Classifying Subreddits

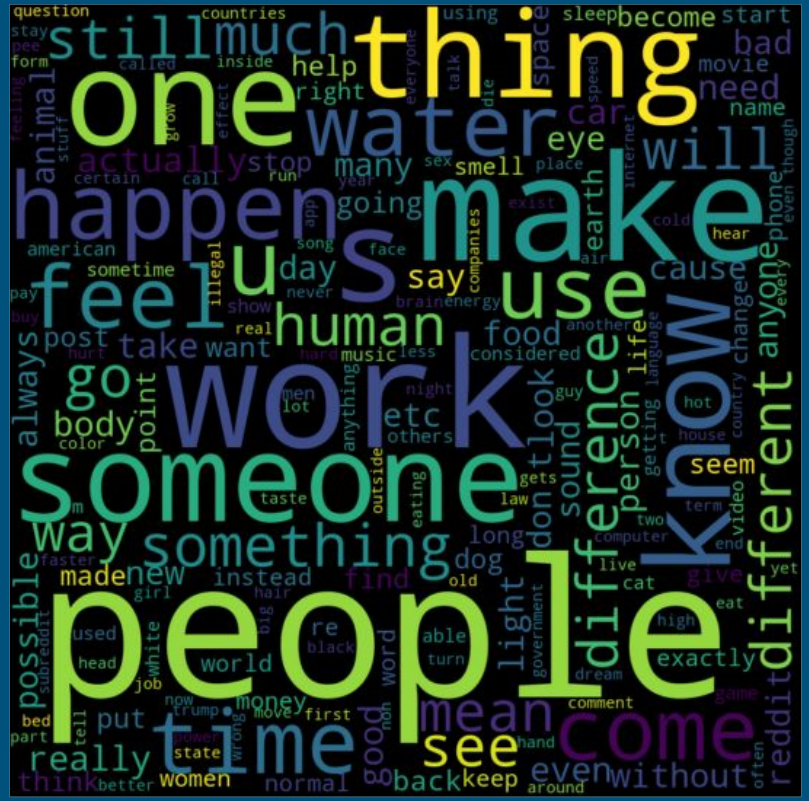
By: Robert Sarno
From: NLP Discovery

Objective:

- Reddit has one of the largest variety of question forums, entertaining or more formal. However, the challenge is that as cool as all these forums are there are a lot of similar subreddits and for new users like older generations its harder to understand what subreddit to post your question to.
- The similar subreddits I chose are r/explainlikeimfive and r/NoStupidQuestions.
- Building a Classification Model that helps new users decide what subreddit to post there questions to.

Most Frequent Words:

- The most frequent words in both subreddits.
- High Frequency Words:
    - people
    - someone
    - work
    - make
    - thing



The Initial Data used:

- Total Observations: 10000
- 5000 observations from r/explainlikeimfive
- 5000 observations from r/NoStupidQuestions
- The baseline case is:
  - 50% are r/explainlikeimfive
  - 50% are r/NoStupidQuestions

Title Frequency in r/explainlikeimfive:

Average Word Count:

