CSSM 502: ADVANCED DATA ANALYSIS WITH PYTHON

DAVID CARLSON

FINAL PROJECT

19.01.2022

EZGİ AYDIN

0078635

E. Aydın
19.01.2022
CSSM 502

TABLE OF CONTENTS:

E. Aydın
19.01.2022
CSSM 502

1.INTRODUCTION

This present report is the outcome of the sentiment analysis performed over "Women's Clothing E-Commerce Reviews" dataset, as a final/term project of CSSM 502: Advanced Data Analysis with Python course. Mainly, the report first covers an exploratory data analysis and then continues with machine learning algorithm training.

I have chosen sentiment analysis as my project because a) While I was in university, I did work as a Marketing Insight Analyst for Ebrandvalue company, which provides advisory over "brand value" based on social media sentiment analysis. However, since I was part of marketing team and I had zero coding experience back then, I did not have chance to see/experience the kitchen of this analysis. Even though what I performed in this project is base level, I wanted to try and set this as a starting point for many to come. b) I have made some progress over the years, and I want to continue with that by learning and experiencing new things and topics. NLP is an area that I have great interest in and motivation to learn deeper about it.

Before continuing with the report, I just want to acknowledge that, this was the best class that I took this semester by far since we were treated as adults and given assignments were on spot to boost our skills in the area. Therefore, I want to take this opportunity to thank Professor Carlson for being progressive, understanding, broadminded, encouraging us to research and so many more. For working students, that means a lot.

This report consists of 10 parts, as presented in the table of contents. First, I will be starting with the exploratory data analysis to familiarize with the data set. Later, I will continue with data processing and machine learning algorithms.

2. SENTIMENT ANALYSIS

According to Towards Data Science website, sentiment analysis is "contextual mining of text which identifies and extracts subjective information in source material and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations.[1]" In other

[1] https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

words, sentiment analysis provides us with an overview of customer’ s opinion of subject matter. In this report’ s dataset, this is the positive and negative reviews of customers over women clothing.

3.GOAL DEFINITION AND DATASET

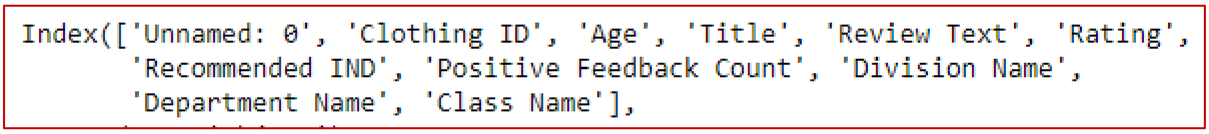
My goals in this project are:

1.To perform a comprehensive exploratory data analysis,

2. Observe whether customer reviews predict customer recommendation,

3.Train and test 4 different models and compare their performance on predicting positive/negative sentiments.

The dataset that I used is “Women Clothing E-Commerce Reviews[2]” gathered from Kaggle. It is a public dataset including 23000 customer reviews and ratings, with some additional demographics. Because this is real commercial data, it has been anonymized, and references to the company in the review text and body have been replaced with "retailer” . The data set includes following columns:



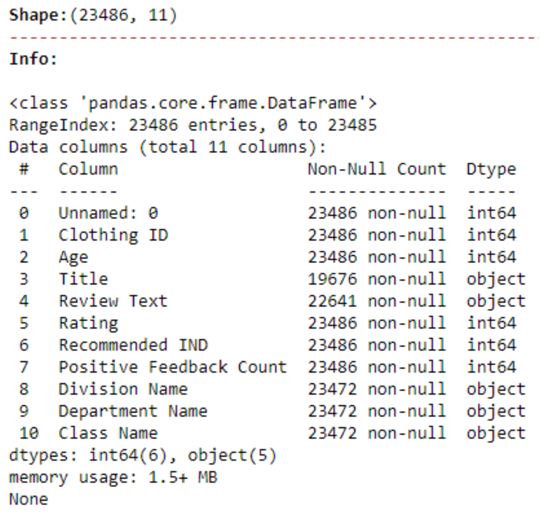
```
Index(['Unnamed: 0', 'Clothing ID', 'Age', 'Title', 'Review Text', 'Rating',
       'Recommended IND', 'Positive Feedback Count', 'Division Name',
       'Department Name', 'Class Name'],
```

- Clothing ID: Integer Categorical variable that refers to the specific piece being reviewed.
- Age: Positive Integer variable of the reviewer’ s age.
- Title: String variable for the title of the review.
- Review Text: String variable for the review body.
- Rating: Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst to 5 Best.
- Recommended IND: Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
- Positive Feedback Count: Positive Integer documenting the number of other customers who found this review positive.
- 8) Division Name: Categorical name of the product high level division.
- 9) Department Name: Categorical name of the product department name.
- 10) Class Name: Categorical name of the product class name.

