COMP262 SEC001 Group3

Project Report

Group 3

Nimish Patel

Sohamkumar Dharmendrabhai Patel

Tin Lap Li

Wenhao Feng

Wing Yan Lau

Table of Contents

[1. Data Exploration 2](#_Toc217959533)

[Dataset Columns 3](#_Toc1205734917)

[Data Size 3](#_Toc554140411)

[Statistics of Rating Score 3](#_Toc1334898399)

[Distribution of Rating 3](#_Toc839161688)

[Count of Review by Products (Top 10 Products) 3](#_Toc1056061306)

[Distribution of Rating by Products (Top 10 Products) 4](#_Toc1977660055)

[Distribution of Rating by Users (Top 10 Users) 4](#_Toc689624127)

[Count of Review by Users (Top 10 Users) 5](#_Toc21953224)

[Statistics of Review Length before Processing 5](#_Toc704358095)

[Statistics of Review Length after Processing 5](#_Toc1565615149)

[Label Count 6](#_Toc163495057)

[Conclusion 7](#_Toc2120547461)

[2. Data Pre-processing 8](#_Toc387104172)

[Pre-processing Steps 8](#_Toc1532264012)

[Remove Duplicate 8](#_Toc738311597)

[Columns Selection 8](#_Toc2044984117)

[Convert To Lower Case 8](#_Toc1872595348)

[Remove Stop Words 8](#_Toc1025945962)

[3. Text Representation 8](#_Toc369601662)

[4. Models 9](#_Toc1556412550)

[4.1 Lexicon Model 1 9](#_Toc169966980)

[a. Assumptions/Heuristics/algorithms used 9](#_Toc169970038)

[b. Explain each model, how it works 9](#_Toc2013985455)

[4.2 Lexicon Model 2 9](#_Toc848532933)

[5. Testing Results Summary 10](#_Toc1693410119)

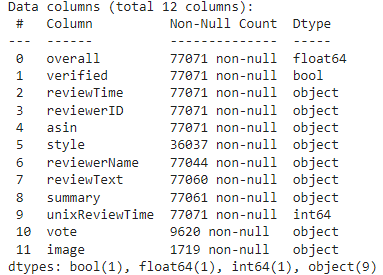
[References 11](#_Toc557107771)

[Appendix 1: Project plan 11](#_Toc900823459)

[Appendix 2: Meeting Register 11](#_Toc1068132636)

# 1. Data Exploration

## Dataset Columns



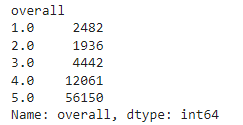
## Data Size

The dataset consists of 77071 rows and 12 columns.

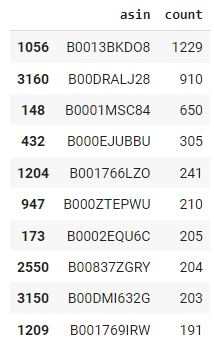
## Statistics of Rating Score

|  |  |
| --- | --- |
| **Metrics** | **Value** |
| Mean | 4.52 |
| Min | 1 |
| Max | 5 |

## Distribution of Rating



## Count of Review by Products (Top 10 Products)



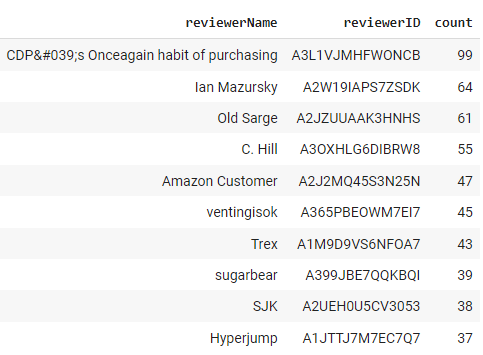
## Distribution of Rating by Products (Top 10 Products)

|  |  |
| --- | --- |
|  |  |

## Distribution of Rating by Users (Top 10 Users)

|  |  |
| --- | --- |
|  |  |

## Count of Review by Users (Top 10 Users)



## Statistics of Review Length before Processing

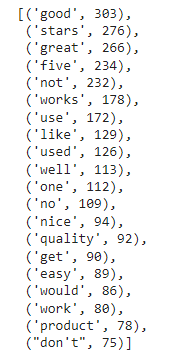
|  |  |
| --- | --- |
| **Metrics** | **Review Length (Words)** |
| Mean | 45 |
| Median | 19 |
| Min | 2 |
| Max | 6024 |

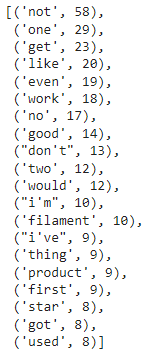
## Statistics of Review Length after Processing

|  |  |
| --- | --- |
| **Metrics** | **Review Length (Words)** |
| Mean | 28 |
| Median | 15 |
| Min | 3 |
| Max | 514 |

## Label Count

|  |  |
| --- | --- |
| **Label** | **Count** |
| Positive | 865 |
| Neutral | 69 |
| Negative | 66 |

Top 20 Words in Positive Reviews after Processing  


Top 20 Words in Negative Reviews after Processing  


## Conclusion

From the above data exploration process, several observations are obtained.

