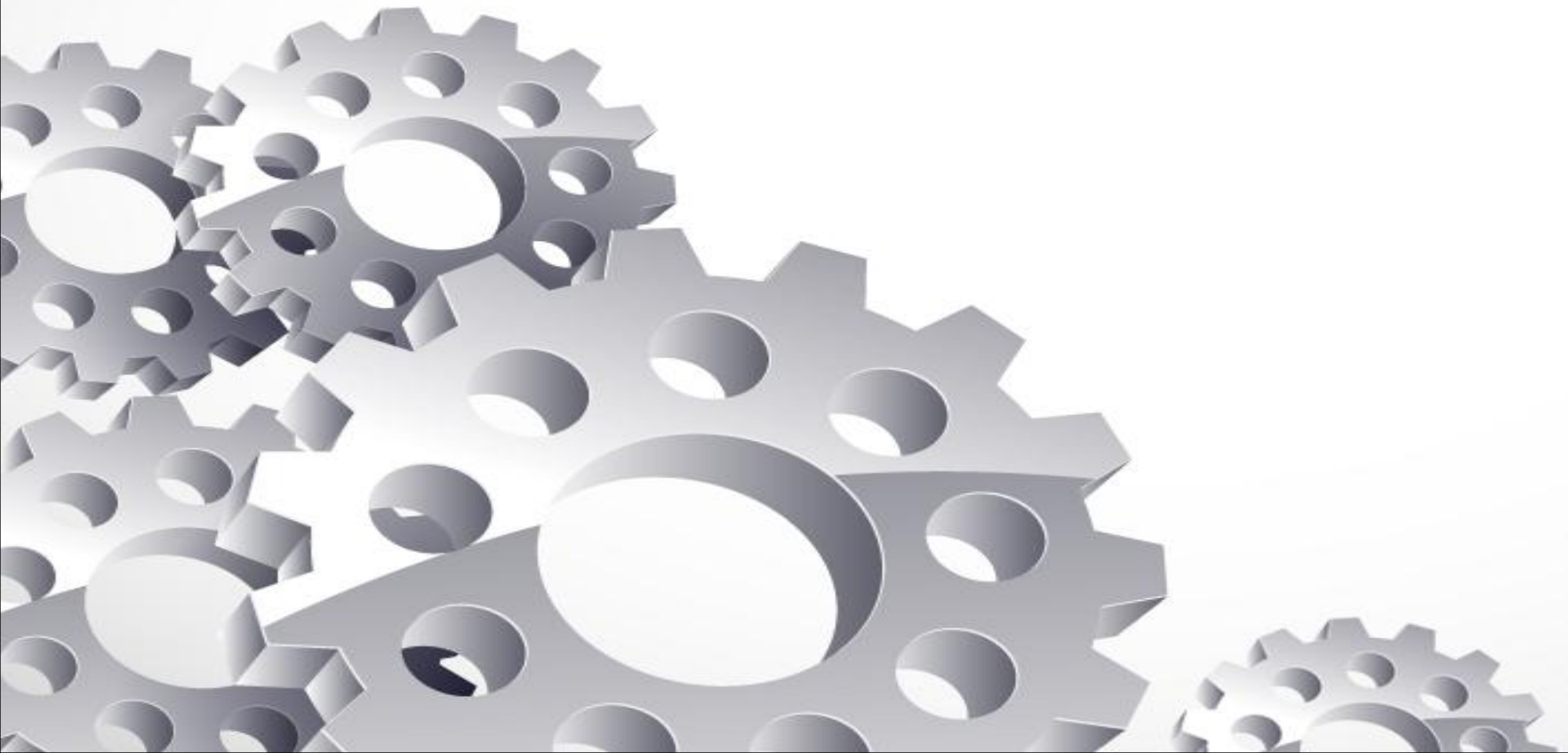
DSI- Project 3

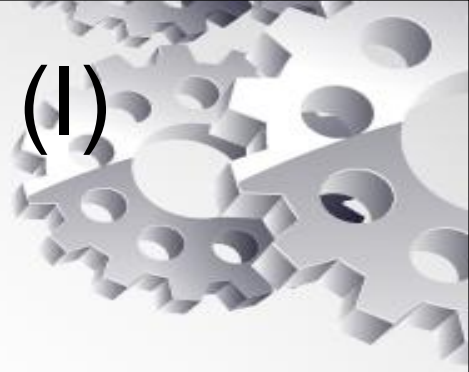
Ivy Chan



Problem Statement

- Documention classification in the past used to be done manually and requires high labour hour.
- As a marketer who is intereted in identify what are the keywords that users in the Reddit community use when discussing on 2 topics: coffee and restaurant so that the keywords can be later used in content creation.
- In this project, We will train a classifier to take in text input and then classify to which group the text belongs to in order to reduce labour hour and increase classification accuracy.

Data Collection: Scrap Data Using Reddit API (I)



- Data of 2 subreddit: Coffee & Restaurant were scrapped using Reddit API.
- After getting our url in json, request is sent to Reddit using a non-deafault value for the key 'User-agent' to prevent Reddit from shutting our script from accessing its API.
- A status code of 200 shows that the request was received and understood ad is being processed.
