Reddit Classification Modeling

Can we classify posts from subreddits r/AskMen and r/AskWomen to their respective threads based on the Title and Body of the post?

Overview

Chose to use post title and body, no comments

Tried lemmatizing text but it actually hurt the model performance

TFIDF Vectorizer worked significantly better than CountVectorizer

Compared Logistic Regression and Random Forest, settled on Random Forest

Challenges

Scraping

Missing or no text in body

No duplicates, adjusting post time period

Pushshift.io

Overfit

High variance between train and test set, Random Forest especially (99% training sets)

Feed more data into model, 5-10% increase in accuracy

Size of Dataset

Over 50k posts in all with over 38k features

Scraping went back 2+ years

Modeling and grid searching took **very** long

Final Model

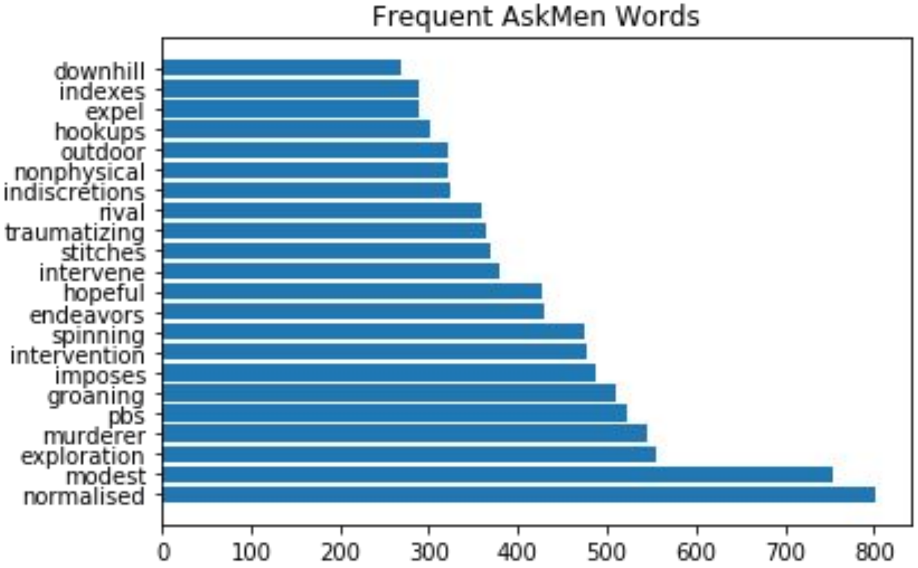
Random Forest

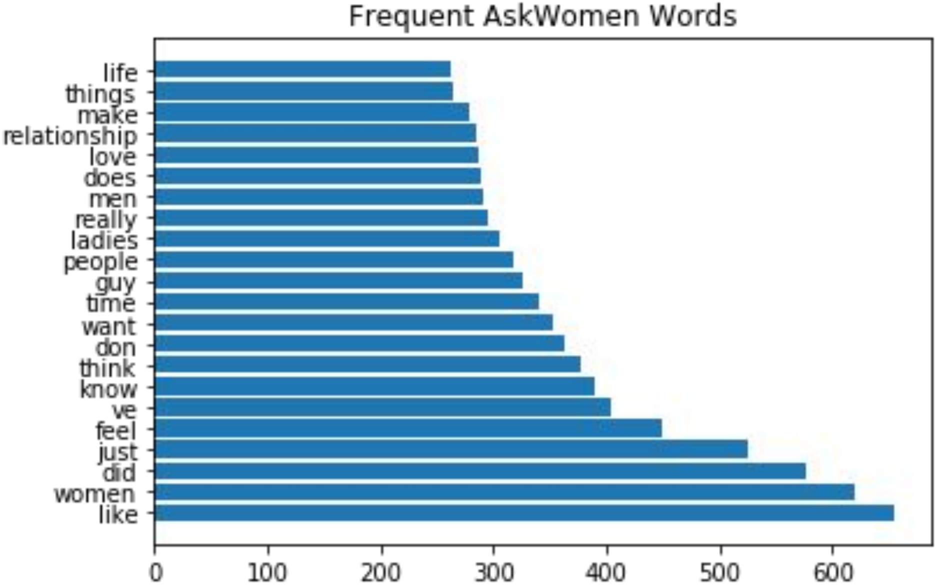
Settled on Random Forest model, n_estimators=100, max_depth=None, max_features=sqrt

Similar accuracy score to Logistic Regression of 0.75

Doesn’t rely on violated assumptions of Logistic Regression (independence of observations/features)

Weighted Words by Subreddit (TFIDF)





Classification Metrics

Accuracy: 0.75

Misclassification Rate: 0.25

Sensitivity: 0.76

Specificity: 0.73

Precision: 0.74

Conclusions and Takeaways

Conclusion:

We can classify posts from r/AskMen and r/AskWomen with 75% accuracy

Further Analysis:

Examine which words were most frequent in each thread

Rerun models using only most frequent words