Classification Of Subreddit Posts

Nicholas Lim

Outline

- Problem Statement
- Data Information
- Data Cleaning & EDA
- Modeling
- Evaluation
- Conclusion

Problem Statement

To classify posts from 2 different subreddits, /r/depression and /r/anxiety.

Motivation: Help reddit users who may be suffering from either depression or anxiety issues but not properly diagnosed to get the correct advice through posting at proper channels

About the data

- Total of 1943 rows of data
  - 998 from r/Anxiety
  - 945 from r/depression

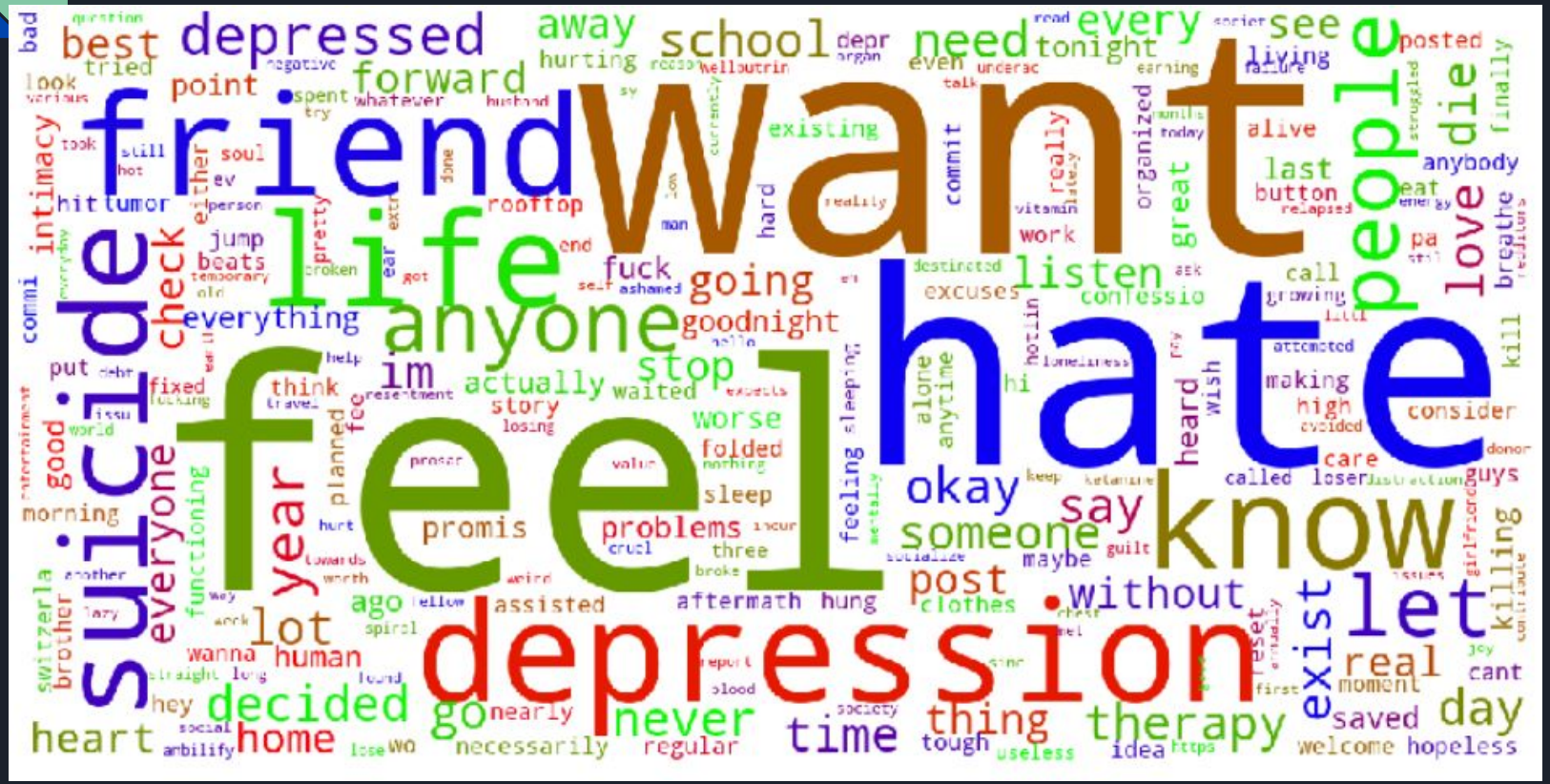
Source:
www.reddit.com/r/depression
www.reddit.com/r/anxiety