- The Json file received is a nested dictionary, access to key 'data' and then key'children' to get the "content".

Data Collection: Scrap Data Using Reddit API (II)

- As each Reddit request will only return 25 posts, a for loop is used to hit Reddit API repeatedly to collect a minimum of 1000 posts from each subreddit. The for loop will have an empty list created to store Arandom sleep duration is included at the end of loop to give anatural break in between request.
- The scraped data is saved in a Pandas Dataframe and exort as csv file.

Subreddit Topic Selected:

- **Coffee**
  - 1245 rows,99 columns
  - Duplicate Post: 288 (23% of the datasets)
  - Selftext Null Value: 53 (4% of the removed duplicate datasets)
  - **Observations: 960**
- **Restaurant**
  - 1228 rows, 104 columns
  - Duplicate Post: 764 (62% of the datasets)
  - Selftext Null Value: 176 (38% of the removed duplicate datasets)
  - **Observation: 464**

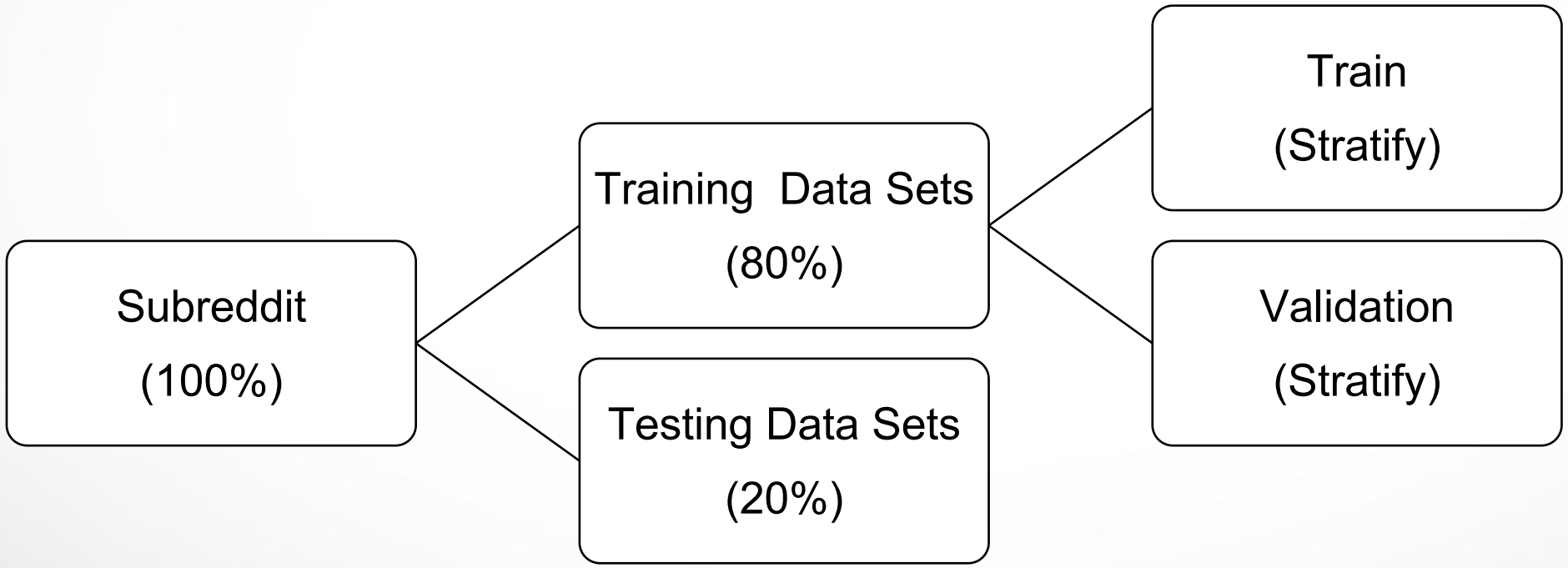
Data Cleaning:

