Mobile Reviews NLP and Clustering

Afnan Alsirhani & Moroj Aldeeb

Introduction

- We utilized unsupervised machine learning techniques to gain more insights into mobile products reviews of amazon.
- The dataset contains approximately 417000 records which explain customers’ ratings , sentiments and reviews over a set of predefined mobile products

objective

- Using unsupervised machine learning techniques along with NLP methods to understand customers' sentiments over different products.
- Explain the differences between customers' reviews based on latent topics contained within to better tailor future marketing and product improvements.

METHODOLOGY

EDA

Wrangling
NLP Preprocessing
Tokenization
Lemmatization/Stemming
Emoji Conversion
Stop Words Removal
Entropy
Vectorization (Count/TF-IDF)

Dimensionality Reduction
- Singular Value Decomposition
- Principal Component Analysis

Topic Modeling
- Latent Semantic Analysis
- LDA (Latent Dirichlet allocation)

Clustering (K-Means)

Classification (SGB classifier, Logistic regression ,Naive Bayes)

Metadata

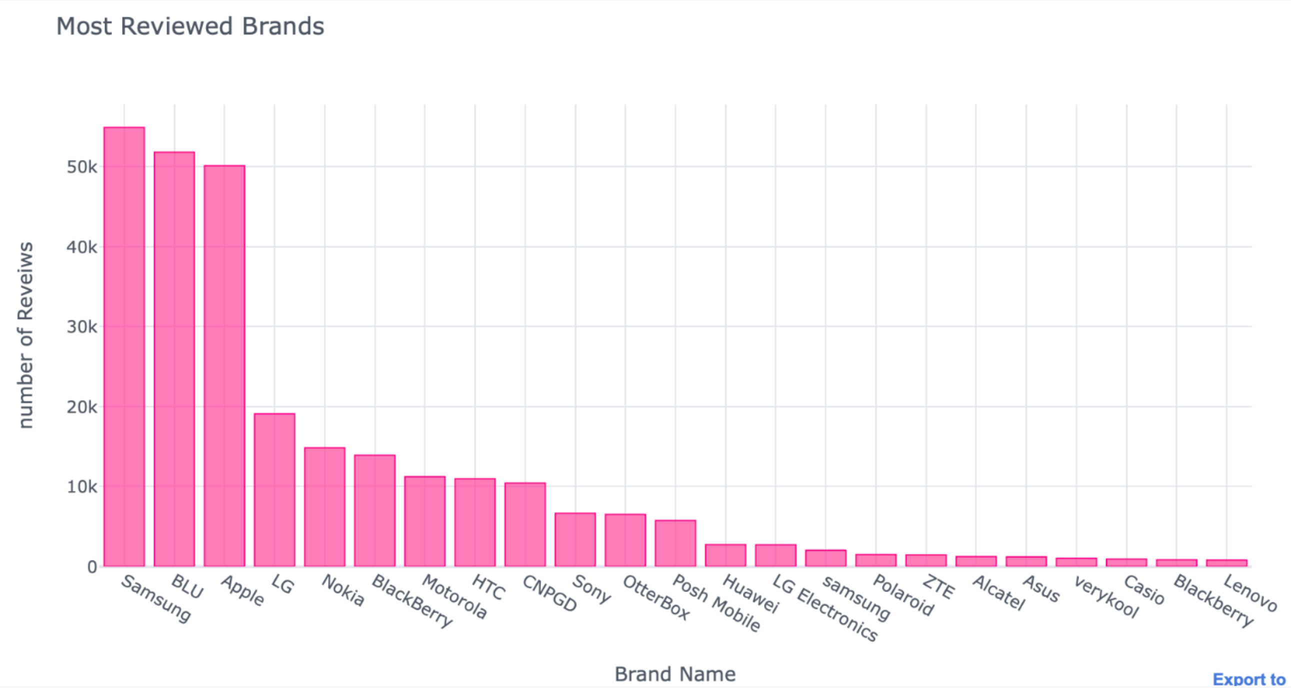
- The data source used was the Kaggle website.
- Data frame shape:-

Original data: 413840 observations and 6 columns → After EDA: 226000 observations and 6 columns → After Feature engineering: 63000 observations and ~ 6000 columns

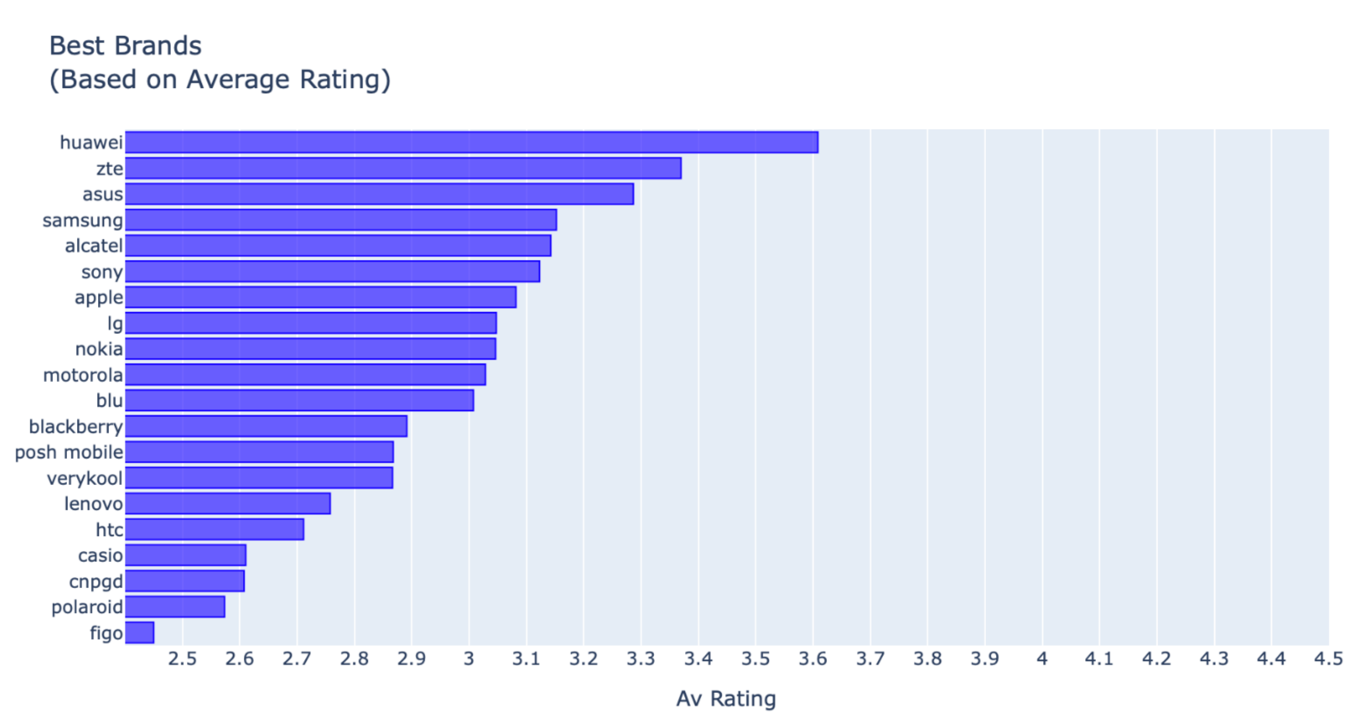
Tools and Libraries

- spaCy
- Corex
- NLTK
- Pandas
- gensim
- Wordcloud
- ScatterText
- Python
- TextBlob