[2] https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

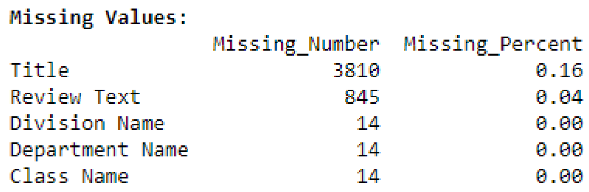
4.EXPLORATORY DATA ANALYSIS

I first performed an exploratory data analysis, which includes overviewing the dataset closely by each column and understanding it in several different aspects. Detailed work can ben found on py. File. Some analysis results are as follow:

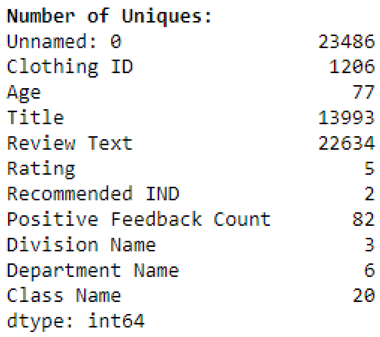


```
Shape:(23486, 11)
--------------------------------------------------------
Info:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23486 entries, 0 to 23485
Data columns (total 11 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Unnamed: 0               23486 non-null  int64
 1   Clothing ID              23486 non-null  int64
 2   Age                      23486 non-null  int64
 3   Title                    19676 non-null  object
 4   Review Text              22641 non-null  object
 5   Rating                   23486 non-null  int64
 6   Recommended IND          23486 non-null  int64
 7   Positive Feedback Count  23486 non-null  int64
 8   Division Name            23472 non-null  object
 9   Department Name          23472 non-null  object
 10  Class Name               23472 non-null  object
dtypes: int64(6), object(5)
memory usage: 1.5+ MB
None
```