- Remove Duplicate Posts
- Fill in null value with Title
- Split both subreddit into 2 parts:
  - Training Datasets- (80%)- For Model Development
    - Coffee- 768
    - restaurant-371
  - Testing Datasets- (20%)- For Performance Evaluation
    - Coffee- 192
    - restaurant - 93
- Map and Label Subreddit: Coffee:0, restaurant:1

Preprocessing:

- Create a function to:
  - Remove html
  - Remove http & www
  - Remove special elements ie. /r/|\n|<|>
  - Remove punctuation, change to lower case
  - Remove stop word
  - Remove Emoji
  - Return as a clean string of words and store in 2 lists
    - Traning_self_text
    - Testing_self_text

Train-Test_Split



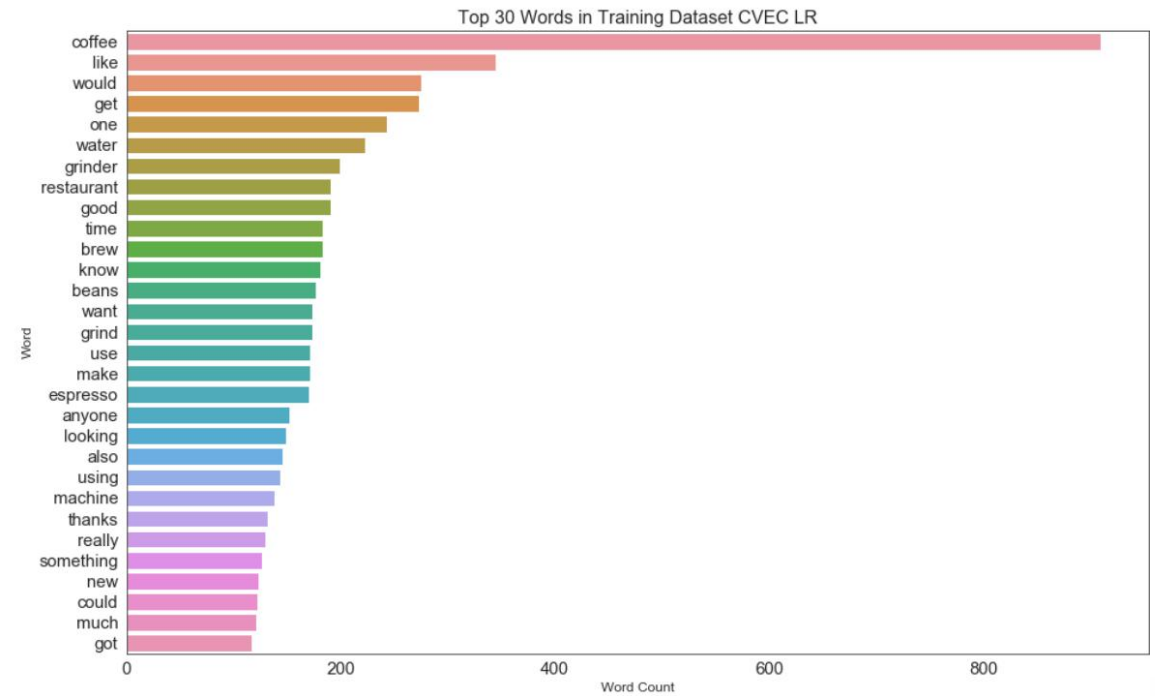
WordCloud

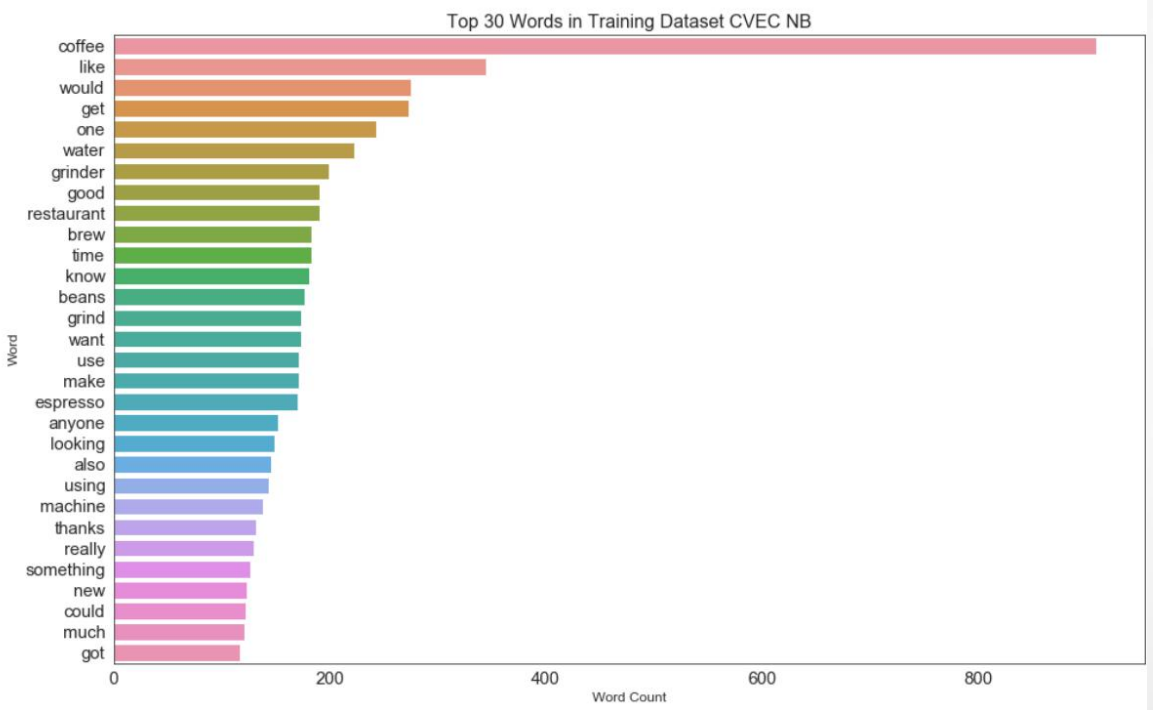
Subreddit Coffee

Subreddit Restaurant

Top 30 Words & Counts for 2 Model

The Sequence Words Count in descending for is slightly different for both model.





Models & Scores:

	Train Score	Validation Score
Baseline Accuracy	Subreddit 0: 0.6742 Subreddit 1: 0.3257	
CountVectorizer Logistic Regression	0.9988	0.9684
CountVectorizer Logistic Regression with Best Parameter	0.9976	0.9508
CountVectorizer Multilinear Naive Bayes	0.9906	0.9508
CountVectorizer Multilinear Naive Bayes with Best Parameter	0.9871	0.9508
Using Best Model to do Prediction with Testing Data	Score on Testing Data: 0.9473	

Conclusion

- The best model chosen in this project is CountVectorizer Multinomial Naive Bayes tuned with hyperparameter withe the final prediction score of 94.73%.
- In this project, the Key Merics we are looking at Accuracy as any misclassification in the subreddit post will result in an ambiguos state as to which group of subject this word belongs to when creating content.