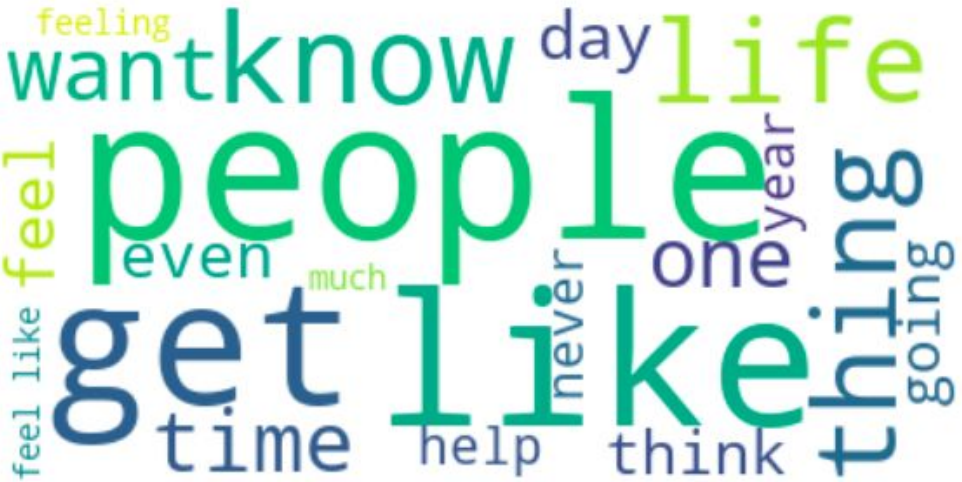
r/MentalHealth & r/Psychology

Subreddits used

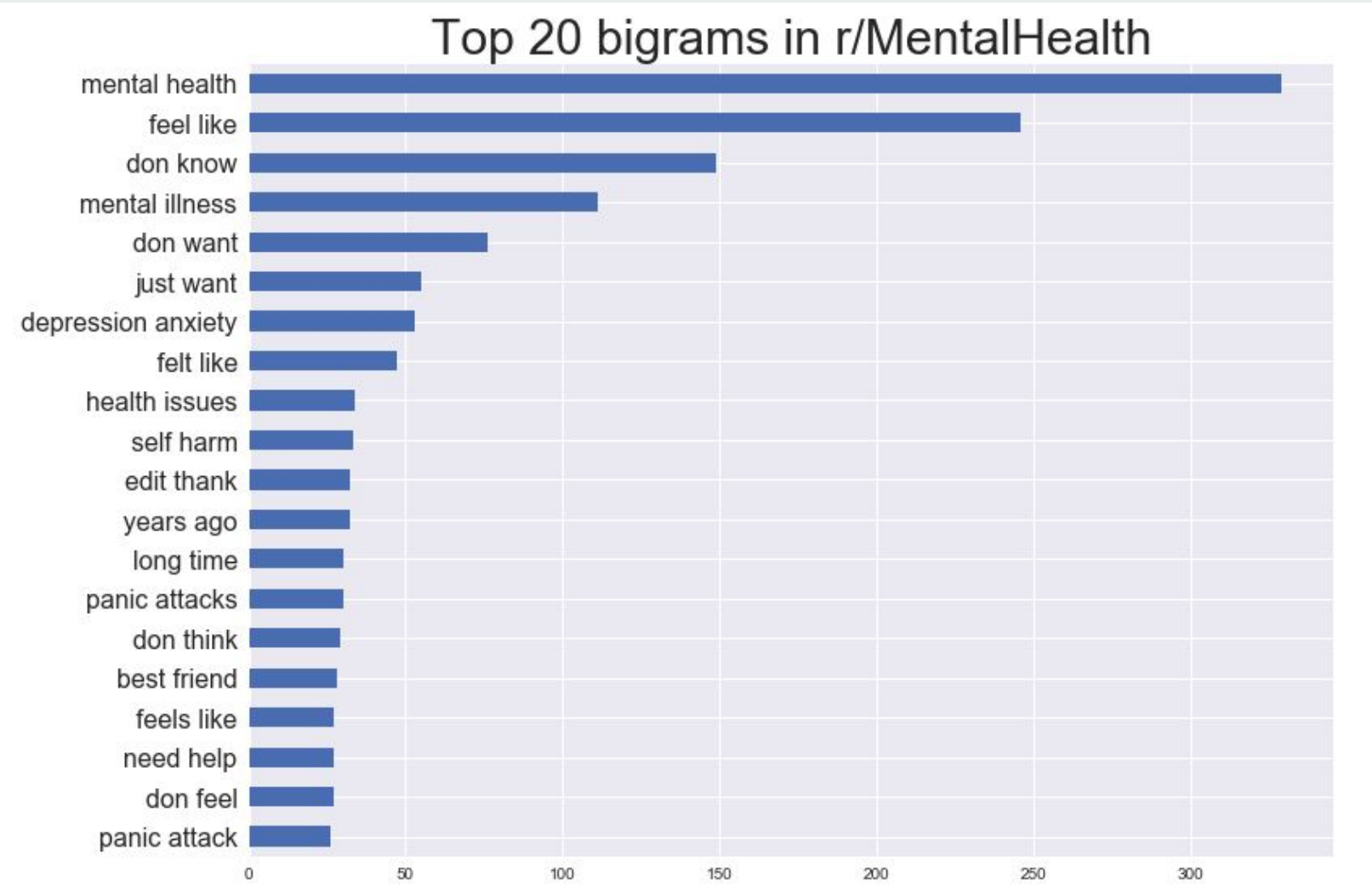
- r/Psychology
- r/MentalHealth

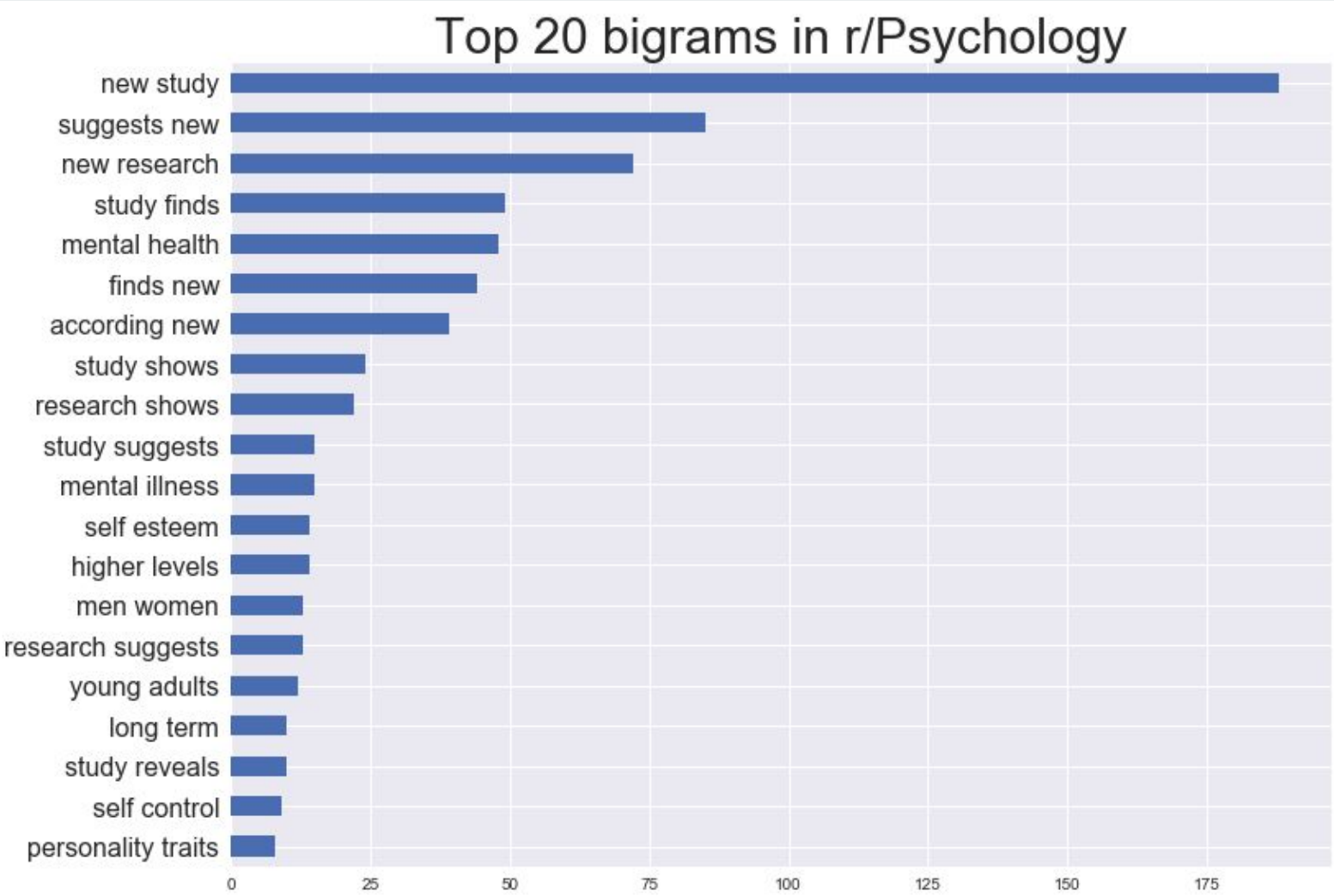
EDA