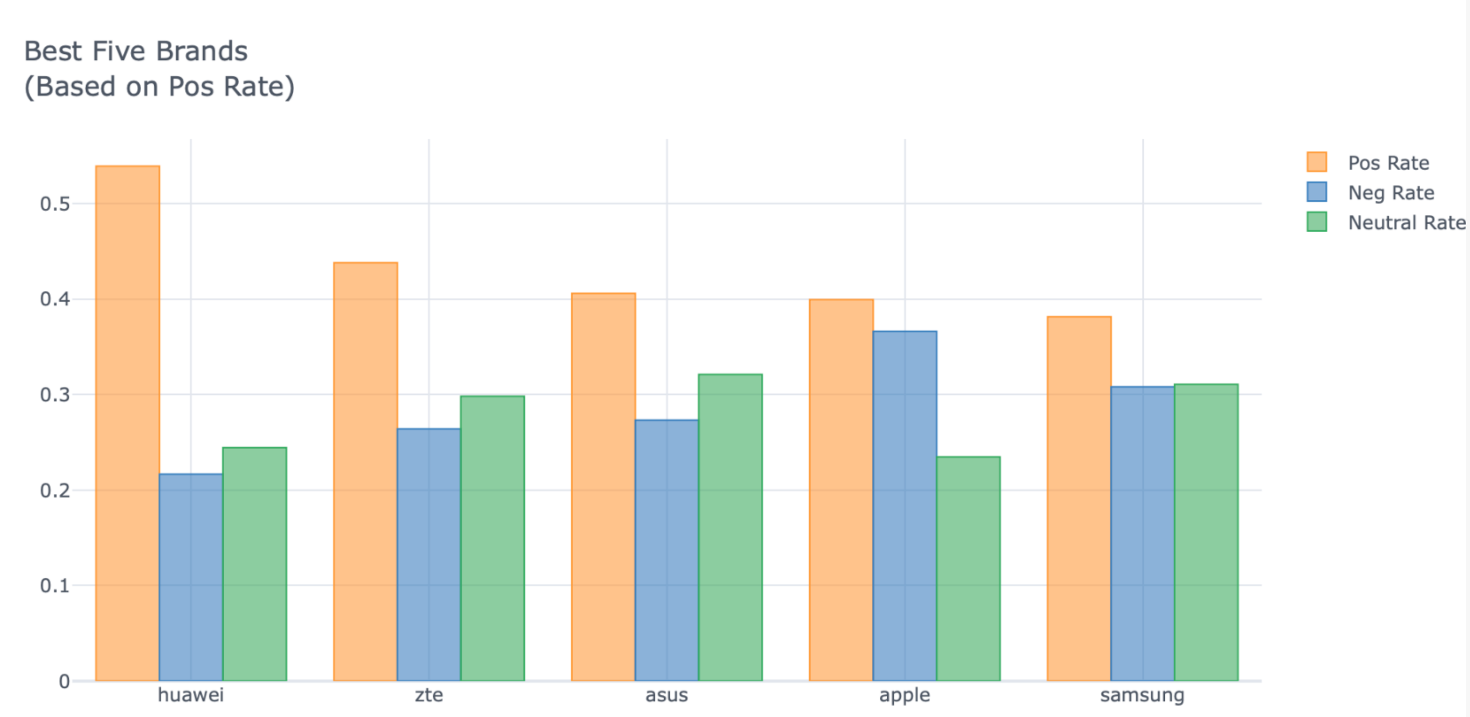
Data Understanding



Data Understanding

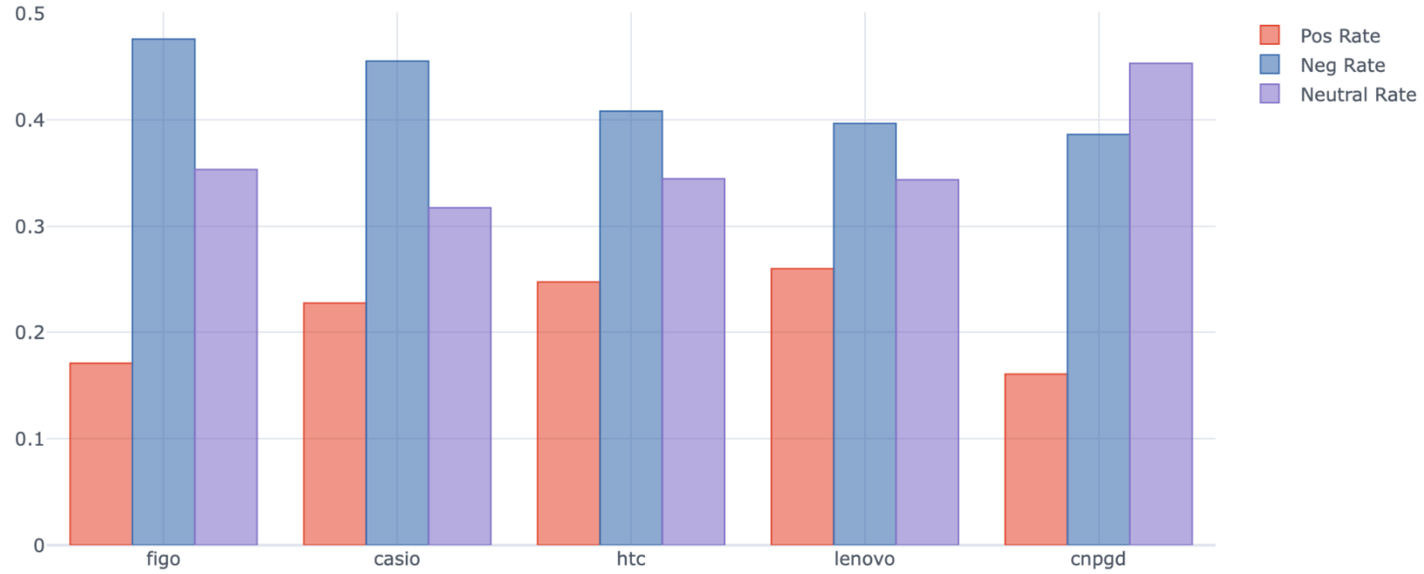


Data Understanding

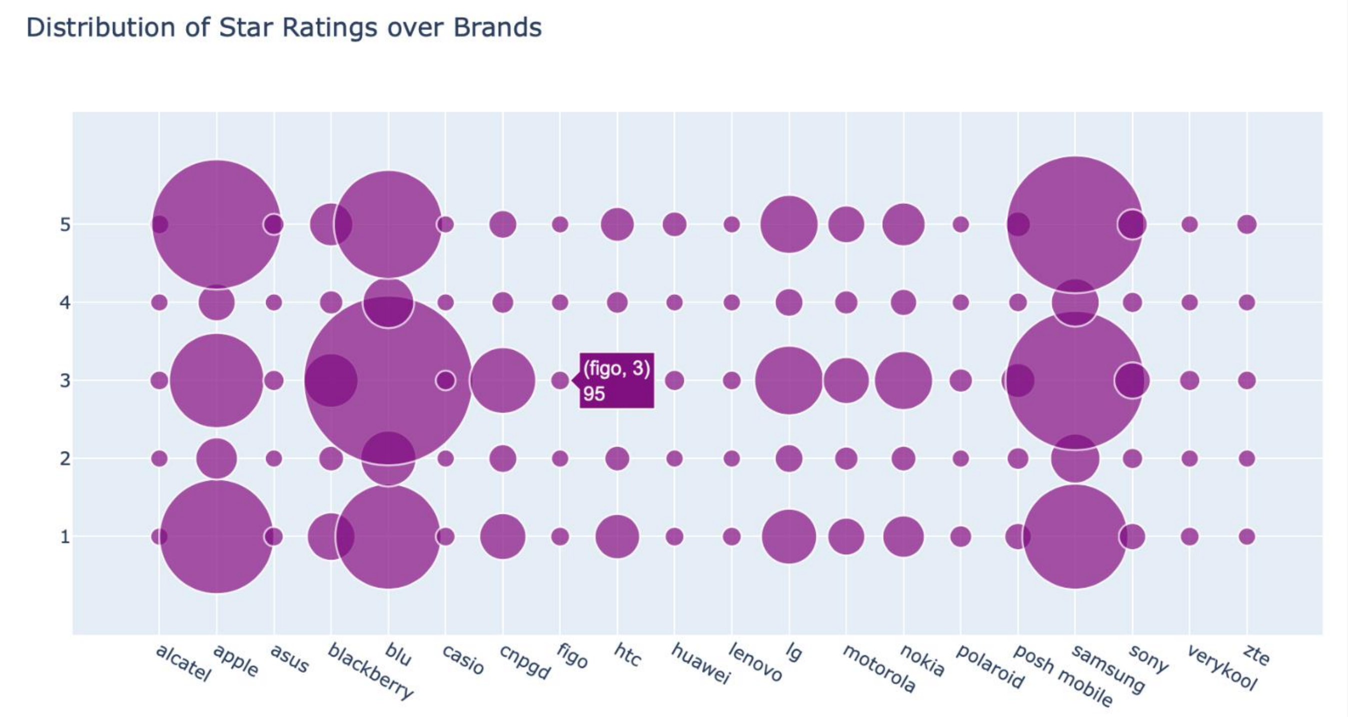


Data Understanding

Worst 5 Brands
(Based on Pos Rate)



Data Understanding



Feature Engineering

- Features selection:
- *We removed most common and rare words in the corpus after NLP Preprocessing*
- Features extraction:

We used Latent Semantic Analysis to generate the most suitable latent dimensions that most explain the data

Topic Modelling

Latent Semantic Analysis (LSA/SVD)	We found 5 main topics for both positive and negative reviews.
Latent Dirichlet Allocation (LDA) model	We found 5 main topics for both positive and negative reviews

Dimensionality Reduction

Principal Component Analysis (PCA)	We found that the best latent components are just 2 principal components for both positive and negative reviews.
Singular value decomposition (SVD)	We utilized SVD for topic modeling, we found different latent dimensions for positive and negative reviews. For Positive Reviews, we found that the best latent dimensions are 2 whereas negative reviews are 4.