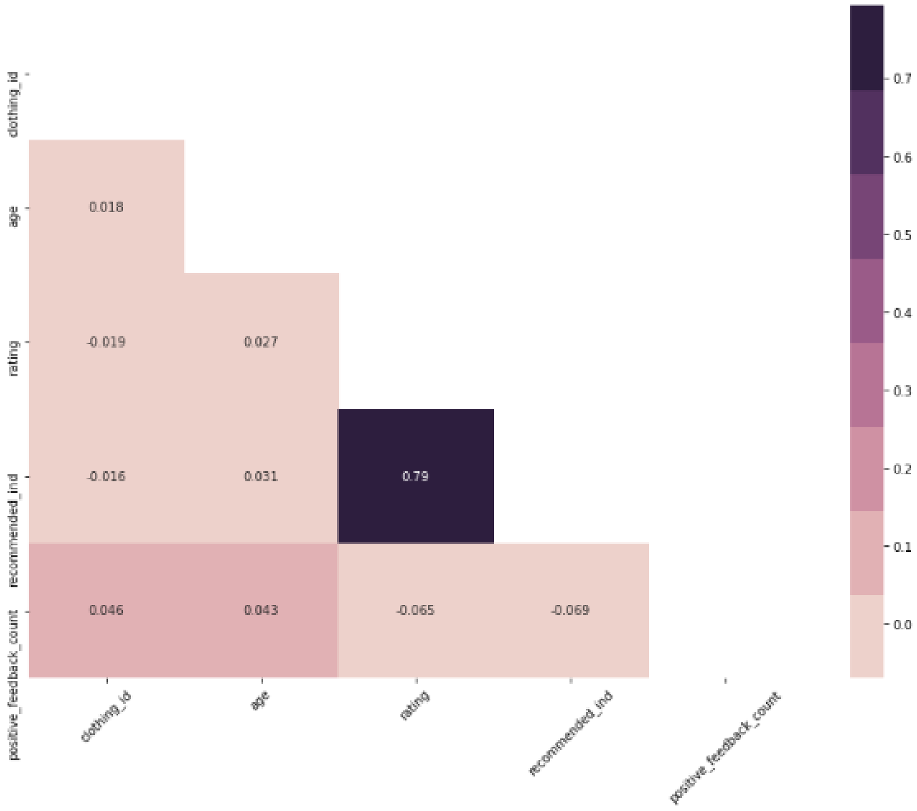


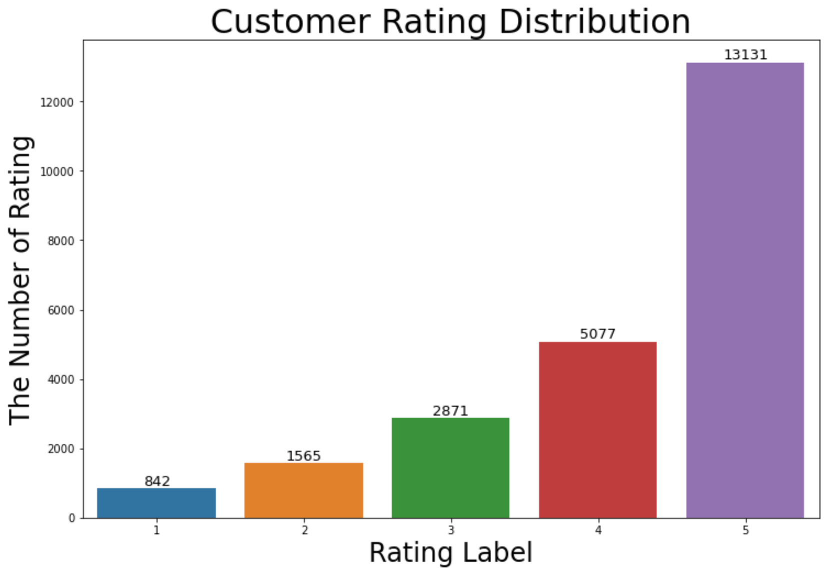
```
Missing Values:
                 Missing_Number  Missing_Percent
Title                      3810             0.16
Review Text                 845             0.04
Division Name                14             0.00
Department Name              14             0.00
Class Name                   14             0.00
```

