CSE508 Information Retrieval

Winter 2024

Assignment-3

Konam Akhil Vamshi

2020513

Considerations:

- I used “Charging cable” as a product for the assignment.
- Merged two datasets, ‘reviews’ and ‘metadata’ on the asin column.
- Used TfidfVectorizer to model review text.
- Generated acronym dictionary from Chatgpt

Preprocessing:

- Dropped null values.
- Dropped duplicate rows.

Total number of rows for the product:

```
Total number of rows for 'Charging cables': 19456
```

Descriptive Statistics of the Product:

```
Number of Reviews: 19456
Average Rating Score: 4.20
Number of Unique Products: 439
Number of Good Ratings: 16541
Number of Bad Ratings: 2915
Number of Reviews corresponding to each Rating:
overall
1     1923
2      992
3     1220
4     2410
5    12911
dtype: int64
```

Preprocessing:

```python
import unicodedata
import re
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize


charging_cable_acronyms = {
    'USB': 'Universal Serial Bus', 'USB-C': 'Universal Serial Bus Type C', 'USB-A': 'Universal Serial Bus Type A',
    'USB-B': 'Universal Serial Bus Type B', 'USB 2.0': 'Universal Serial Bus Version 2.0', 'USB 3.0': 'Universal Serial Bus Version 3.0',
    'USB PD': 'USB Power Delivery', 'PD': 'Power Delivery', 'mAh': 'Milliamp Hour', 'A': 'Ampere',
    'V': 'Volt', 'W': 'Watt', 'Wh': 'Watt Hour', 'QC': 'Quick Charge', 'LED': 'Light Emitting Diode',
    'OEM': 'Original Equipment Manufacturer', 'Li-ion': 'Lithium Ion', 'NiMH': 'Nickel Metal Hydride',
    'AC': 'Alternating Current', 'DC': 'Direct Current', 'MFI': 'Made For iPhone/iPad',
    'AWG': 'American Wire Gauge', 'HDMI': 'High Definition Multimedia Interface', 'DP': 'DisplayPort',
    'TB': 'Thunderbolt', 'PVC': 'Polyvinyl Chloride', 'TPE': 'Thermoplastic Elastomer',
    'EMI': 'Electromagnetic Interference', 'RFI': 'Radio Frequency Interference', 'IP': 'Ingress Protection',
    'IoT': 'Internet of Things', 'SBC': 'Single Board Computer', 'CE': 'Conformité Européenne',
    'FCC': 'Federal Communications Commission', 'UL': 'Underwriters Laboratories', 'RoHS': 'Restriction of Hazardous Substances',
    'PSE': 'Product Safety Electrical Appliance & Material', 'BSMI': 'Bureau of Standards, Metrology and Inspection',
    'KCC': 'Korea Certification Commission', 'SAA': 'Standards Association of Australia',
    'ERP': 'Energy-Related Products', 'CCC': 'China Compulsory Certificate', 'C-tick': 'Australian Certification Mark for Electromagnetic Compatibility',
    'GFCI': 'Ground Fault Circuit Interrupter', 'MOSFET': 'Metal Oxide Semiconductor Field Effect Transistor',
    'PCB': 'Printed Circuit Board', 'SMD': 'Surface Mount Device', 'DIP': 'Dual In-line Package',
    'BT': 'Bluetooth', 'RF': 'Radio Frequency', 'IC': 'Integrated Circuit', 'ESD': 'Electrostatic Discharge',
    'FOM': 'Figure of Merit', 'PPTC': 'Polymeric Positive Temperature Coefficient', 'NTC': 'Negative Temperature Coefficient',
    'PTC': 'Positive Temperature Coefficient', 'LDO': 'Low Dropout', 'SOC': 'State of Charge',
    'SOH': 'State of Health', 'LVP': 'Low Voltage Protection', 'OVP': 'Over Voltage Protection',
    'OCP': 'Over Current Protection', 'OTP': 'Over Temperature Protection', 'OPP': 'Over Power Protection',
    'SCP': 'Short Circuit Protection', 'DPDM': 'Dual Role Power Data Management', 'DRP': 'Dual Role Power',
    'D+': 'Data Plus', 'D-': 'Data Minus', 'SOP': 'Start of Packet', 'EOP': 'End of Packet',
    'CC': 'Configuration Channel', 'Vbus': 'Voltage Bus', 'GND': 'Ground'
}

lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))

def expand_acronyms(text, acronym_dict):
    words = word_tokenize(text)
    expanded_words = [acronym_dict.get(word, word) for word in words]
    return ' '.join(expanded_words)

def preprocess_text(text, acronym_dict):
    # Remove HTML tags
    text = BeautifulSoup(text, "html.parser").get_text()

    # Remove accented characters
    text = unicodedata.normalize('NFKD', text).encode('ascii', 'ignore').decode('utf-8', 'ignore')

    # Expand acronyms
    text = expand_acronyms(text, acronym_dict)

    # Remove special characters
    text = re.sub(r'[^a-zA-Z0-9\s]', '', text)

    # Lemmatization and remove stop words
    text = ' '.join([lemmatizer.lemmatize(word) for word in text.split() if word not in stop_words])

    return text.lower()

# Apply preprocessing to reviewText
df_reviews_headphones['processed_reviewText'] = df_reviews_headphones['reviewText'].apply(preprocess_text)
```