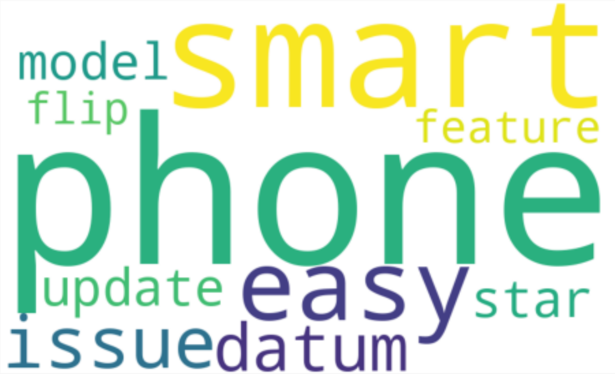
Count Vectorizer vs TF/IDF

- We used Count Vectorizer and Term Frequency / Inverse Document Frequency (TF/IDF) and we found that TF/IDF gives better results for subsequent topic modelling.

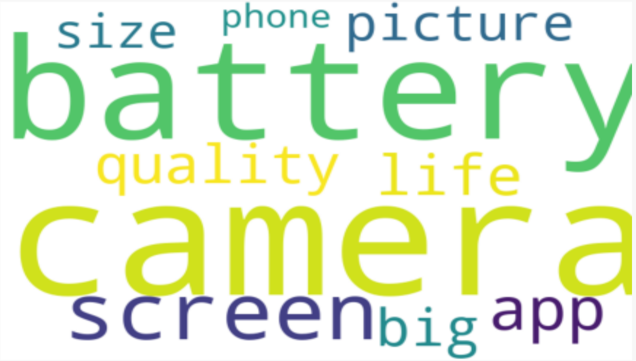
Word cloud for positive reviews

Word cloud for each topic in positive reviews

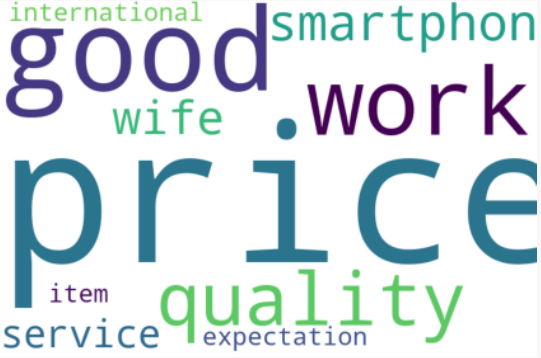
Topic 0

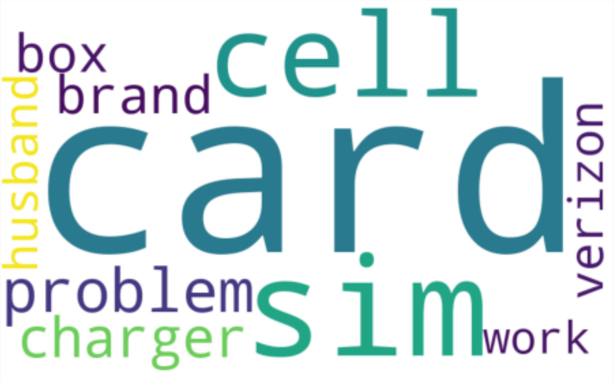


Topic 1



Topic 2

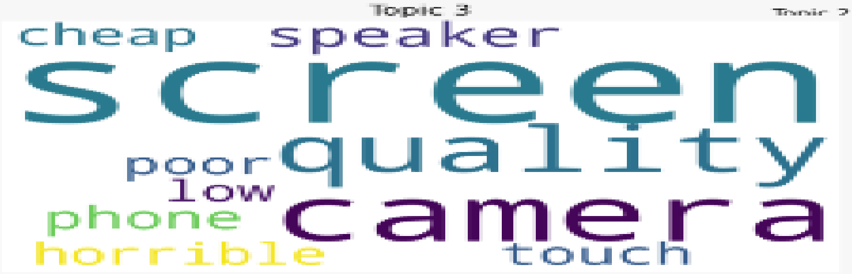




Topic 4

Word cloud negative reviews

Word cloud for each topic in negative reviews



CorEx

```
0: screen,big,size,protector,touch,large,bright,resolution,brightness,inch
1: speaker,loud,phone,video,feature,button,case,picture,text,music
2: quality,performance,sound,build,impressed,functionality,superb,stutter,capture,fluid
3: memory,ram,battery,camera,app,life,update,storage,game,high
4: card,sim,dual,micro,slot,local,microsd,travel,straight,ready
5: charger,cable,new,review,work,version,day,box,able,charge
```