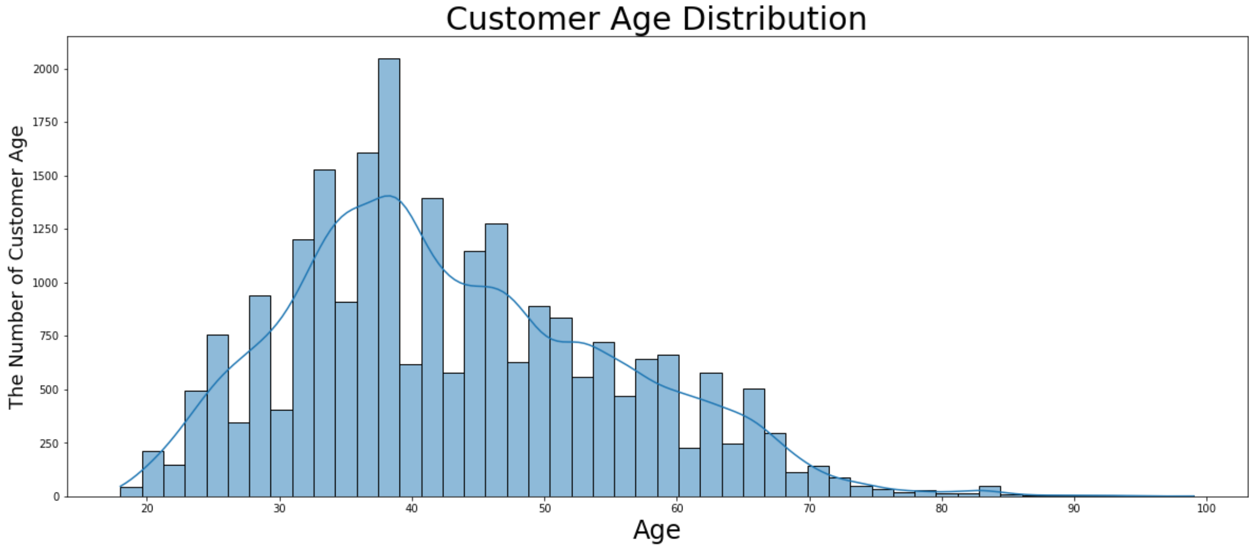


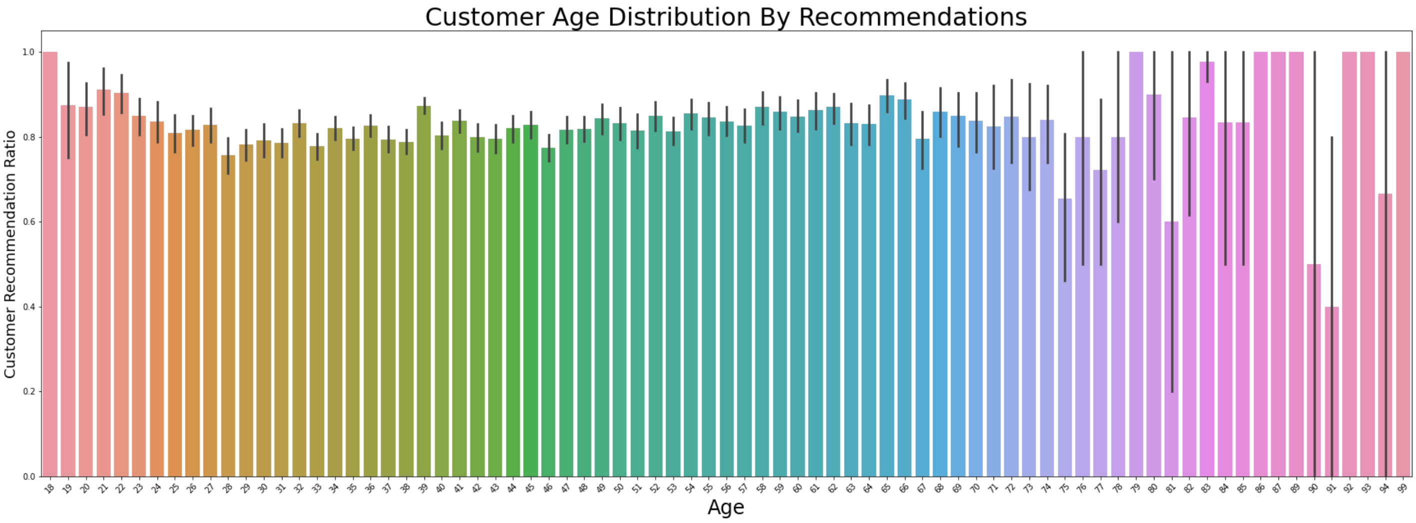
```
Number of Uniques:
Unnamed: 0                 23486
Clothing ID                 1206
Age                           77
Title                      13993
Review Text                22634
Rating                         5
Recommended IND                2
Positive Feedback Count       82
Division Name                  3
Department Name                6
Class Name                    20
dtype: int64
```

While observing the dataset, I looked for correlation between columns and the analysis resulted 0.79 correlation between Rating and Recommended IND columns. Additional to correlation heatmap, I will be sharing other exploratory data analysis results here with visualized versions.

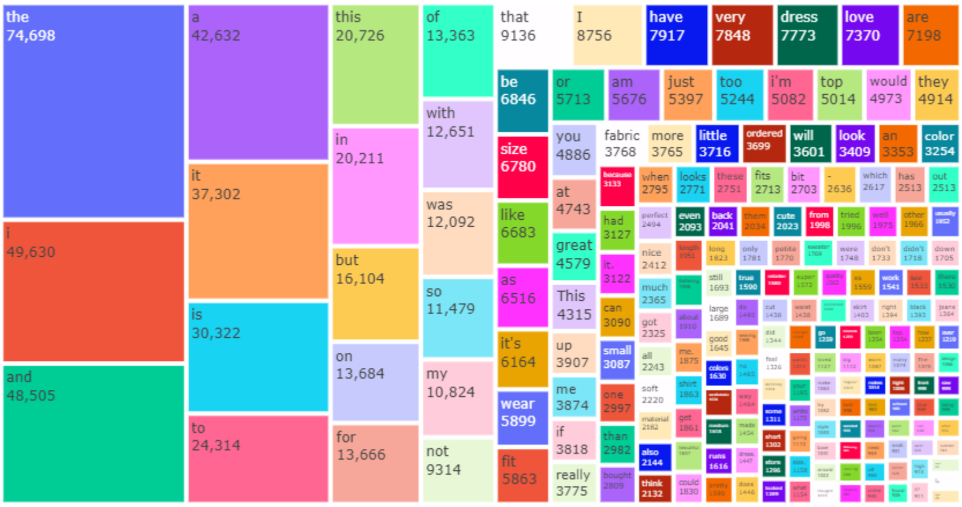



Customer Rating Distribution
The Number of Rating
Rating Label
0
2000
4000
6000
8000
10000
12000
842
1565
2871
5077
13131
1
2
3
4
5


Customer Age Distribution
The Number of Customer Age
Age
0
250
500
750
1000
1250
1500
1750
2000
20
30
40
50
60
70
80
90
100


Customer Age Distribution By Recommendations
Customer Recommendation Ratio
Age
0.0
0.2
0.4
0.6
0.8
1.0
18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 89 90 91 92 93 94 99

Top Frequent 200 Words in the Dataset (Before Cleaning)