EDA:

```
Top 20 most reviewed brands:
 brand
iTEKIRO                220
UPBRIGHT                59
HQRP                    29
Live2Pedal              28
NiceTQ                  27
ReadyPlug               27
Life-Tech               25
Conwork                 23
Generic                 22
ANiceS                  21
Super Power Supply      20
                        17
Factory Direct          17
CoverON                 16
Cabepow                 14
MyNetDeals              14
WILLTOP                 14
ALPHA TECH              13
TUSITA                  13
iTKEIRO                 13
Name: count, dtype: int64
```

```
Top 20 least reviewed brands:
 brand
Everus                             1
Shenzhen TOZ Technology co., LTD   1
Acasis                             1
Equinux                            1
LANSUNS                            1
Voroar                             1
CJRSLRB                            1
enKo Products                      1
Fully                              1
LEPOWER                            1
Motorola FRS                       1
KssFire                            1
CTYRZCH                            1
Bodelin                            1
amovee                             1
KaLaiXing                          1
Comkia                             1
MISSJIRA                           1
Aplusphone                         1
SZ                                 1
Name: count, dtype: int64
```

```
Count of ratings for the product over 5 consecutive years:
 year
2014    2496
2015    5059
2016    6373
2017    3279
2018    1105
dtype: int64
```

```
Most positively reviewed Charging Cable ASIN and average rating:
             overall
asin
B0173MPK8O      5.0
```

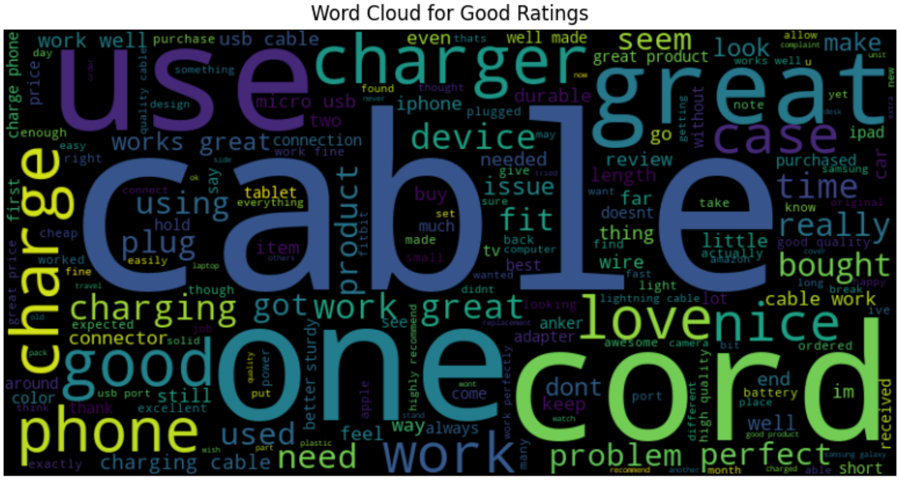
For Bad Ratings:

- The word "charge" is quite prominent, suggesting issues with charging capabilities.
- "Phone" and "cable" also stand out, which may indicate problems with phone cables.
- Words like "stopped," "broke," "cheap," and "return" suggest dissatisfaction with product durability, quality, and perhaps a desire to return the product.
- "Work" and "worked" alongside negative terms suggest that the products often failed to work as expected.

