Predicting Subreddit by Post

GA Data Science Immersive
Sara Soueidan
April 24th, 2020

Guess the Subreddit

“Made everything bagels! Pretty stoked for a first attempt!”

“Rhubarb and apple turnovers. Made from scratch, except for the pastry, obviously.”

“I just had a fried rice revelation.”

“Help with my cheesecake.”

“What am I doing wrong? Is this me or the pan?”

“My cat is an a@@hole….”

Guess the Subreddit

BAKING "Made everything bagels! Pretty stoked for a first attempt!"

BAKING "Rhubarb and apple turnovers. Made from scratch, except for the pastry, obviously."

COOKING "I just had a fried rice revelation."

COOKING "Help with my cheesecake."

COOKING "What am I doing wrong? Is this me or the pan?"

BAKING "My cat is an a@@hole…."

The Data

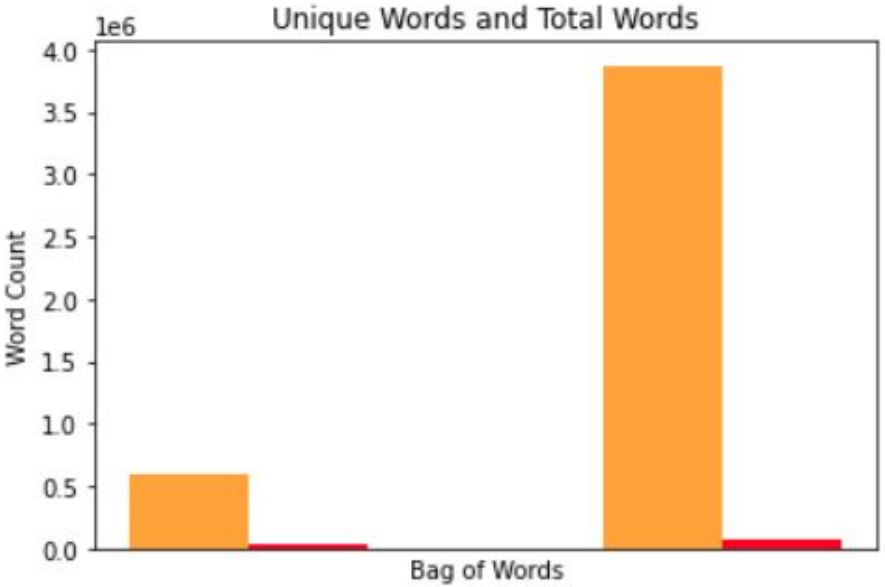
FEATURES

title
created_utc
selftext
id
media_only

CLASSIFIER

subreddit

The Evaluation



Total Words

Unique Words

Sentence Length

The Clean

Remove stopwords+

Lemmatize

Filter for urls

The Models

Logistic Regression

Training Accuracy - 93%

Testing Accuracy ~ 92%

Naive Bayes

Training Accuracy - 91%

Testing Accuracy ~ 91%

KNN

Training Accuracy - 99%

Testing Accuracy ~ 87%

Conclusions

Simple (shorter) is better

Grouping words has strong impact

Start large, tune small

Questions?

References

- https://towardsdatascience.com/another-twitter-sentiment-analysis-with-python-part-8-dimensionality-reduction-chi2-pca-c6d06fb3fcf3
- https://www.kdnuggets.com/2019/01/solve-90-nlp-problems-step-by-step-guide.html
- https://pdf.sciencedirectassets.com/278653/1-s2.0-S1877705814X00020/1-s2.0-S1877705814003750/main.pdf?
- https://towardsdatascience.com/text-classification-using-k-nearest-neighbors-46fa8a77acc5
- https://medium.com/@annabiancajones/sentiment-analysis-on-reviews-feature-extraction-and-logistic-regression-43a29635cc81
- Lessons