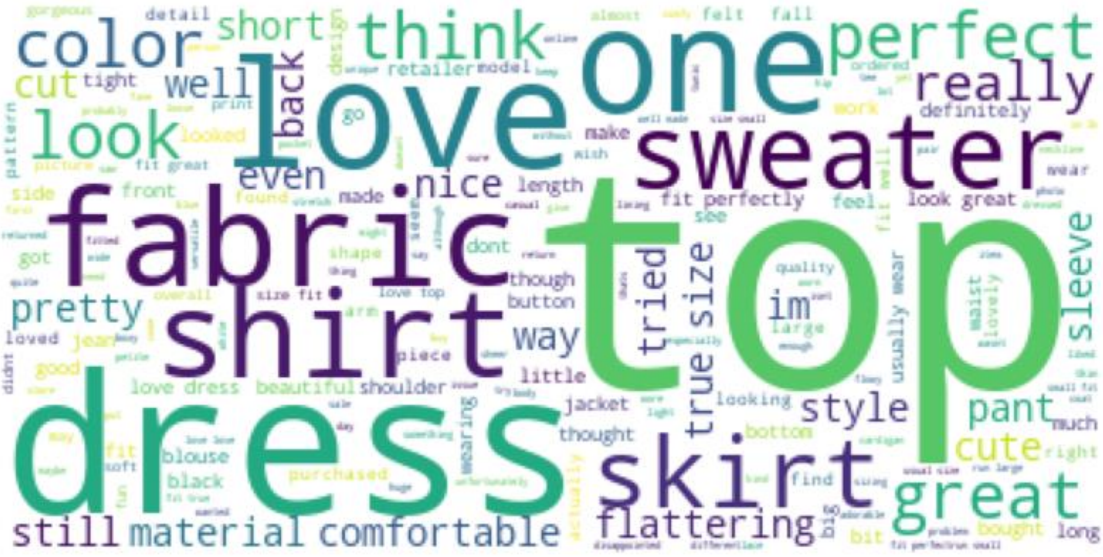
5.FEATURE SELECTION & DATA CLEANING

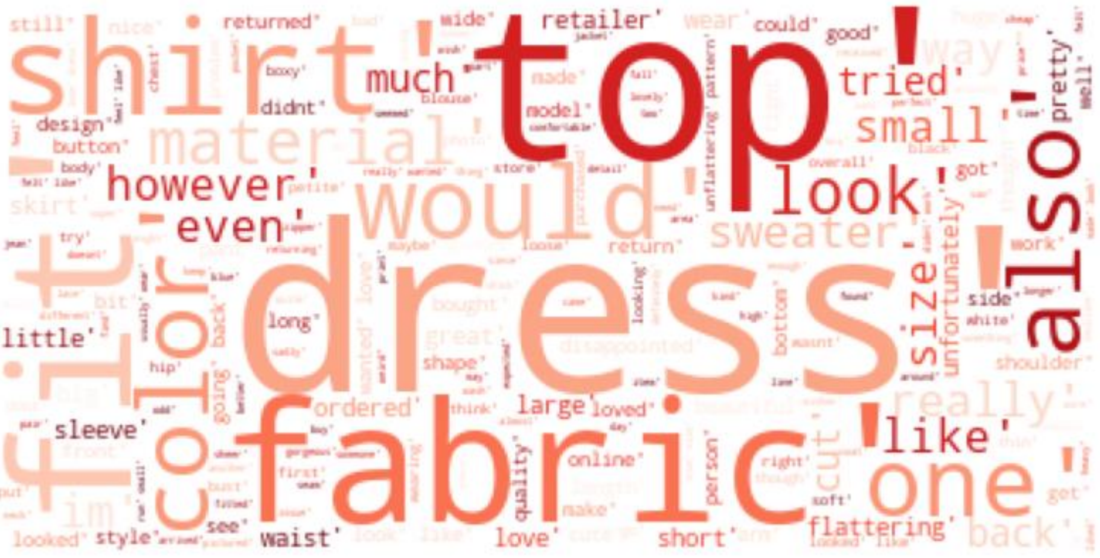
Since I am only using 2 features from this dataset, which are "Review Text" and "Recommended IND", I have dropped other unnecessary columns. After that I have performed some missing value analysis and since the number was relatively small, I dropped those rows too.