Component 1 (topic 1) seems to be about Screen Features

Component 2 (topic 2) seems to be about Sound Features

Component 3 (topic 3) seems to be about Quality

Component 4 (topic 4) seems to be about Memory

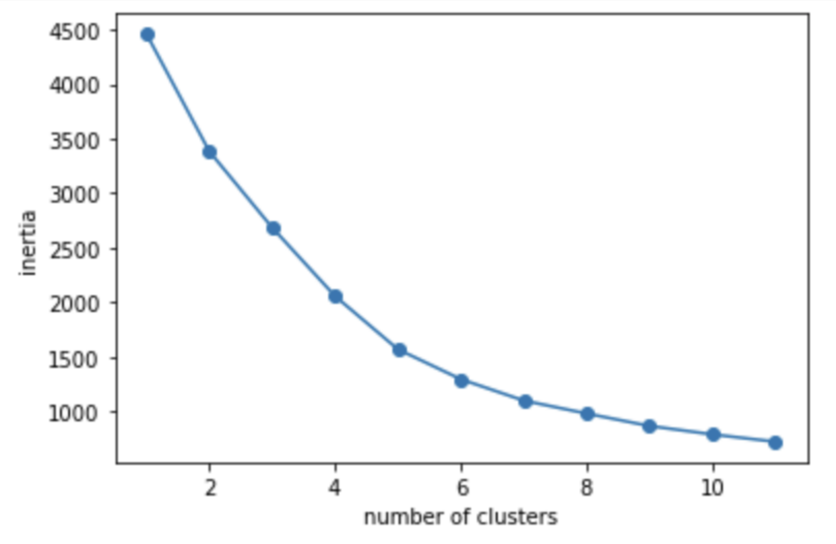
Component 5 (topic 5) seems to be about Card

Component 6 (topic 6) seems to be about Mobile Accessaries

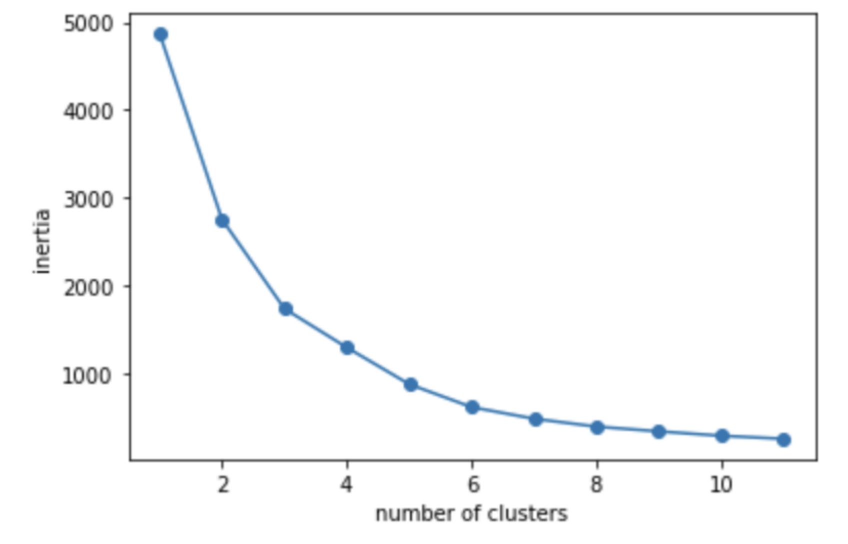
Clustering

- By performing clustering on positive reviews and negative reviews separately, we found unequal latent dimensions between these two categories.
- Positive reviews can be represented sufficiently by only two latent dimensions whereas negative reviews could be represented by more than two latent dimensions with very low overlap.

Elbow Differences between Positive and Negative Reviews



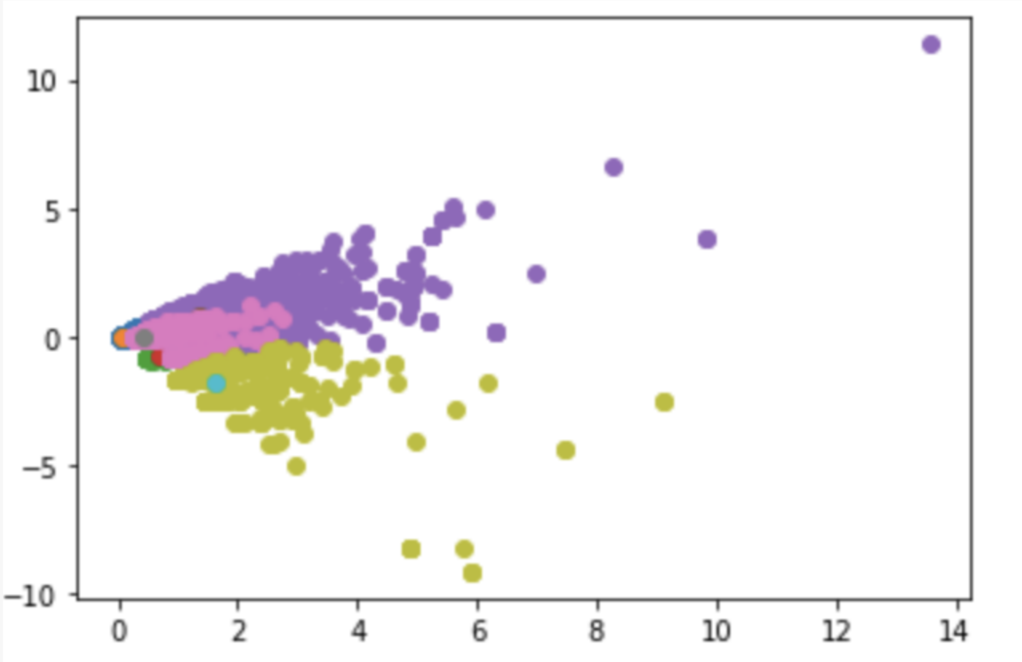
Positive Reviews



Negative Reviews

Positive Reviews (K-Means)

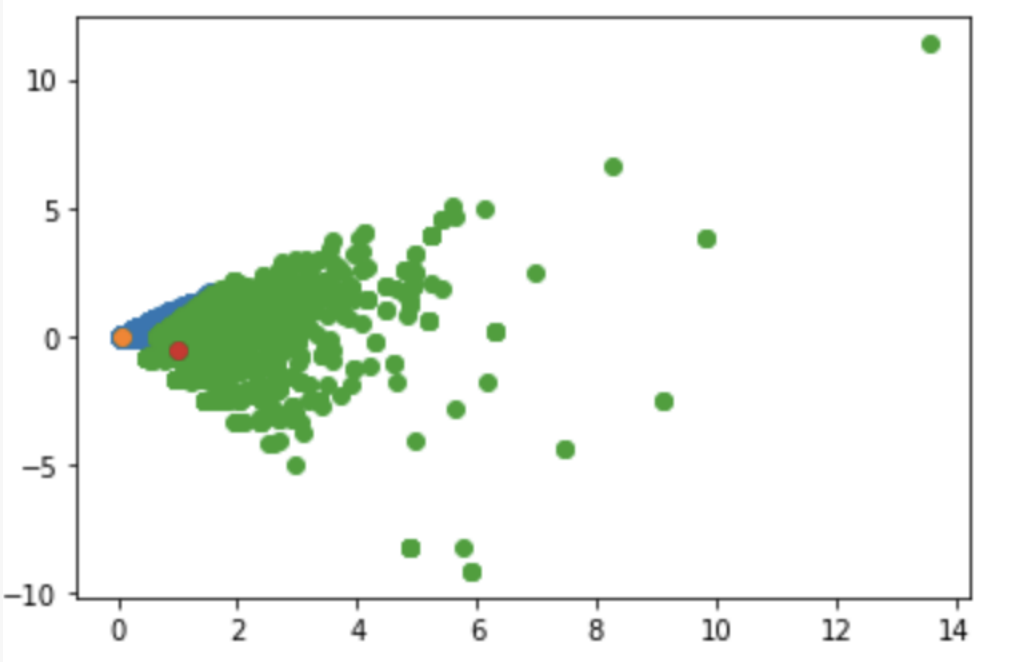
- Positive Reviews with 5 latent dimensions.
- Obvious overlap among clusters in higher dimensions.



