Sentiment Analysis on Customer Reviews

Comp 472

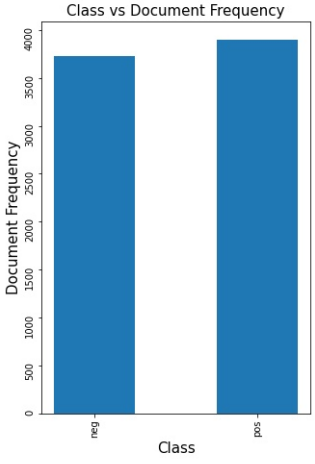
Nadia Sheikh

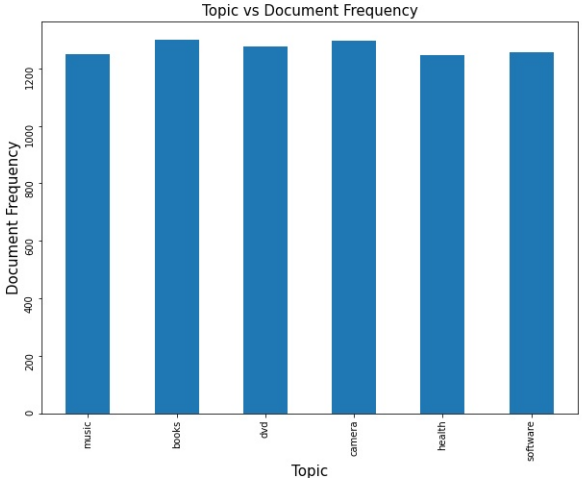
February 22, 2021

Overview

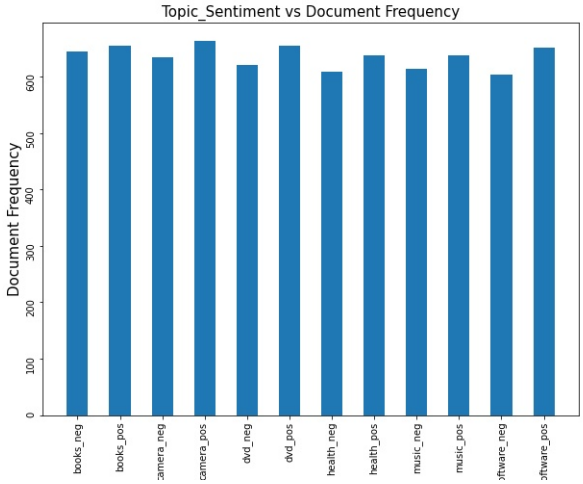
1. Exploratory Data Analysis
2. Preprocessing
3. Naive Bayes
4. Decision Tree
5. Analysis

Balanced Data Set?

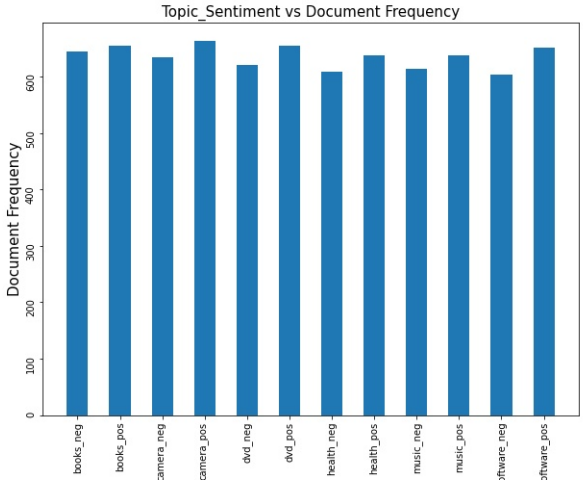




Balanced Data Set?



Balanced Data Set?