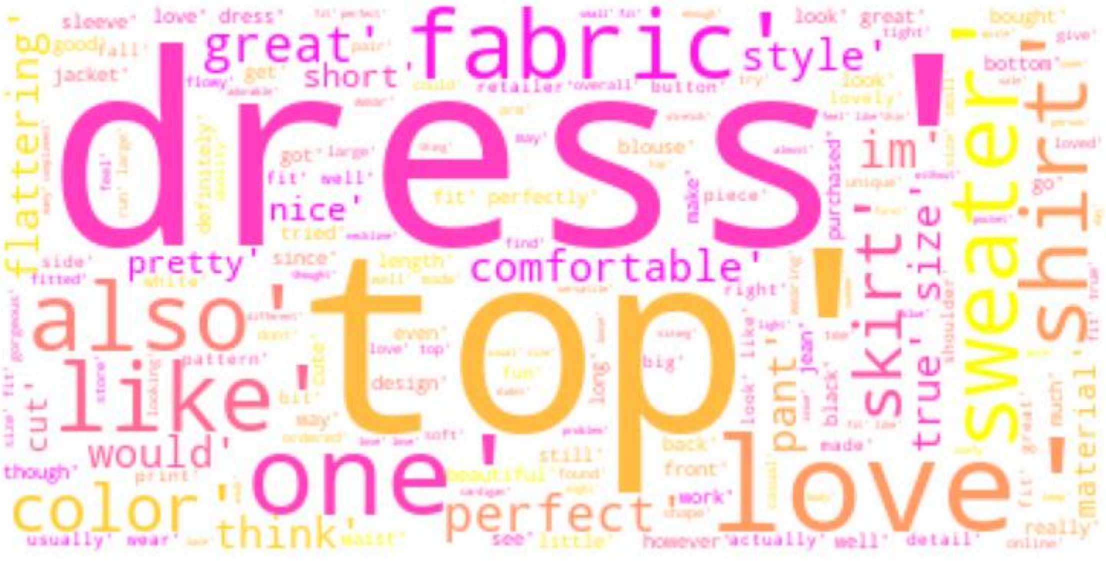
6.TEXT MINING

Text is an unstructured form of data, therefore various types of noise are present in it, which causes inability to analyze without any pre-processing. The process of making this text data ready for analysis is called text preprocessing. In this project, to preprocess the data I have first removed all punctuations, numbers, stop words, outliers, tokenized text and applied lemmatizer (Lexicon Normalization). After preprocessing, I created word clouds for all, positive and negative words.

7.WORDCLOUD


top
love
one
dress
fabric
shirt
sweater
skirt
color
perfect
think
short
really
look
great
style
pant
cute
sleeve
true size
tried
nice
even
pretty
way
still
material
comfortable
flattering
well
cut
back
definitely
little
jacket
thought
blouse
black
fit perfectly
look great
usually wear
love dress
beautiful
shoulder
looking
bottom
button
though
piece
im
large
quality
waist
lovely
front
side
got
loved
good
tight
detail
retailer
model
design
felt
fall
wish
make
length
feel
see
wear
find
big
bit
bought
long
purchased
soft
fun
fit
found
made
shape
dont
size fit
love top
work
ordered
print
looked
picture
fit great
pattern
go
actually
wearing
didnt
overall
jean


top'
dress'
fabric'
shirt'
fit'
would'
material'
also'
color'
look'
one'
size'
sweater'
tried'
small'
however'
even'
much'
really'
like'
unfortunately'
pretty'
cut'
back'
flattering'
large'
little'
sleeve'
long'
shape'
ordered'
design'
button'
returned'
retailer'
could'
good'
way'
still'
nice'
wide'
didnt'
boxy'
model'
made'
blouse'
body'
skirt'
try'
bit'
hip'
going'
wanted'
bought'
great'
disappointed'
return'
unflattering'
pattern'
looking'
bottom'
side'
white'
shoulder'
work'
got'
black'
overall'
loved'
think'
online'
quality'
person'
right'
soft'
get'
first'
im'
see'
waist'
love'
short'
style'
looked'
put'
make'
wear'
chest'
petite'
store'
detail'
well'


dress'
top'
fabric'
love'
one'
also'
like'
sweater'
shirt'
skirt'
great'
style'
comfortable'
color'
would'
perfect'
think'
flattering'
true'
size'
pant'
nice'
pretty'
short'
im'
cut'
material'
bottom'
lovely'
look'
fit'
well'
fit' perfectly'
good'
fall'
jacket'
sleeve'
love' dress'
look' great'
bought'
tight'
retailer'
button'
got'
large'
blouse'
definitely'
purchased'
make'
piece'
tried'
length'
since'
side'
pattern'
design'
bit'
even'
love' top'
fun'
long'
big'
jean'
black'
made'
still'
beautiful'
back'
front'
work'
cute'
though'
print'
usually' wear'
see'
little'
however'
actually'
well'
detail'
really'
online'
waist'
much'
shoulder'
white'
find'
soft'
may'
ordered'
flowy'
get'
first'
feel'
run' large'
gorgeous'
fitted'
store'
fit' true'