Positive Reviews (K-Means)

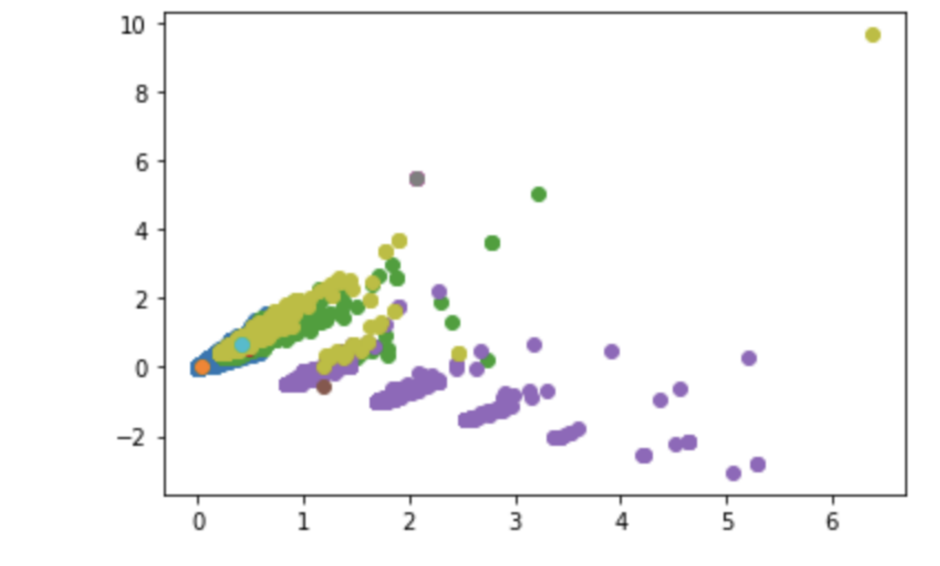
- Positive Reviews with only 2 latent dimensions/components.

- Very low overlap and the two latent dimensions are quite separable.



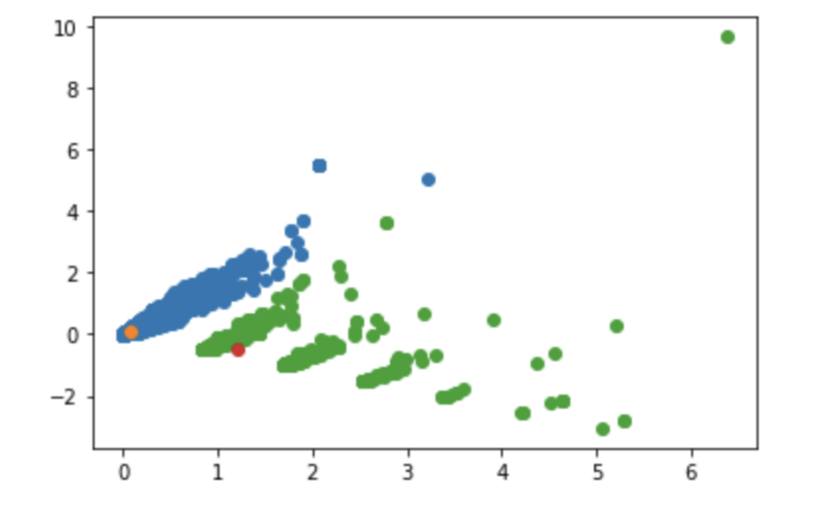
Negative Reviews(K-Means)

- Negative reviews contain more latent dimensions/components that contain very low overlap.

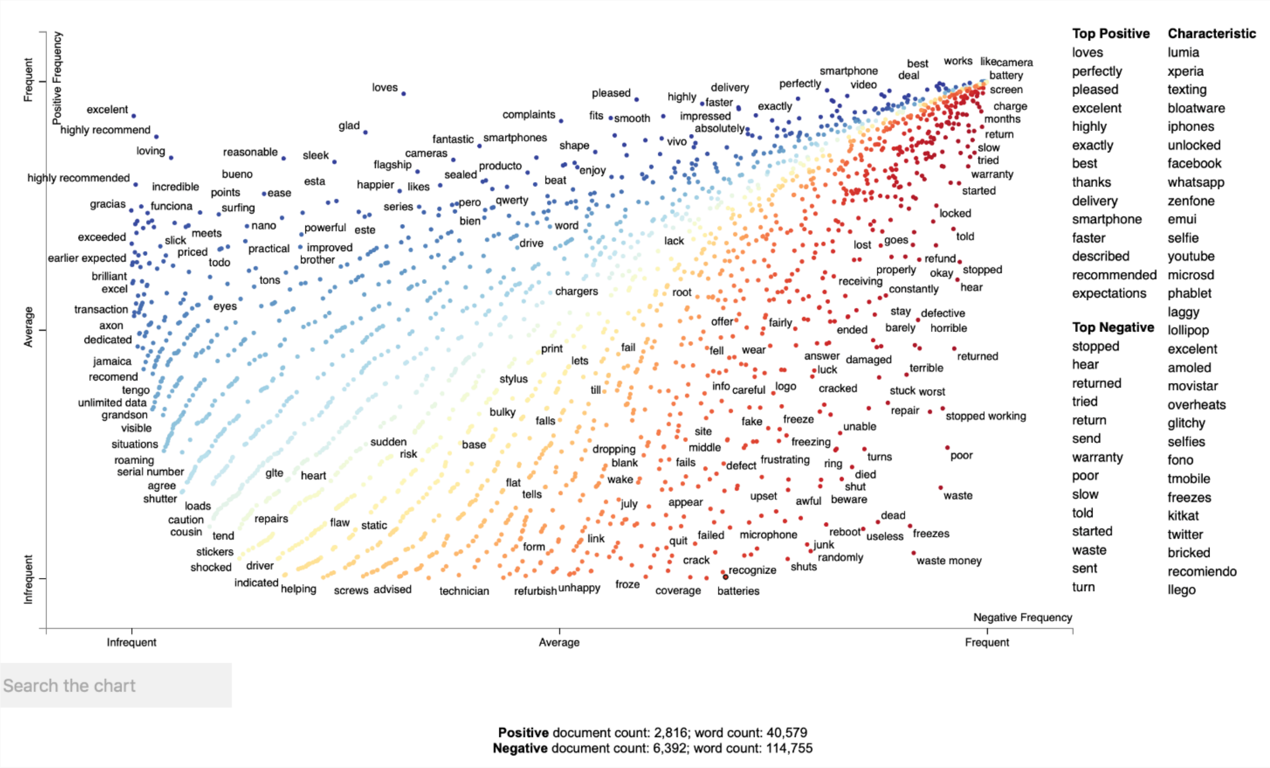


Negative Reviews (K-Means)