Data Cleaning

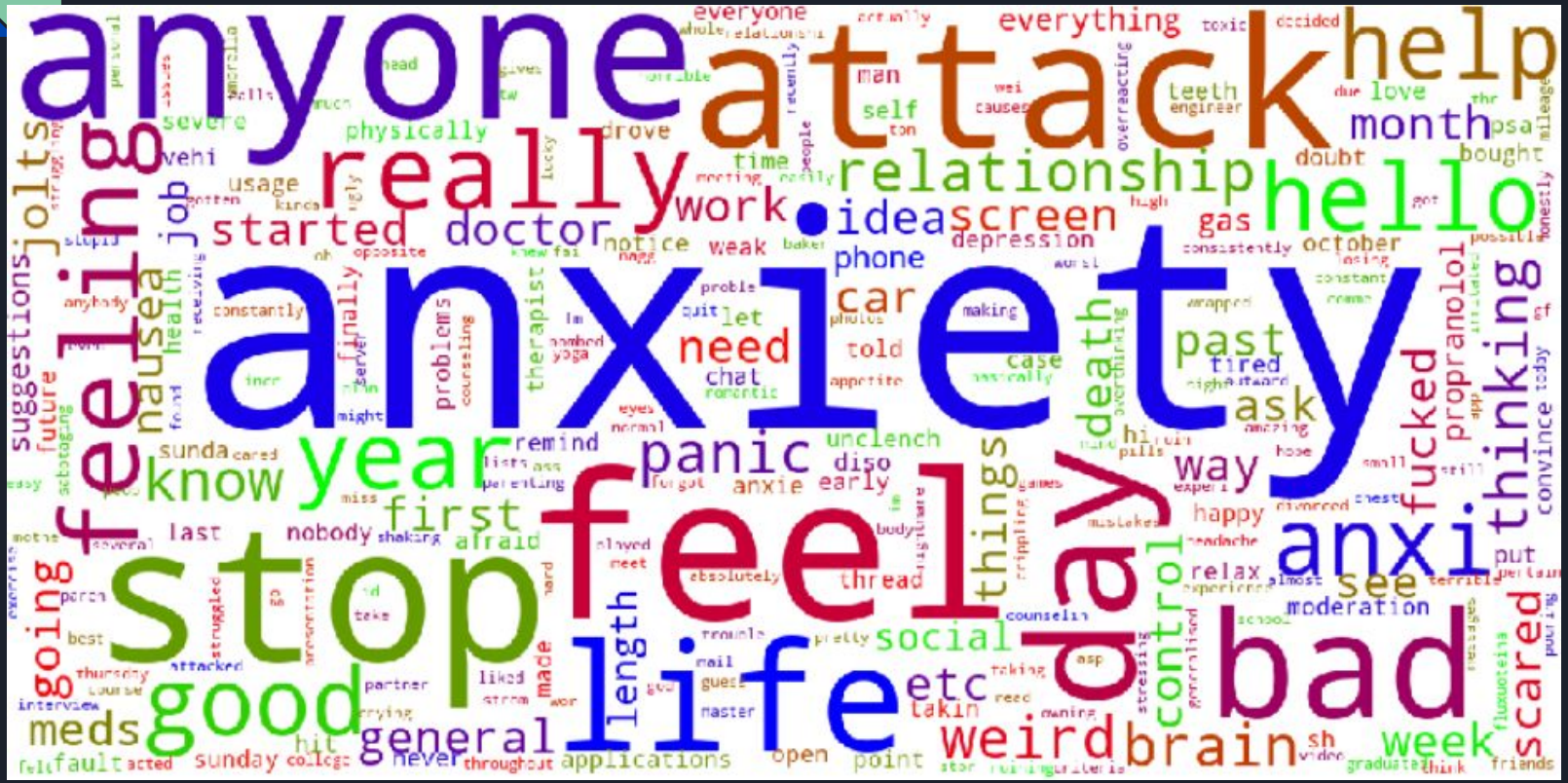
Following was done to the data

- Filled posts that only contained images/video clips with blank quotes (‘’)
- Title and content of posts were merged to one single string of text
- Removed non-letters (numerics, new line separators, punctuations)
- Stop words removed using nltk’s library of stopwords

Frequent Words in r/depression



Frequent Words in r/Anxiety



Modeling

Model	Train Score (CountVectorized)	Test Score (CountVectorized)	Train Score (TfidfVectorized)	Test Score (TfidfVectorized)
Logistic Regression	0.8277	0.821	0.8538	0.8416
K-Nearest Neighbours	0.7076	0.6749	0.7804	0.7737
Naive Bayes (Multinomial)	0.8476	0.8313	0.8428	0.8025
Decision Tree	0.7955	0.7922	0.8016	0.7984
Random Forest	0.8627	0.8374	0.86	0.8477
Extra Tree	0.8298	0.7963	0.8469	0.8004

Model Evaluation

Model	Correct /r/depression posts predicted	Correct /r/anxiety posts predicted
Logistic Regression with TfidfVectorization	197/236	212/250
Random Forest with TfidfVectorization	198/236	214/250

Conclusion

- Random Forest with TfidfVectorizer worked fairly well with an accuracy score of close to 85%, even though both subreddits were fairly similar in nature.
- Logistic Regression with TfidfVectorizer also works equally well as well with an accuracy score of 84%
- Scope can be expanded to include the following to further improve the models:
    - Include lemmatization, stemming and spell checks to have a general feel of the posts
    - Include more subreddits (eg. bipolar) in our classification model. This may be further extended to be used as an initial diagnosis of any mental issues that the user might be suffering from.
    - Tuning of parameters for random forest to get a better score. However, this requires a longer amount of time to tune to get the perfect parameters.
    - Consider either boosting or bagging to get a more optimal outcome.