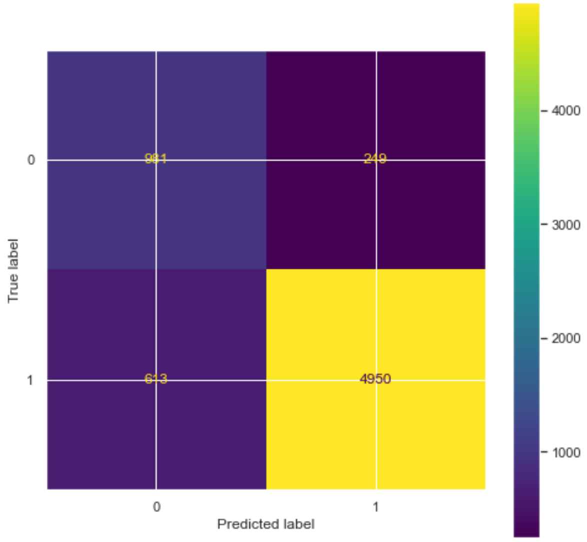
8.TRAIN/TEST SPLIT & VECTORIZATION

Before performing any modeling, I needed to apply vectorization to convert text documents into numeric feature vectors and train test split. From the tokenization step, review text column is in token form, so I used scikitlearn count vectorizer to convert that text into a matrix of token counts. To be able to evaluate from multiple approaches, I performed 2 different vectorization methods:1. Count Vectorization, 2. TF-IDF Vectorization. The reason for this is that I read that count vectorization is based on frequency representation and a simpler form of vectorization, on the other hand TF-IDF takes weight into account. So, if there is any, we will be able to observe their differences as well as classification algorithms. After this step, I will be building models using 4 classification algorithms.

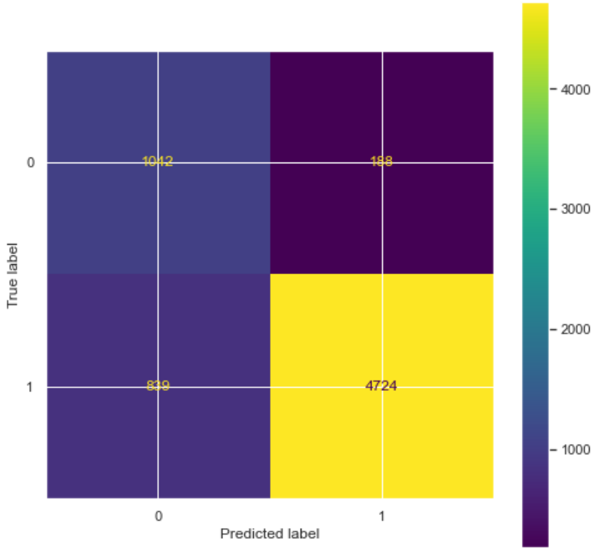
9.MACHINE LEARNING MODELS

After train and test split and vectorization I' ve created models with 4 classification algorithms, which are Logistic Regression, Naive Bayes, Support Vector Machine and Random Forest. Following are the heatmaps referencing models.

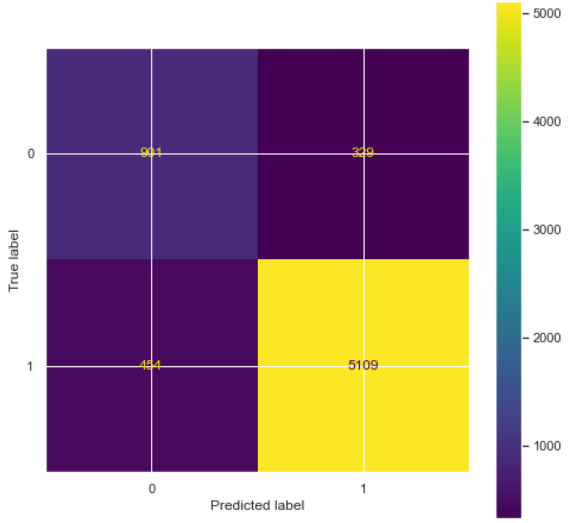
1)Logistic Regression with Count Vectorizer



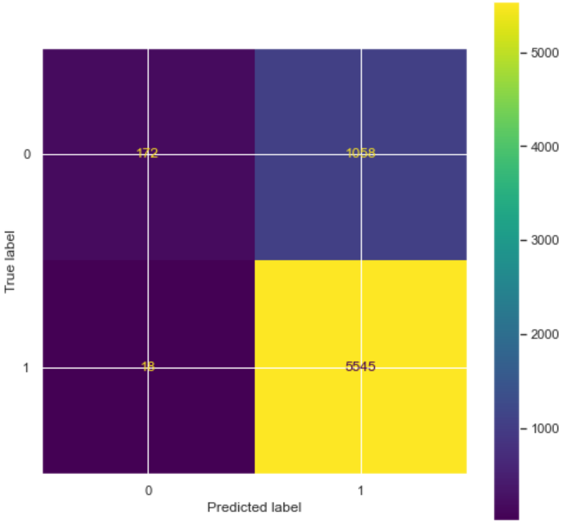
2) Logistic Regression With TF-IDF Vectorizer