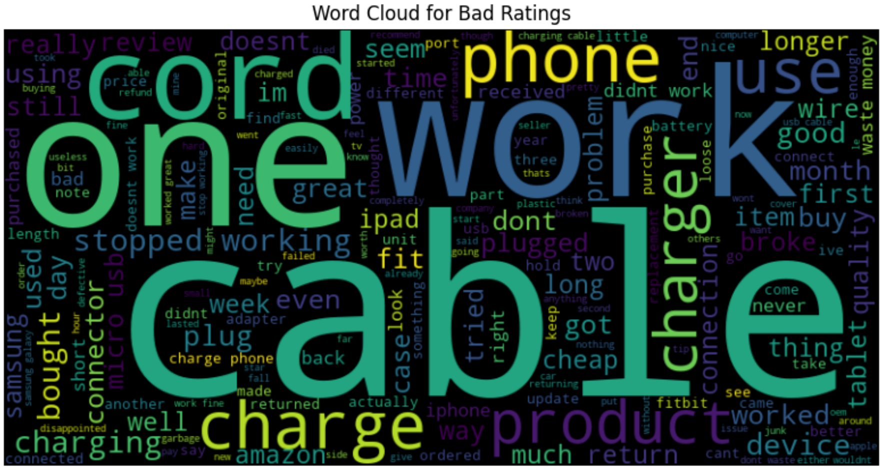
For Good Ratings:

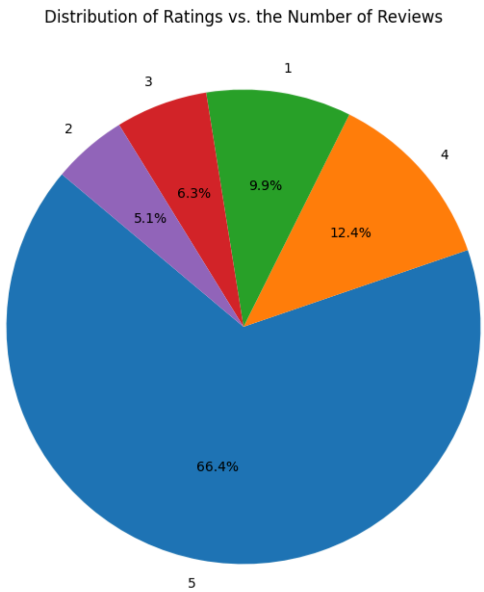
- Positive words like "great," "good," "works," "nice," and "perfect" indicate satisfaction with the products.
- "Charge" and "charging" appear here as well but in a positive light, suggesting successful charging experiences.
- "Phone" and "device" are visible, which could mean that the items in question are related to phone or device accessories or usage.
- The presence of "fit" and "well made" implies that customers found the products to be of a good fit and well-crafted.

Word Cloud for Good Ratings



Word Cloud for Bad Ratings






```
Year with maximum reviews: 2016
```

```
Year with the highest number of customers: 2016
```

Performance of models:

Model: Logistic Regression

	precision	recall	f1-score	support
Average	0.47	0.24	0.31	313
Bad	0.74	0.71	0.73	640
Good	0.92	0.97	0.94	3911
accuracy			0.89	4864
macro avg	0.71	0.64	0.66	4864
weighted avg	0.87	0.89	0.88	4864

```
Model: K-Nearest Neighbors
              precision    recall  f1-score   support

     Average       0.35      0.06      0.10       313
         Bad       0.62      0.35      0.45       640
        Good       0.86      0.97      0.91      3911

    accuracy                           0.83      4864
   macro avg       0.61      0.46      0.49      4864
weighted avg       0.79      0.83      0.80      4864
```

```
Model: Linear SVM
              precision    recall  f1-score   support

     Average       0.44      0.27      0.33       313
         Bad       0.73      0.74      0.74       640
        Good       0.93      0.95      0.94      3911

    accuracy                           0.88      4864
   macro avg       0.70      0.66      0.67      4864
weighted avg       0.87      0.88      0.88      4864
```

```
Model: Naive Bayes
              precision    recall  f1-score   support

     Average       0.21      0.03      0.05       313
         Bad       0.72      0.65      0.68       640
        Good       0.89      0.97      0.93      3911

    accuracy                           0.87      4864
   macro avg       0.61      0.55      0.55      4864
weighted avg       0.83      0.87      0.84      4864
```

