Subreddit Classification: AskWomen vs AskMen

BY RACHAEL FRIEDMAN

Background

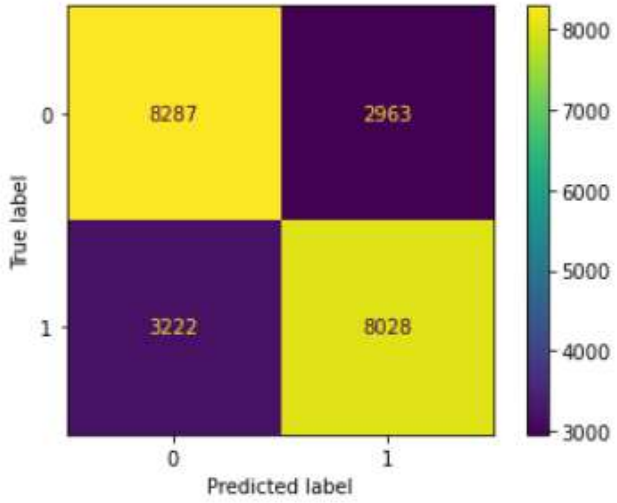
- A digital marketing company has hired me to build a model that best predicts which subreddit a post came from
- This company is interested in the **AskWomen and AskMen subreddits** because they advertise on both
- Questions to answer through research:
  - Are there differences in topics of post title on these two threads?
  - How well can models classify which subreddit a post comes from?
  - How can this information help the company market better?

Methodology

- Data collection: using Pushshift's API to collect post titles for 50,000 subreddits each
- Data cleaning and exploratory data analysis
- Modeling:
  - Pre-Processing- lemmatization, stemming, stop words, Count Vectorizer, TF-IDF Vectorizer
  - Classification models - Naïve Bayes, Random Forest Classifier, Logistic Regression, Support Vector Classification
- Model Selection:
  - How do the model's accuracy scores for training and testing groups compare to baseline and other models?
  - How much time and computing resources does the model require to run?
  - What conclusions can we make from the results?

Model Selection

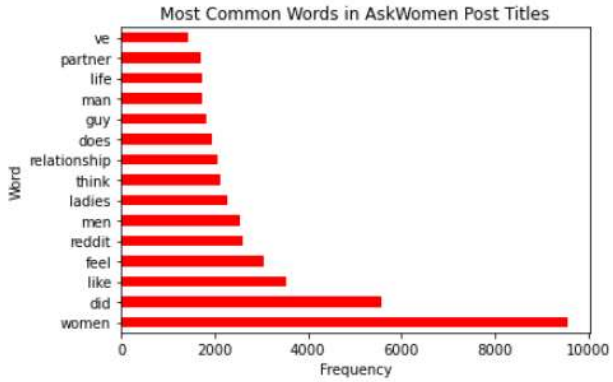
- Bernoulli Naïve Bayes is the best performing model
- Best parameters for Count Vectorizer pre-processing:
  - No stop words
  - 10,000 max features
  - Extract unigrams and bigrams
- Outperforms baseline score of 0.5
- Most consistent accuracy scores between testing and training groups
- Takes relatively less time to fit model

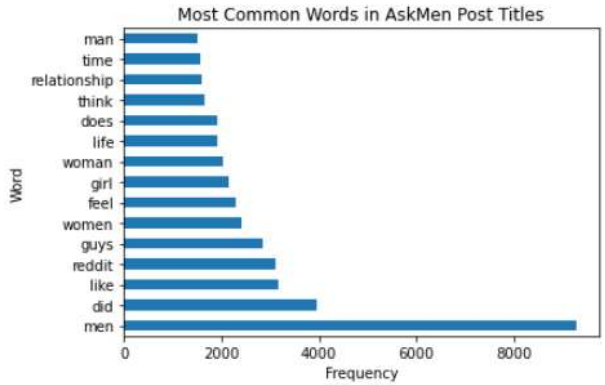



```
training score:  0.7521037037037037
testing score: 0.7251111111111112
```

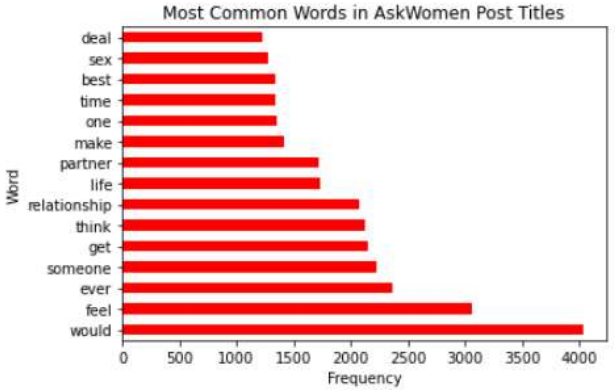
Key Findings

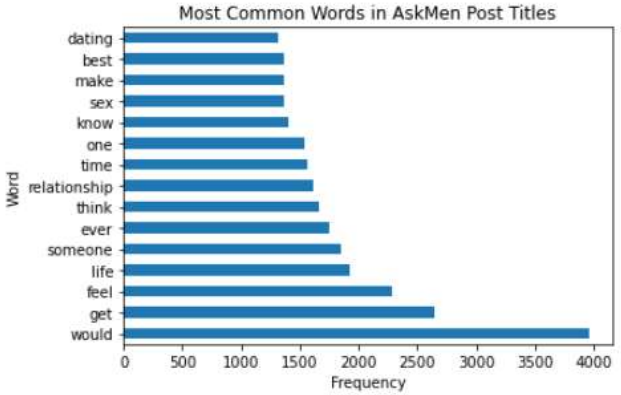
AskWomen and AskMen (English stop words excluded)





AskWomen and AskMen (English + custom stop words excluded)





Recommendations

- We hit our upper limit on the accuracy rate for classifying posts between the two subreddits and can feel confident that different methods were explored
- Topics of conversation between AskWomen and AskMen are similar
- Words used most frequently in both subreddits focus on interpersonal relationships
- Similar digital marketing may be used for both subreddits
- Pipelines for data cleaning and production model can be used to compare other subreddits for marketing purposes

Questions?

References

- Pushshift's API - https://github.com/pushshift/api
- AskWomen subreddit - https://www.reddit.com/r/AskWomen/
- AskMen subreddit - https://www.reddit.com/r/AskMen/