Classifying Reddit Posts
Beer vs. Wine

Pemika Nakata
DSI Project 3

Background

My startup has created an application that allows consumers to view more information about alcoholic beverage by hovering their phone camera over the bottle. To further improve our application, we wanted to expand the use of our machine learning algorithm and find a way to classify human-generated content, so we can provide suggestions about the drink consumers were referring to.

Problem Statement

Given that we can gather data from Reddit, our objective now is to find which classification model is best at differentiating beer subreddit posts from wine subreddit posts as these are our targeted alcohol categories.

The two classifier model that will be constructed are **Multinomial Naive Bayes model** and **Logistic Regression model.** The evaluation metric that will be used is **Accuracy** as both models need to correctly classify posts into the respective subreddits.

Data Collection

- We acquired the dataset using Reddit's API
- The two subreddits I used:
  - r/beer
    - 1634 rows, 2 columns (titles, selftext)
  - r/wine
    - 1344 rows, 2 columns (titles, selftext)

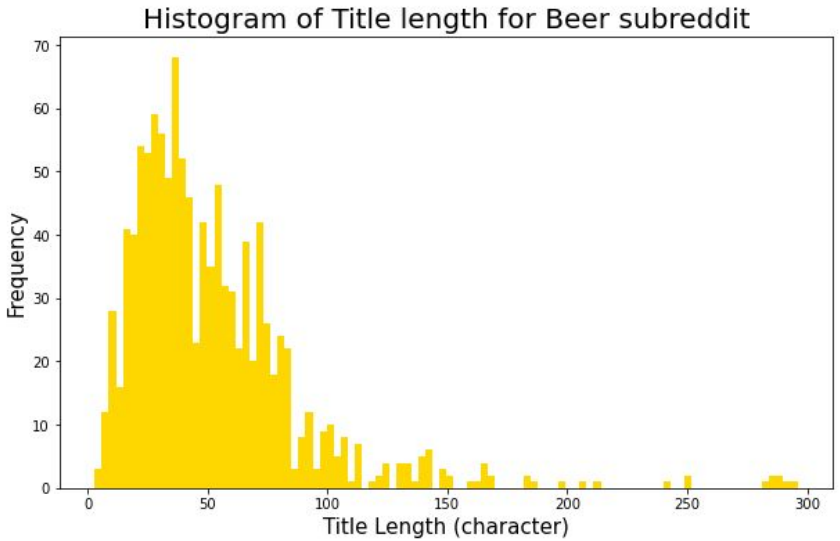
Sample Post

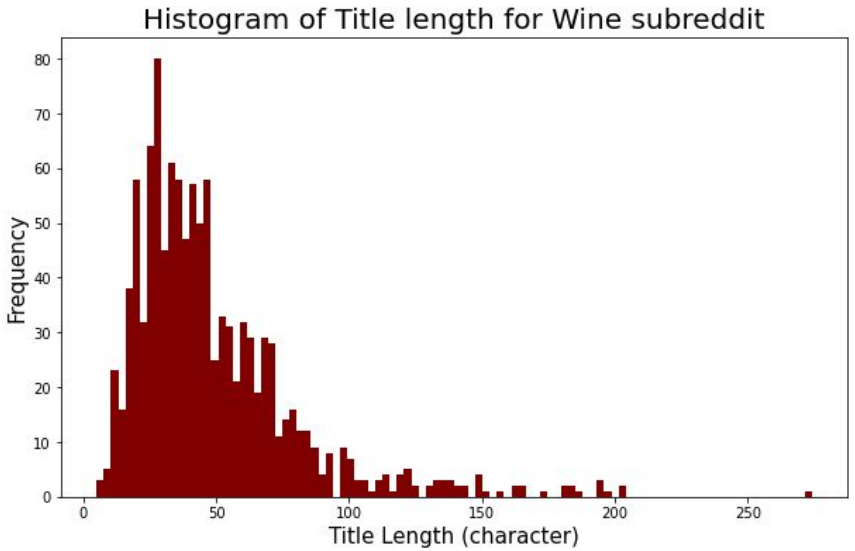
- "I've moved back to Canada and am trying to explore BC wines, but how to do that when I can't go tasting?"
- "I have a bit of a strange situation, is anybody else like this? I feel opposite to most beer drinkers, at least in North America, if I drink warm/room temp beer I feel alright, but if I drink ice cold beer, my stomach feels heavy and I feel full and like I'm going to vomit. Anybody else like me?"