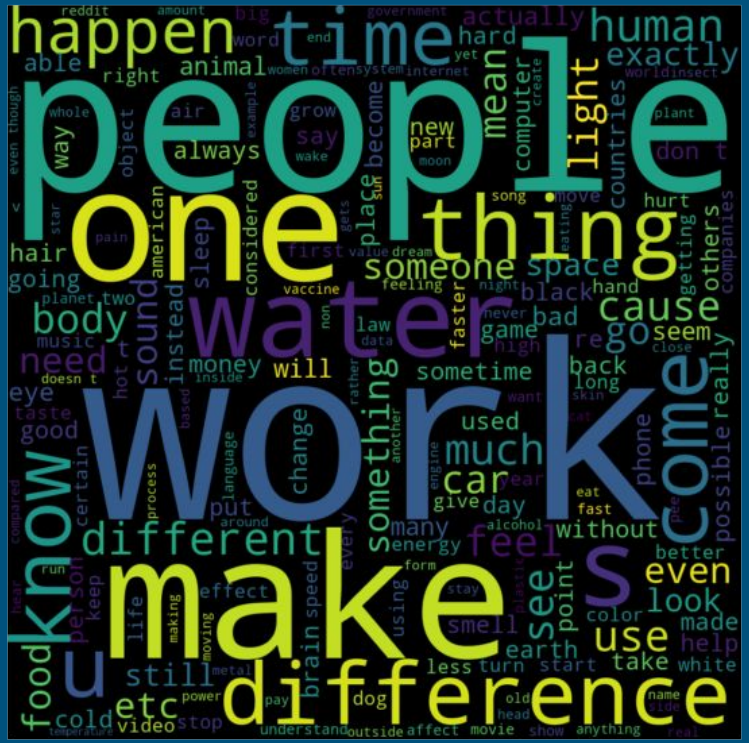
- 13 words

Average Character Count

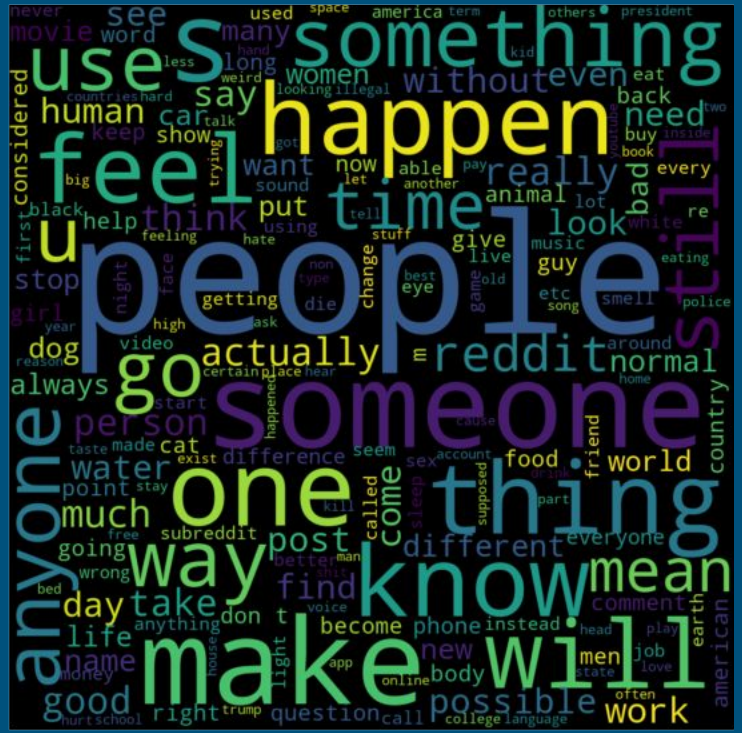
- 73 characters

Words most used:

- People, work, make



Title Frequency in r/NoStupidQuestions:



Average Word Count:

- 14 words

Average Character Count

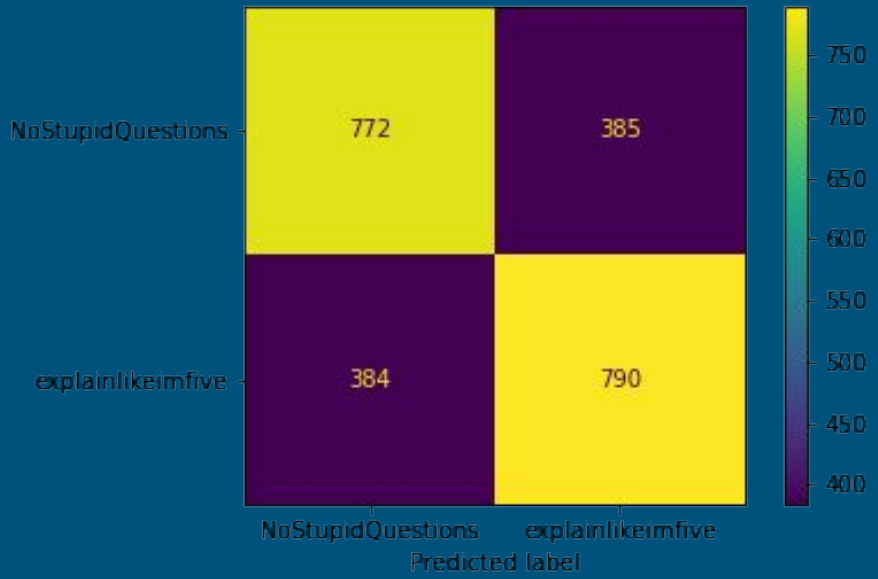
- 75 characters

Words most used:

- People, make, someone

Initial Model:

- The initial model I designed was simple CountVectorizer with Logistic Regression.
- Had an accuracy of 92% on train data and 67% on test data.
- Sensitivity of .67
- Specificity of .67
- With just using the title feature.



Improving my Classification for both Subreddits:

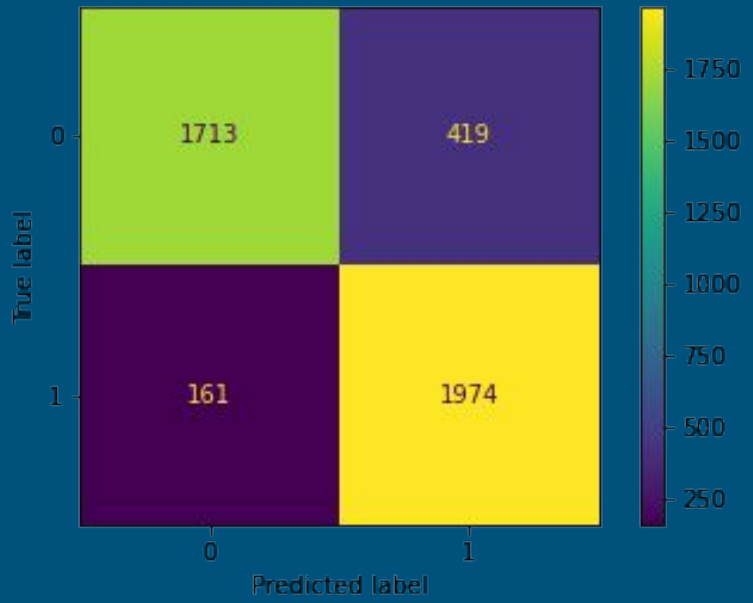
- Additional Features
  - Title character length
  - Title word count
  - NLTK Sentiment Analysis - positive, negative, neutral, compound scores
- Correcting title grammar
- Removing high frequency words in both subreddits
- Including more data

Final Model:

- The model i decided to construct incorporated a TFIDVectorizer with ADA Boosted Classifier, XGB Boosted Classifier, Bagging Classifier, Random Forest Classifier, and Extra Tree Classifier.
- I grouped all these models in to a ensemble model and gave Random Forest and Extra Tree 66% more weight than the other classifiers.
- With a lot of fine tuning and adding more training data I was able to get a good accuracy.

Final Model Performance:

- The model I was able to obtain had a training accuracy of 100% and a testing accuracy of 87%.
- With a Misclassification Error of 0.36.
- Sensitivity of .92
- Specificity of .80



Conclusion:

- In conclusion my model is great at predicting what subreddit a user should post their question to but not amazing. With a 0.86 balanced accuracy score, 0.87 f1 score, and 0.87 roc auc score there is definitely room for improvement.
- I would recommend using my model to help users understand where there posts should go especially for new users.

Thank you!

Questions