The rating score of review are mainly distributed to “5.0” having 56150 reviews which is significantly high comparing to that of “1.0”, “2.0” and “3.0” having not more than 5000 reviews respectively. This means that the dataset is unbalanced, having more positive reviews than negative.

In total 5334 of products, products with ID, B0013BKDO8, B00DRALJ28 and B0001MSC84 have more than 600 reviews, while all other products have around 300 or below reviews. Most of the ratings from the top 10 products are positive, falling on score “5.0”.

Reviewer with name, “CDP's Onceagain habit of purchasing”, contributes 99 reviews, who is the top contributor. Only 4 users contribute more than 50 reviews; other users are below 50. Most of the ratings from the top 10 users are positive, falling on score “5.0” and “4.0”.

Before preprocessing, the mean, median, minimum and maximum of review length are 45, 19, 2 and 6024 correspondingly. After preprocessing, the mean, median, minimum and maximum of review length are 28, 15, 3 and 514 correspondingly. The minimum length after preprocessing is higher than before is due to concatenation of “reviewText” and “Summary”.

The top 10 words appeared in positive reviews are as below:

good, stars, great, five, not, works, use, like, used, well

The top 10 words appeared in negative reviews are as below:

not, one, get, like, even, work, no, good, don’t, two

To conclude, the dataset is unbalanced on positive data.

# 2. Data Pre-processing

## Pre-processing Steps

The following pre-processing steps are carried out:

1. Remove duplicate reviews
2. Randomly select 1000 records as sample
3. Transform “overall” column to “rating\_tag” column by the following rules
   1. If value is larger than 3, consider as “pos” which means positive
   2. If value is equal to 3, consider as “neu” which means neutral
   3. If value is smaller than 3, consider as “neg” which means negative
4. Select column “reviewText” and “summary” as corpus to be analyzed
5. Concatenate column “reviewText” and “summary”
6. Convert corpus to lower case
7. Remove stop words

## Remove Duplicate

Duplicated reviews are found in the dataset. Before removal, there are 77071 reviews, after that, 72968 records left in dataset. There are 4103 rows of duplicated reviews.

## Columns Selection

“reviewText” and “summary” columns are selected for analyzation. “reviewText” contains the most context of the review, and “summary” contains less context but can still be used for analysis.

## Convert To Lower Case

This is a standardizing procedure of the text for computer to easier understand human input; thus give higher accuracy. Moreover, this step helps in removing the stop word set using stop word library for steps afterwards.

## Remove Stop Words

There are common words that are used within sentences without any meaning. Those words are called stop words. As they do not have value to the dataset, they can be eliminated for cleaning purposes. In this sentiment analysis case, some of the negative words are removed from the stop word set such as “not” and “no” to preserve the negative context of the data.

# 3. Text Representation

We opted to use the TF-IDF technique for text representation because of its ability to address the weighted frequency of each term by considering its overall frequency across the entire document and inversely measuring how rare the term is across the corpus. This helps to minimize the impact of irrelevant words on sentimental analysis results. Additionally, TF-IDF generates a sparse matrix representation that reduces the likelihood of overfitting in the model.

It is worth noting that other text representation techniques such as Bag of Words (BoW) and Count Vectorizer are also commonly used in natural language processing tasks. However, BoW only considers the frequency of each term in a document, without considering its overall frequency across the entire corpus. On the other hand, Count Vectorizer computes each term's frequency across the whole corpus but does not account for the rarity of the term in the current document. In contrast, TF-IDF provides a more comprehensive approach to text representation by combining both the frequency and rarity of each term.

# 4. Models

## 4.1 Lexicon Model 1

### a. Assumptions/Heuristics/algorithms used

The dataset only contains around 70,000 rows of data, so our problem does not need to scalable. Rule-based AI models can operate with simple basic information and data, while machine learning system requirement more data than rule-based models. Second, the dataset is about review of a product which is not a specific knowledge domain, so the Vader model should works best when applied to social media text or review.