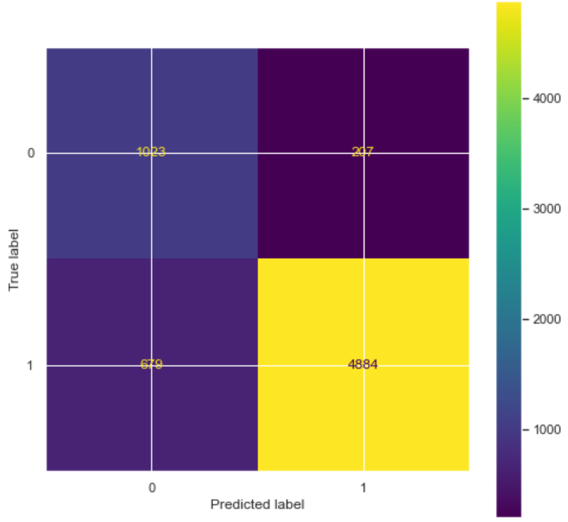
3) Naive Bayes with Count Vectorizer



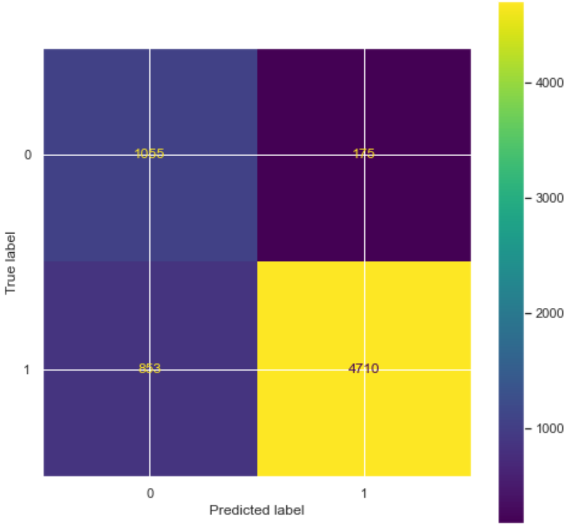
4) Naive Bayes with TF-IDF Vectorizer



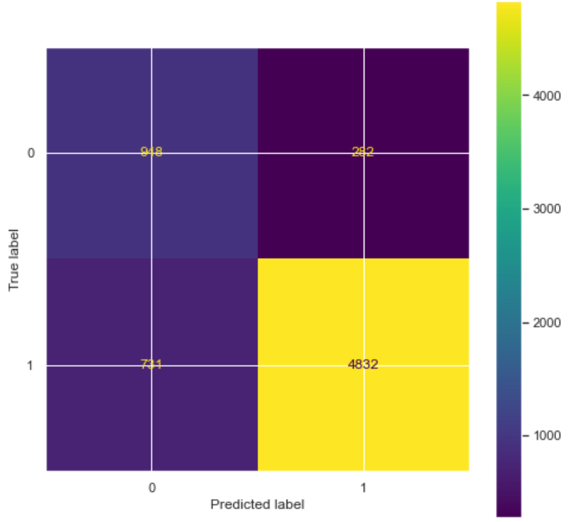
5) SVM With Count Vectorizer



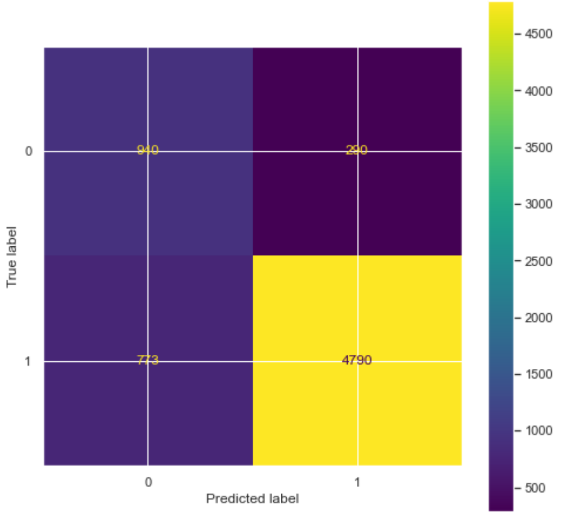
6) SVM With TF-IDF Vectorizer



7) Random Forest with Count Vectorizer



8) Random Forest With TF-IDF Vectorizer



10.COMPARISON OF MODELS AND EVALUATION

As conclusion, since the result of each model is relatively similar to each other, I think the best model could be chosen based on which indicator we want to value most, from accuracy and precision. Additionally, I did not observe a significant affect by changing the vectorization method, since scores are also similar in that case.

E. Aydın
19.01.2022
CSSM 502

RESOURCES:

https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

https://www.analyticsvidhya.com/blog/2021/06/part-5-step-by-step-guide-to-master-nlp-text-vectorization-approaches/

https://www.guru99.com/nltk-tutorial.html

https://www.kaggle.com/edchen/text-vectorization

https://medium.com/@paritosh_30025/natural-language-processing-text-data-vectorization-af2520529cf7

https://www.dataschool.io/comparing-supervised-learning-algorithms/

https://towardsdatascience.com/how-to-turn-text-into-features-478b57632e99

https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/