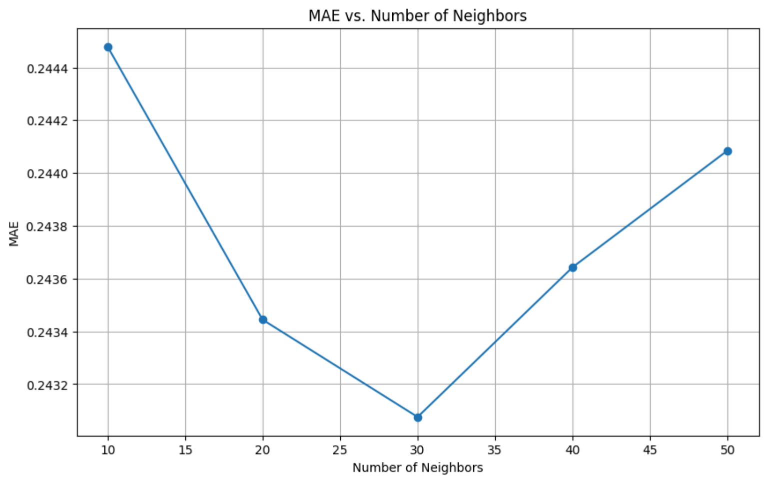
```
Model: Decision Tree
              precision    recall  f1-score   support

     Average       0.44      0.02      0.04       313
         Bad       0.63      0.41      0.50       640
        Good       0.86      0.97      0.91      3911

    accuracy                           0.84      4864
   macro avg       0.64      0.47      0.48      4864
weighted avg       0.80      0.84      0.80      4864
```

- All models perform best at identifying 'Good' ratings, likely due to a larger number of examples to learn from.
- 'Average' ratings are consistently the hardest to predict for all models, indicating a challenge in distinguishing moderate sentiments.
- 'Bad' ratings are easier for models to identify than 'Average', hinting at more distinct linguistic features in negative reviews.
- Class imbalance impacts performance, with models favoring the majority 'Good' class and performing poorly on the minority classes.
- The macro average scores are lower than weighted averages, revealing that models are not as effective when classes are evenly considered.

Collaborative filtering:

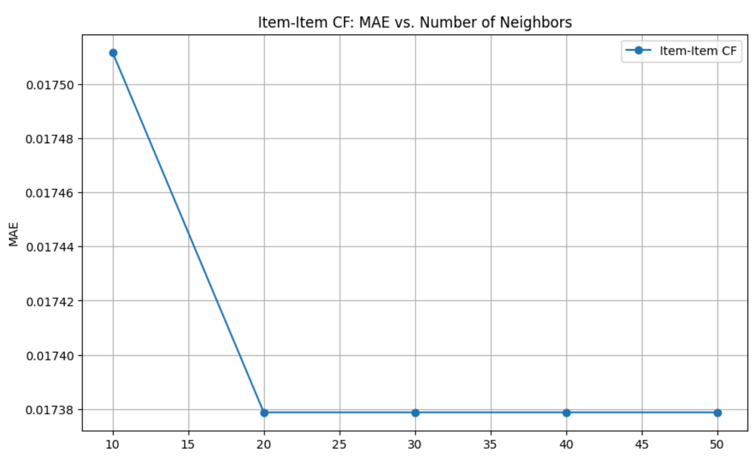
User-User:

MAE for 10 neighbors: 0.24447957718442326
MAE for 20 neighbors: 0.24344435465172745
MAE for 30 neighbors: 0.2430754553426476
MAE for 40 neighbors: 0.2436431699374561
MAE for 50 neighbors: 0.24408568254043855



Item-Item:

MAE for 10 neighbors (Item-Item CF): 0.017511671335200745
MAE for 20 neighbors (Item-Item CF): 0.01737878096975224
MAE for 30 neighbors (Item-Item CF): 0.01737878096975224
MAE for 40 neighbors (Item-Item CF): 0.01737878096975224
MAE for 50 neighbors (Item-Item CF): 0.01737878096975224



```
Top 10 Products by User Sum Ratings:
 asin
B00R1EPGRA    12396.5
B00TIT3KYC    11978.0
B0177MQWC8     2099.0
B00MY05GNA     1921.0
B00SVNGKJ8     1860.0
B01DNTWGYM     1192.0
B01ANLA6OU     1041.0
B0177L6A4O      982.0
B01HCT3GCU      935.0
B01FVUHW1I      838.5
```