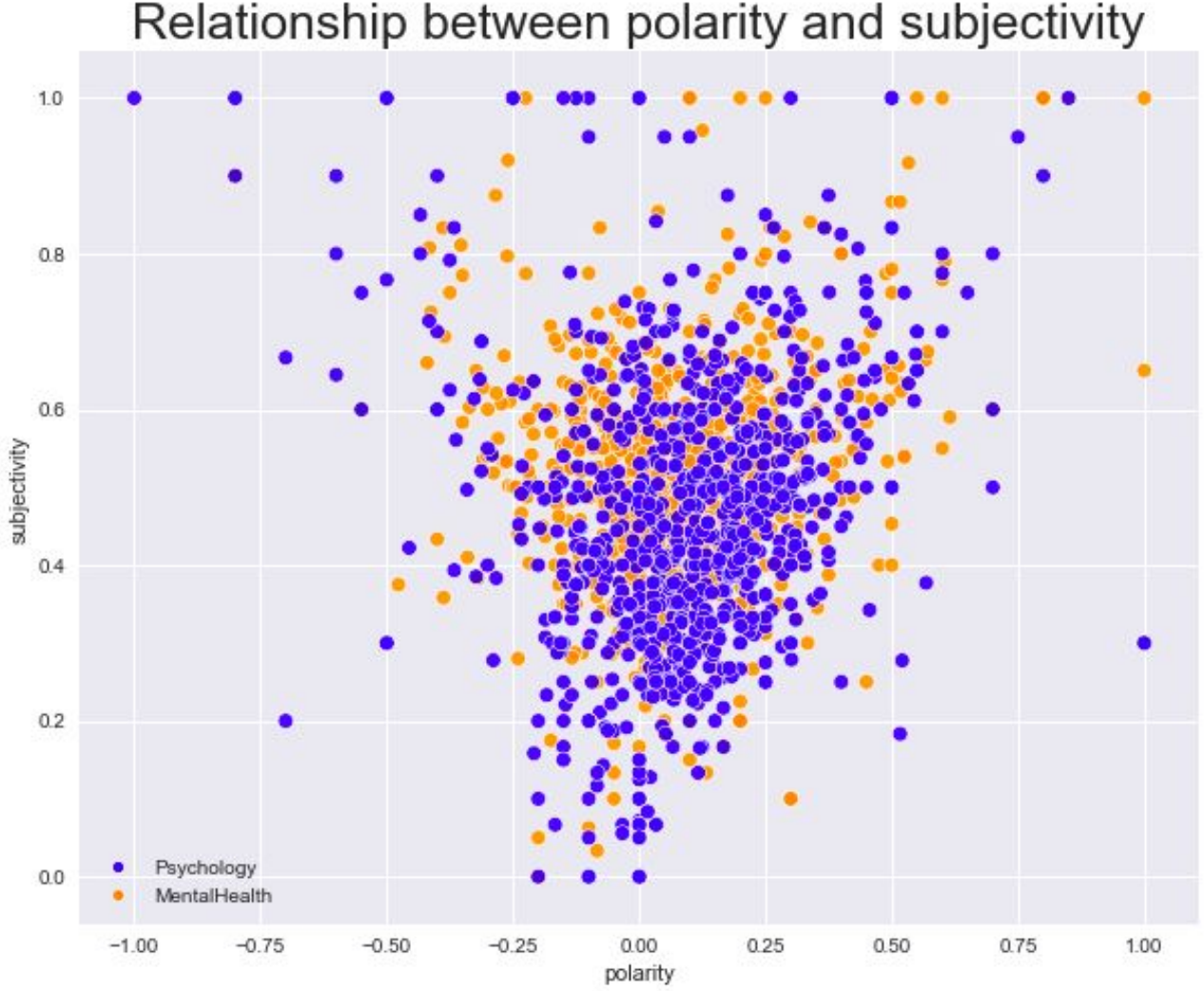
r/MentalHealth

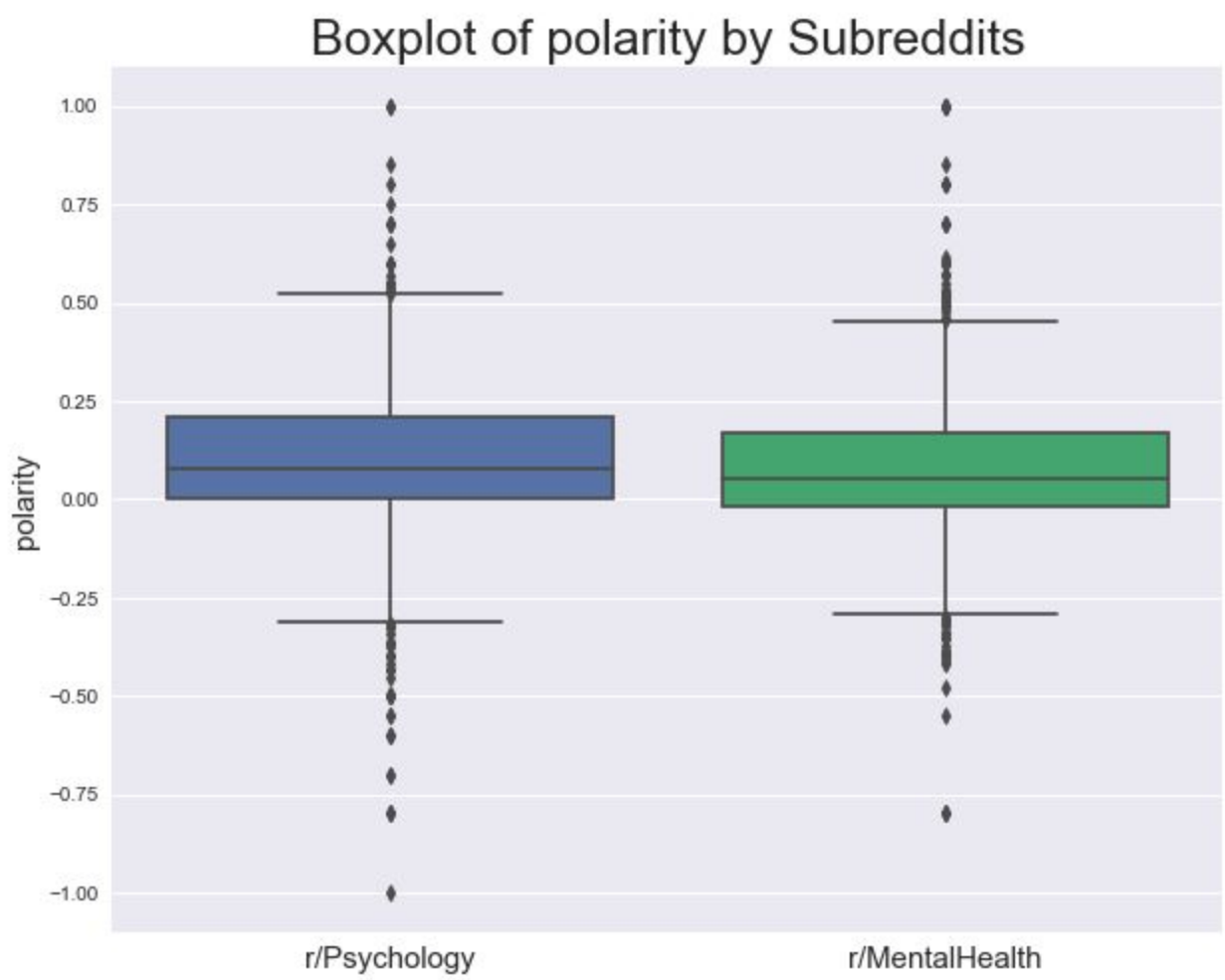


r/Psychology


Top 20 bigrams in r/MentalHealth
mental health
feel like
don know
mental illness
don want
just want
depression anxiety
felt like
health issues
self harm
edit thank
years ago
long time
panic attacks
don think
best friend
feels like
need help
don feel
panic attack
0
50
100
150
200
250
300


Top 20 bigrams in r/Psychology
new study
suggests new
new research
study finds
mental health
finds new
according new
study shows
research shows
study suggests
mental illness
self esteem
higher levels
men women
research suggests
young adults
long term
study reveals
self control
personality traits
0
25
50
75
100
125
150
175


Relationship between polarity and subjectivity
subjectivity
polarity
1.0
0.8
0.6
0.4
0.2
0.0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Psychology
MentalHealth