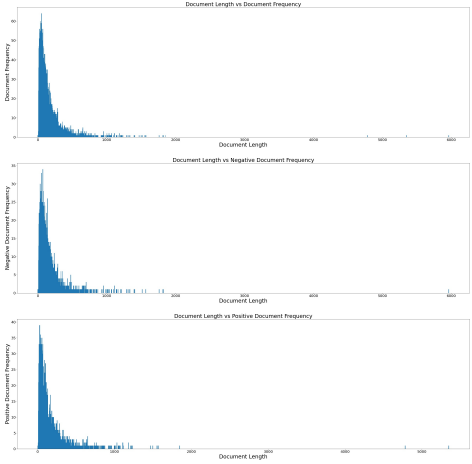


Document Length

Statistics	All	Negative Documents	Positive Documents
Smallest	0	0	3
Largest	5964	5964	5349
Median	97	102	92
Mean	149	152	146
Mode	55	164	73
Standard Deviation	191	191	27

Table: Performance metrics after hyper-parameter turning

Naive Bayes: Distribution of Classes



Vocabulary

Tuning Parameters

Metric	Count
Vocabulary Size	45889
Negative Document Vocabulary Size	29877
Positive Document Vocabulary Size	31068
Common Vocabulary	15056
Negative Document Unique Vocabulary	14821
Positive Document Unique Vocabulary	16012

Table: Performance metrics after hyper-parameter turning

Preprocessing Steps

- Data Cleaning
- Split Data
- Feature Extraction
  - Count Vectorization
  - TFIDF Vectorization
- Label Encoding

Naive Bayes without Hyper-parameter Tuning

Metrics	Count Vectorization	TFIDF Vectorization
Precision	0.8021857923497268	0.8005982053838484
Recall	0.7734457323498419	0.8461538461538461
F1	0.7875536480686696	0.8227459016393442
Accuracy	0.7923439958049292	0.8185631882538018

Table: Performance metrics prior to hyper-parameter turning

Naive Bayes after Hyperparameter Tuning

Tuning Parameters

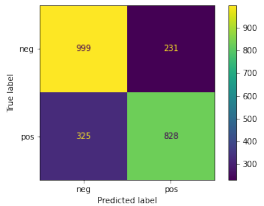
- Parameter: Alpha
- Method: RandomizedSearchCV
- Space: uniform(0,1)
- Iterations: 50

Metrics	**Count Vectorization** alpha(0.832619845547938)	**TFIDF Vectorization** alpha(0.359507900573786)
Precision	0.9819819819819819	0.9891774891774892
Recall	0.9188619599578504	0.9631190727081138
F1	0.9493739793140991	0.9759743726641752
Accuracy	0.951232302045097	0.9764027267960147

Table: Performance metrics after hyper-parameter turning

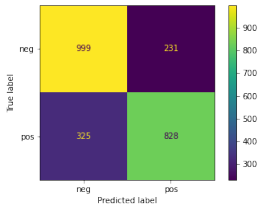
Naive Bayes(Counter Vector): Test Results

Metrics	Count Vectorization
Precision	0.7818696883852692
Recall	0.7181266261925412
F1	0.7486437613019891
Accuracy	0.7666806546370122



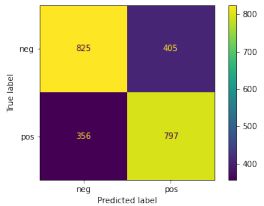
Naive Bayes (TFIDF Vector): Test Results

Metrics	TFIDF Vectorization
Precision	0.8083832335329342
Recall	0.7025151777970512
F1	0.7517401392111369
Accuracy	0.775493075954679

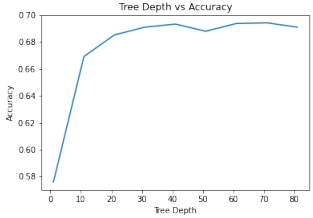


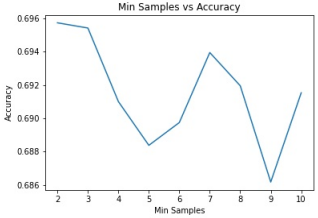
Decision Tree (Count Vector): Test Results