Data Cleaning

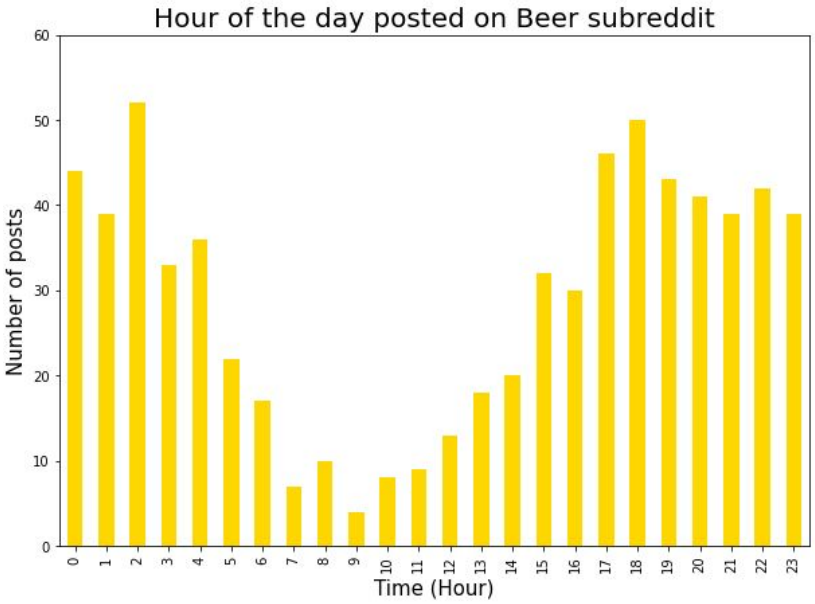
- Concatenate title and selftext into one column
- Drop duplicate posts
- Drop null values

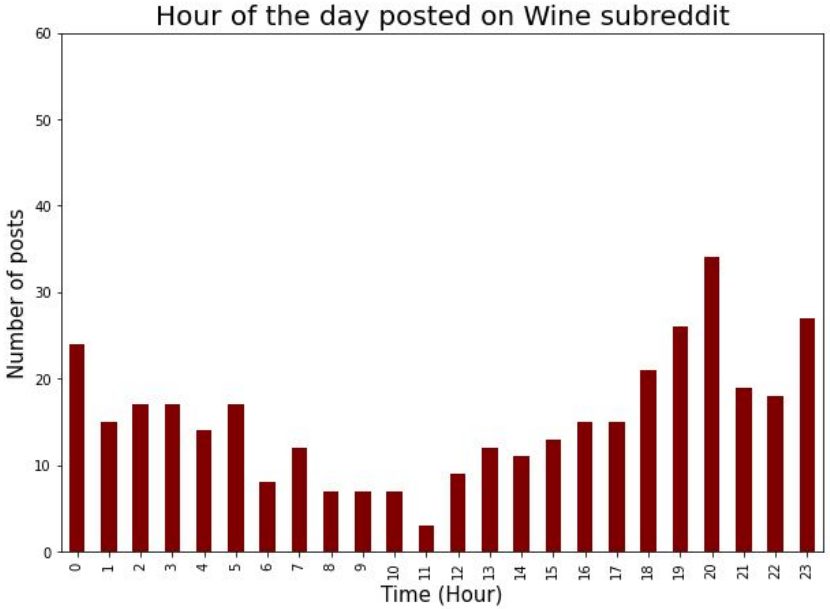
Exploratory Data Analysis - Length of Title