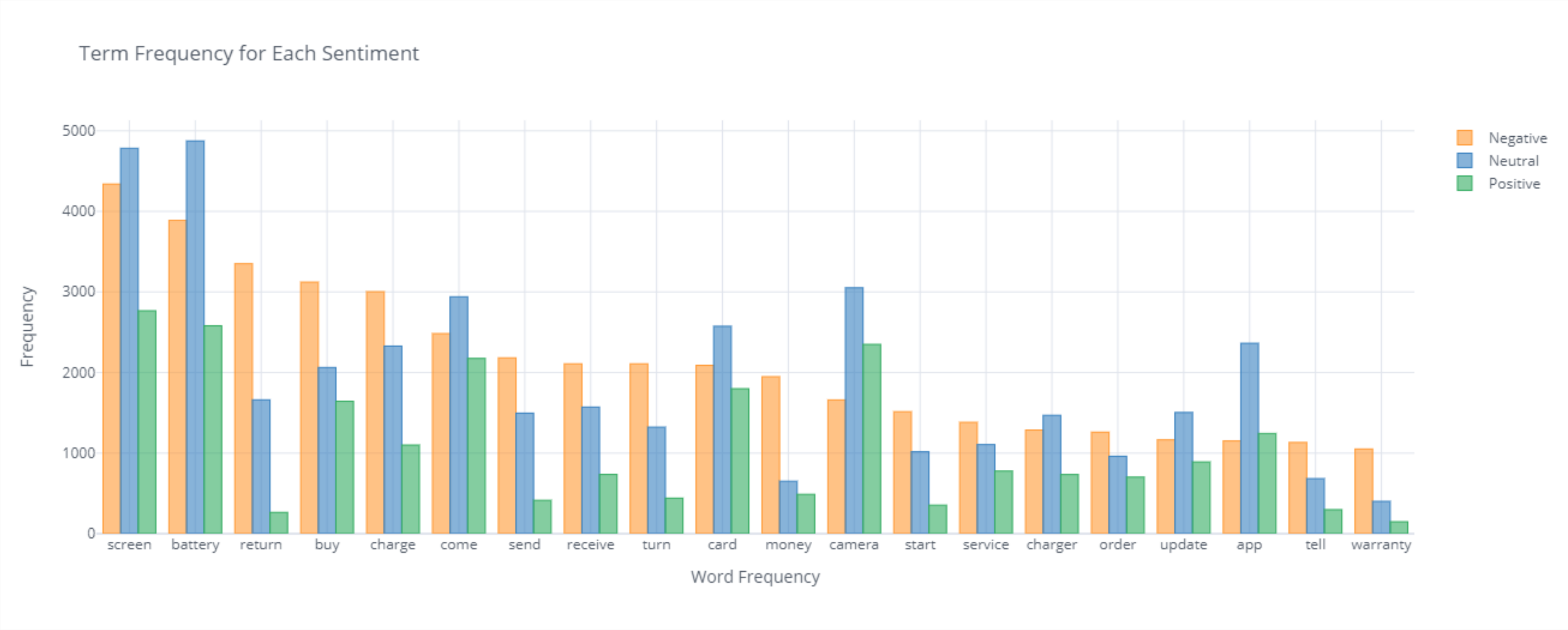
- Negative reviews could be fully represented by only two latent dimensions/components.