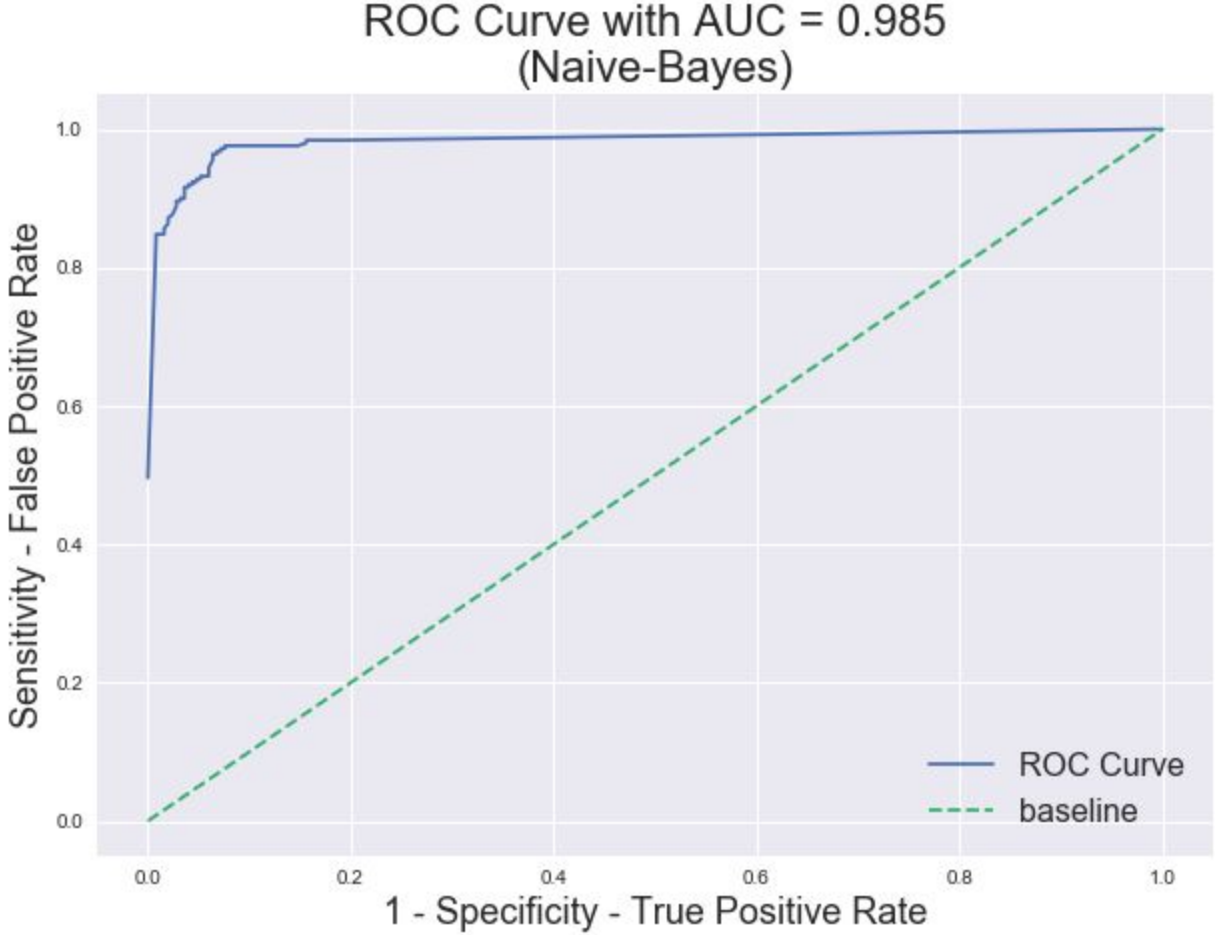

Boxplot of polarity by Subreddits
polarity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
r/Psychology
r/MentalHealth

Modelling

Summary of modelling

Model	Training Score	Test Score	ROC AUC Score
Logistic Regression	0.999	0.948	0.98
Naive-Bayes	0.962	0.946	0.985
Random Forest	1	0.944	0.99

r/Psychology	r/MentalHealth
benevol	zoloft
increas risk	know wa
wa link	know think
gap	know thi
bia	know struggl
favour	know peopl
studi reveal	know need
moral	know love
differ men	know like
perform better	know life
enhanc	know happen


ROC Curve with AUC = 0.985
(Naive-Bayes)
Sensitivity - False Positive Rate
1 - Specificity - True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve
baseline

Misclassification

- 26-27 False Positive and False Negative

Misclassified as r/Psychology

- "U.S. Suicide Rates Are the Highest They've Been Since World War II [U.S. suicide rates] are at their highest since World War II, according to federal data and the opioid crisis, widespread social media use and high rates of stress may be among the myriad contributing factors."

- Usually posted in r/Psychology, very objective kind of headline

Misclassified as r/MentalHealth

- "In light of the very tragic Connecticut Elementary School shootings, everyone is now bringing up gun control again. What no one is talking about (and never seems to talk about) is helping to increase mental health healthcare in the country. And it's pissing me off."
- Usually posted in r/MentalHealth, very subjective posts

Conclusions and Limitations

- Naive-Bayes Multinomial is the best classifier
  - Good balance of bias and variance tradeoff
- Bag-of-words approach used
- Each word is considered independent
- Does not have context

Recommendations

- Increase number of posts from other subreddits
- Identify and input more features into the model
    - Looking at linguistic features (emojis, POS tagging, Entity recognition etc.)
    - Using sentiment analysis
    - Length of posts