Exploratory Data Analysis - Posting time

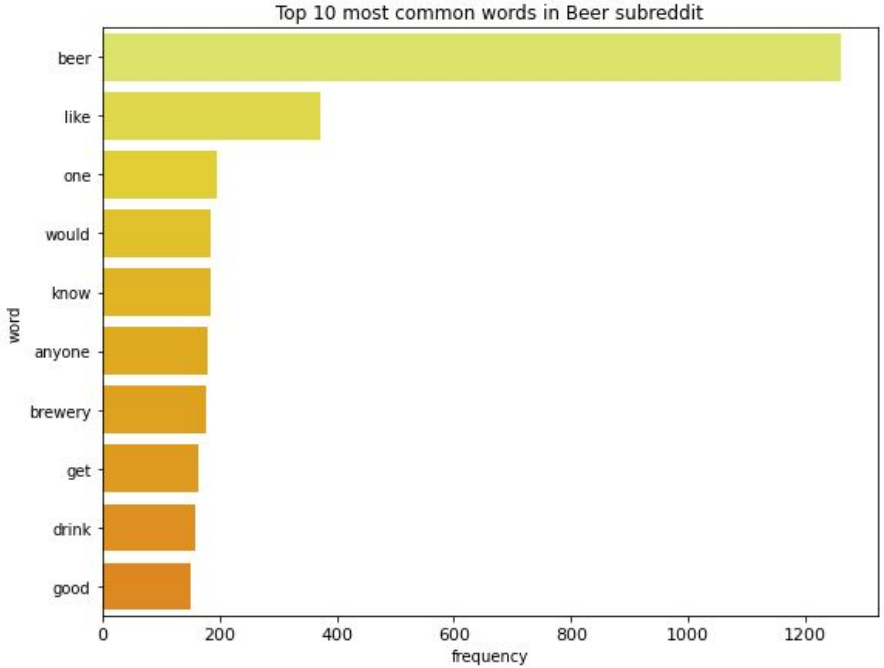


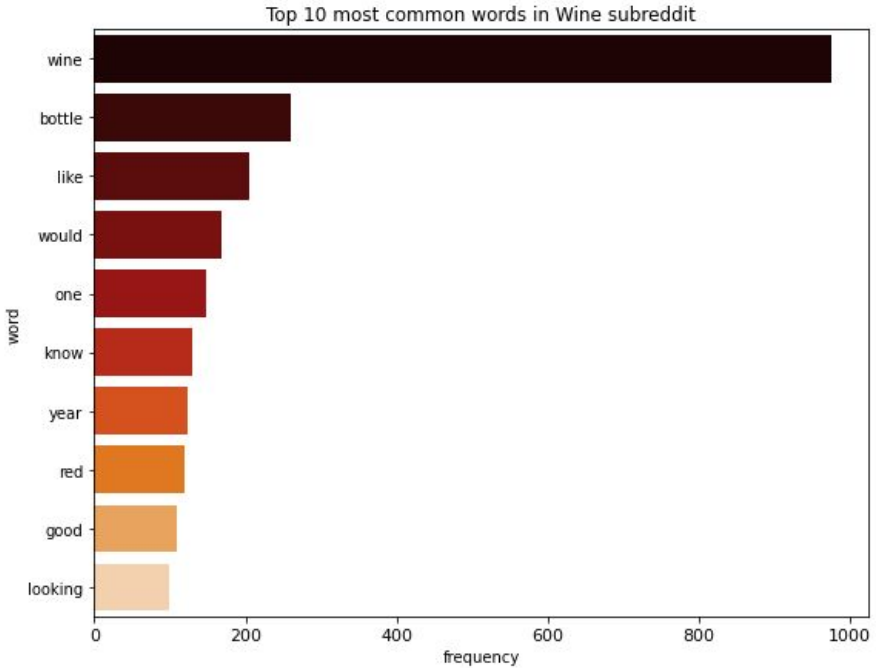


Preprocessing

- Remove
  - HTML, http, https
  - commonly used domains
  - slash, hyphen, hyphenated numerals
  - new-line, tabs, carriage returns, and apostrophes that are followed by a space character
  - Non-letters
  - Stopword - “english”
- Convert
  - Lowercase
- Lemmatize

Exploratory Data Analysis - Common words





Modeling

2 Models

- Logistic Regression and Multinomial Naive Bayes

Each model was run 3 times:

- Count Vectorizer (Default):
  - Countvectorizer
  - Stopwords = english

- Count Vectorizer & TFIDF Vectorizer (Tuned) :
  - Optimized GridSearch
  - Testing different hyperparameters:
    - Logistic Regression Model
      - Penalty: [‘l1’,’l2’]
      - Solver: ['liblinear','sag','saga']
      - C': np.logspace(-5,0,100)
    - Multinomial Naive Bayes Model
      - Pngram_range: [(1,1),(2,2)]
      - max_features: [8000, 9000, 10000]
      - min_df: [1,2]
      - max_df: [.9, .95]
      - alpha:np.linspace(0.5, 1.5, 2)

Evaluation

Dataset	Type of vectorizer	Type of model	Accuracy Score
Train	Count Vectorizer (default)	Logistic Regresion	0.993
Test	Count Vectorizer (default)	Logistic Regresion	0.901
Train	Count Vectorizer (default)	Multinomial Naive Bayes	0.976
Test	**Count Vectorizer (default)**	**Multinomial Naive Bayes**	**0.943**
Train	Count Vectorizer (tuned)	Logistic Regresion	0.966
Test	Count Vectorizer (tuned)	Logistic Regresion	0.914
Train	Count Vectorizer (tuned)	Multinomial Naive Bayes	0.979
Test	Count Vectorizer (tuned)	Multinomial Naive Bayes	0.941
Train	TFIDF Vectorizer (tuned)	Logistic Regresion	0.987
Test	TFIDF Vectorizer (tuned)	Logistic Regresion	0.911
Train	TFIDF Vectorizer (tuned)	Multinomial Naive Bayes	0.972
Test	TFIDF Vectorizer (tuned)	Multinomial Naive Bayes	0.931

Conclusion

With our objective of finding the best classification model to differentiate wine subreddit posts to beer subreddit posts, we found that our default Multinomial Naive Bayes model with count vectorizer gives the best accuracy score of 94.3%.

I concluded that the best model for my Startup to proceed for further development would be Multinomial Naive Bayes model with default hyperparameters and count vectorizer.

Recommendations

Moving forward, steps that I can take to tune and tweak my model for better accuracy score would be to:

- Feature engineer more features such as characters or word lengths
- Explore other features that could play significant role in predicting outcome such as post comments
- Try using other models such as Decision Trees, Random Forests
- Try adding the keyword "wine" and "beer" to the list of stopwords to see how well the model can classify without these key predictors
- Create my own list of stop words such as "bottle", "anyone"
- Explore more evaluation metrics apart from accuracy score such as ROC AUC score

Thank you!


BLUE MOON
BELGIAN WHITE
Budweiser
Corona Extra
ESPERADO
Estrella Damm
Heineken
Modelo
BIRRA MORETTI
PERONI
San Miguel
SOL
STELLA ARTOIS


Dão
ASSYRTIKO