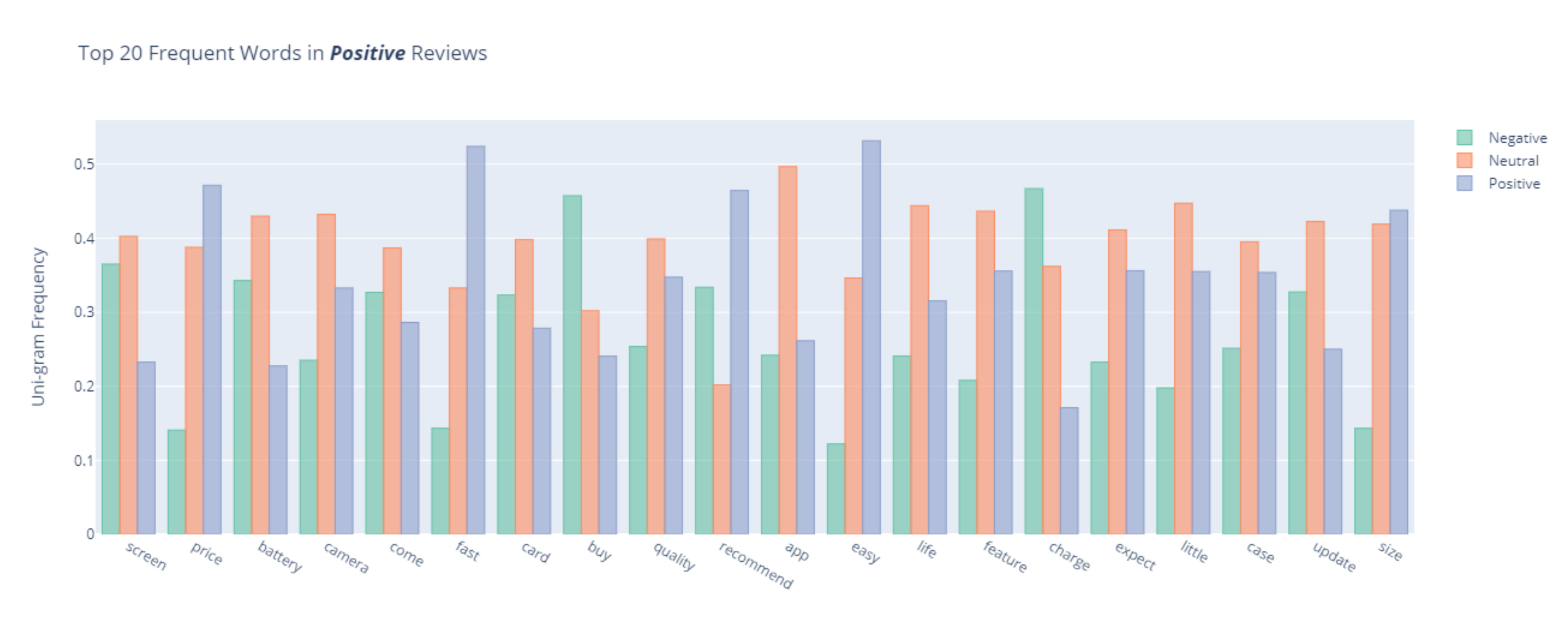
Scatter Text



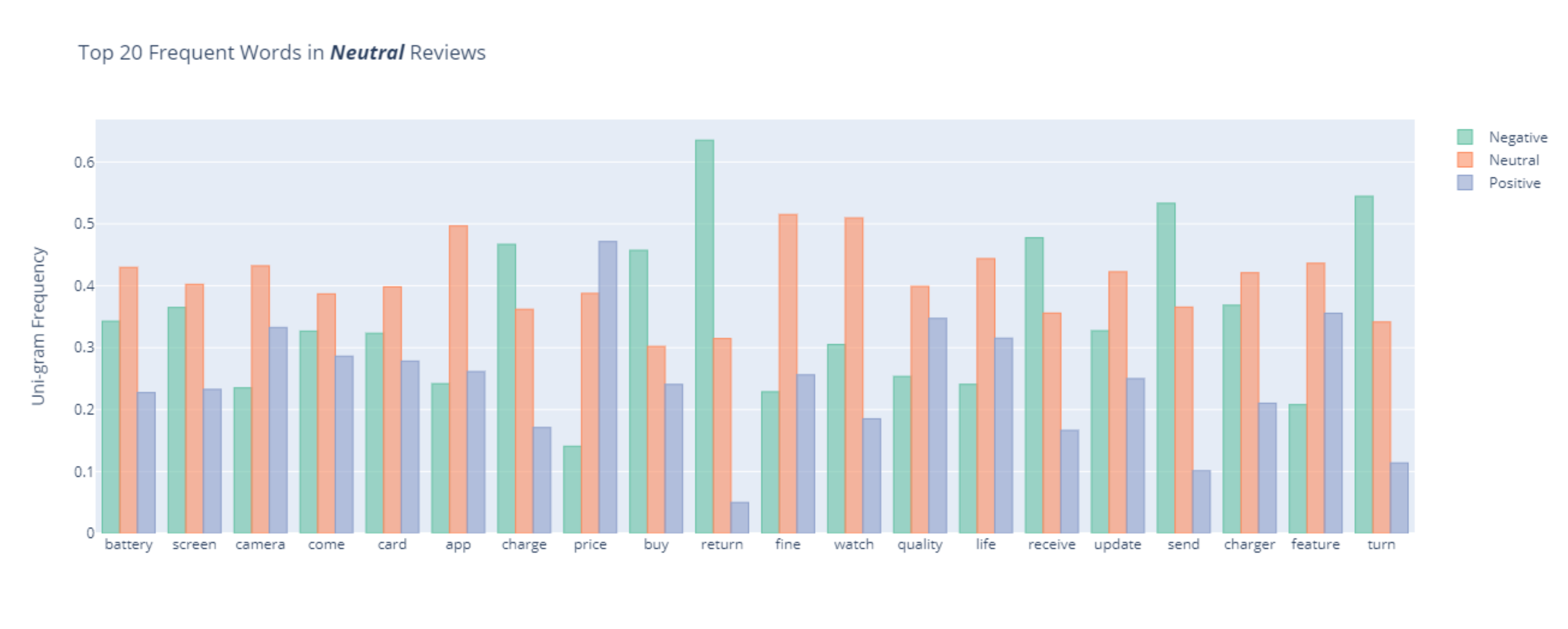
Term Frequency for Each Sentiment



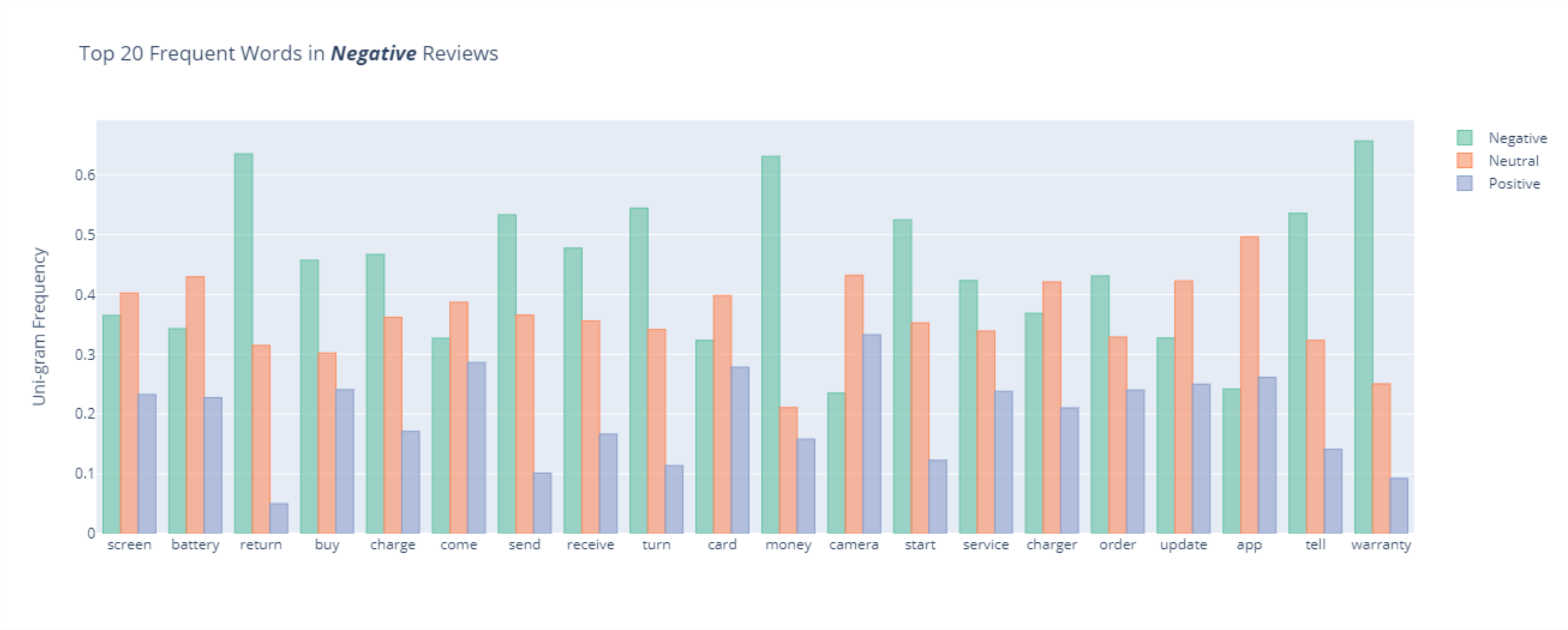
Uni-gram Frequency (Positive Reviews)



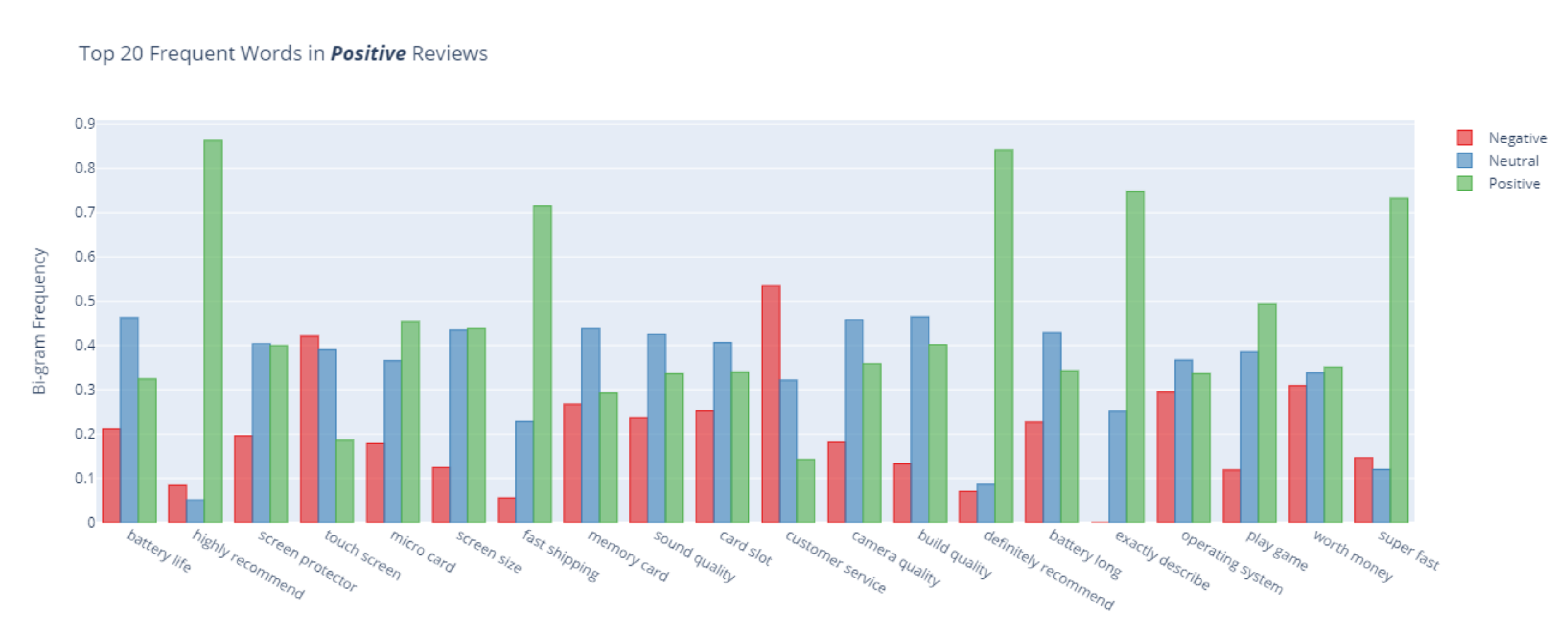
Uni-gram Frequency (Neutral Reviews)