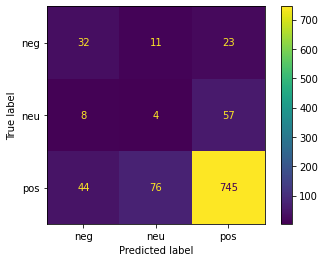
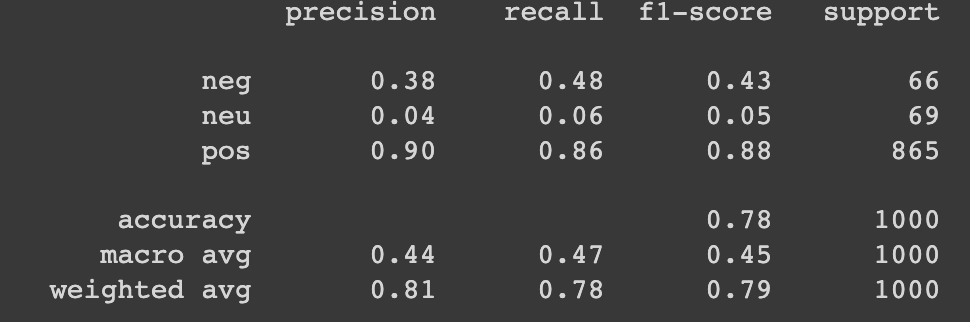
### b. Explain each model, how it works

Each word in the VADER lexicon is rated as positive or negative, and in many cases, how positive or negative. Sentiment ratings from 10 independent human raters were used to rate over 9,000 token features on a scale from "[–4] Extremely Negative" to "Extremely Positive", with allowance for "Neutral (or Neither, N/A)". VADER produces four sentiment metrics from these word ratings: positive, neutral, negative, and compound. The first three metrics represent the proportion of the text that falls into each category while the compound metric represents an overall score for the text ranging from -1 (most negative) to +1 (most positive). Vader also takes into account the context in which words are used, as well as punctuation and capitalization, in order to improve the accuracy of its sentiment analysis. This makes it particularly effective at analyzing social media text, which often includes informal language and unconventional grammar.

## 4.2 Lexicon Model 2

TextBlob’s sentiment analysis consists of polarity and subjectivity. Polarity scores lies between –1 and 1, where –1 refers to negative sentiment and 1 refers to positive sentiment. Subjectivity scores between 0 and 1, it rates if the text contains personal opinion. The model also caters emoji and exclamation marks when analyzing sentiment of the text.

During our experiment, TextBlob shows 78.1% overall accuracy. The figures show Textblob has high on positive comments. But it shows a tendency to lean towards positive result when the actual sentiment is neutral.



# 5. Testing Results Summary

**Overall accuracy:**

|  |  |
| --- | --- |
| **TextBlob** | **Vader** |
| 78.1% | 78.9% |

**F1-score:**

|  |  |  |
| --- | --- | --- |
|  | **TextBlob** | **Vader** |
| Negative | 0.43 | 0.45 |
| Neutral | 0.05 | 0.07 |
| Positive | 0.88 | 0.89 |

The results found Vader generally performed slightly better than TextBlob on sentiment analysis. Both models show great performance on positive sentiments, but low performance on neutral and negative sentiments.

# References

Using VADER to handle sentiment analysis with social media text

https://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html

# Appendix 1: Project plan

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Due Date** | **Person in Charge** | **Expected Finish Date** |
| Data Exploration & Pre-processing | Week 5 | Wing | 6 Feb, 2023 |
| Text representation | Week 6 | Nimish & Sohamkumar | 10 Feb, 2023 |
| Lexicon Model 1 | Week 8 | Wenhao | 10 Mar, 2023 |
| Lexicon Model 2 | Week 8 | Tin | 10 Mar, 2023 |
| Machine Learning Model 1 | Week 12-14 | Wing | 7 Apr, 2023 |
| Machine Learning Model 2 | Week 12-14 | Nimish & Sohamkumar | 7 Apr, 2023 |
| Enhancement | Week 14 | Nimish & Sohamkumar | 16 Apr, 2023 |
| Test Result Comparison | Week 14 | Tin | 16 Apr, 2023 |
| Conclusion | Week 14 | Wenhao | 16 Apr, 2023 |

# Appendix 2: Meeting Register

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Meeting** | **Date** | **Time** | **Agenda** | **Attendees** |
| First Meeting | 7 February, 2023 | 11:00am – 11:30am | Division of work and project timeline | Everyone |
| Second Meeting | 13 February, 2023 | 4:00pm – 4:20pm | Discussion Lexicon model and text representation | Everyone |