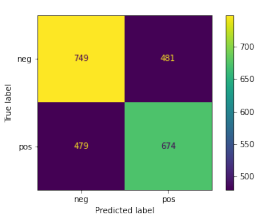
Metrics	Count Vectorization
Precision	0.6630615640599001
Recall	0.6912402428447528
F1	0.6768577494692145
Accuracy	0.6806546370121695



Improving Decision Tree Performance

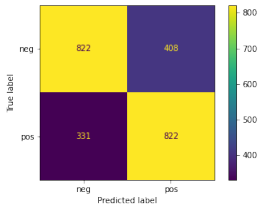




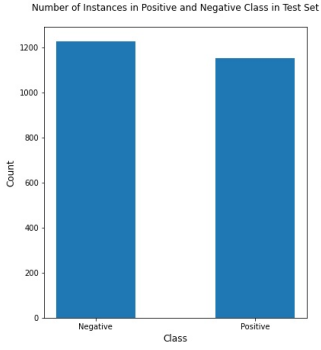


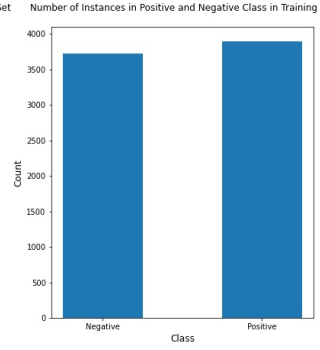
Improved Decision Tree (Count Vector): Test Results

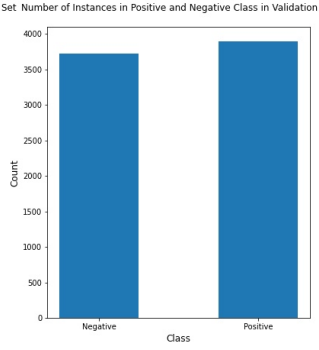
Metrics	Count Vectorization
Precision	0.6682926829268293
Recall	0.7129228100607112
F1	0.6898866974402014
Accuracy	0.6898866974402015



Naive Bayes: Analysis



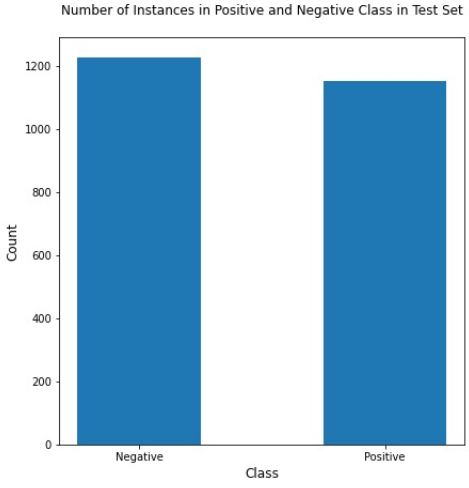


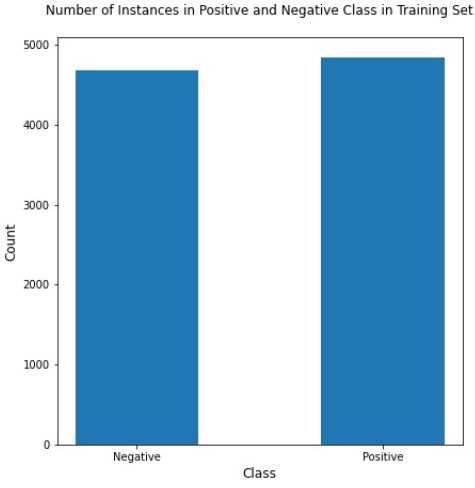


Naive Bayes: Analysis

Refer to nb_incorrect_10.txt
Refer to incorrect_intersection.png
Refer to correct_intersection.png

Decision Tree: Analysis





Decision Tree: Analysis