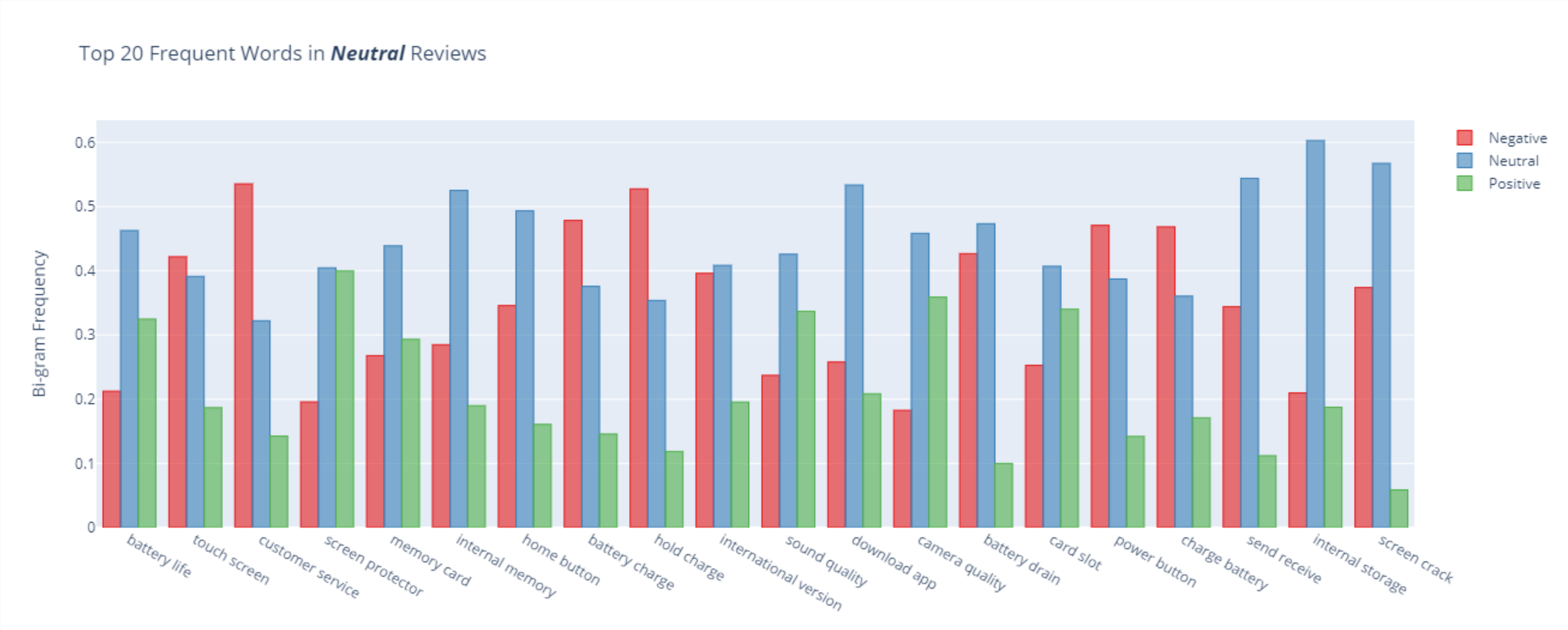
Uni-gram Frequency (Negative Reviews)



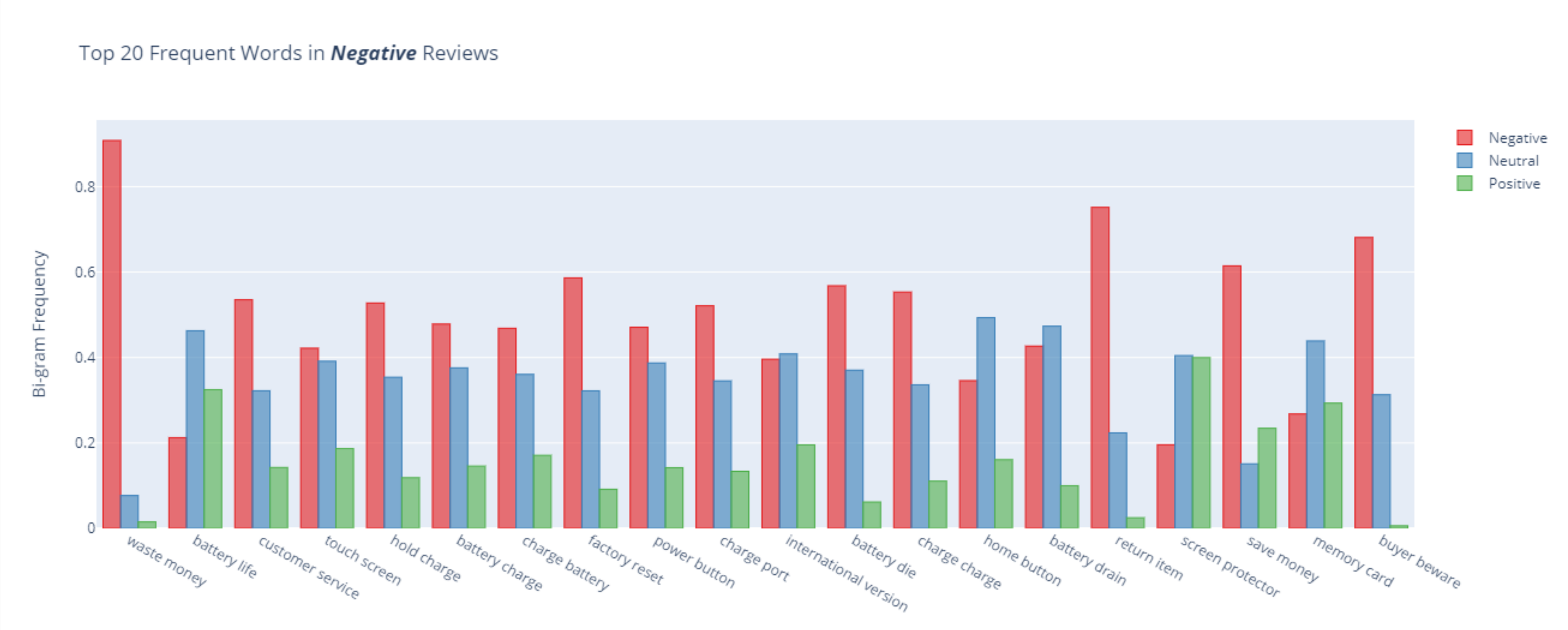
Bi-gram Frequency (Positive Reviews)



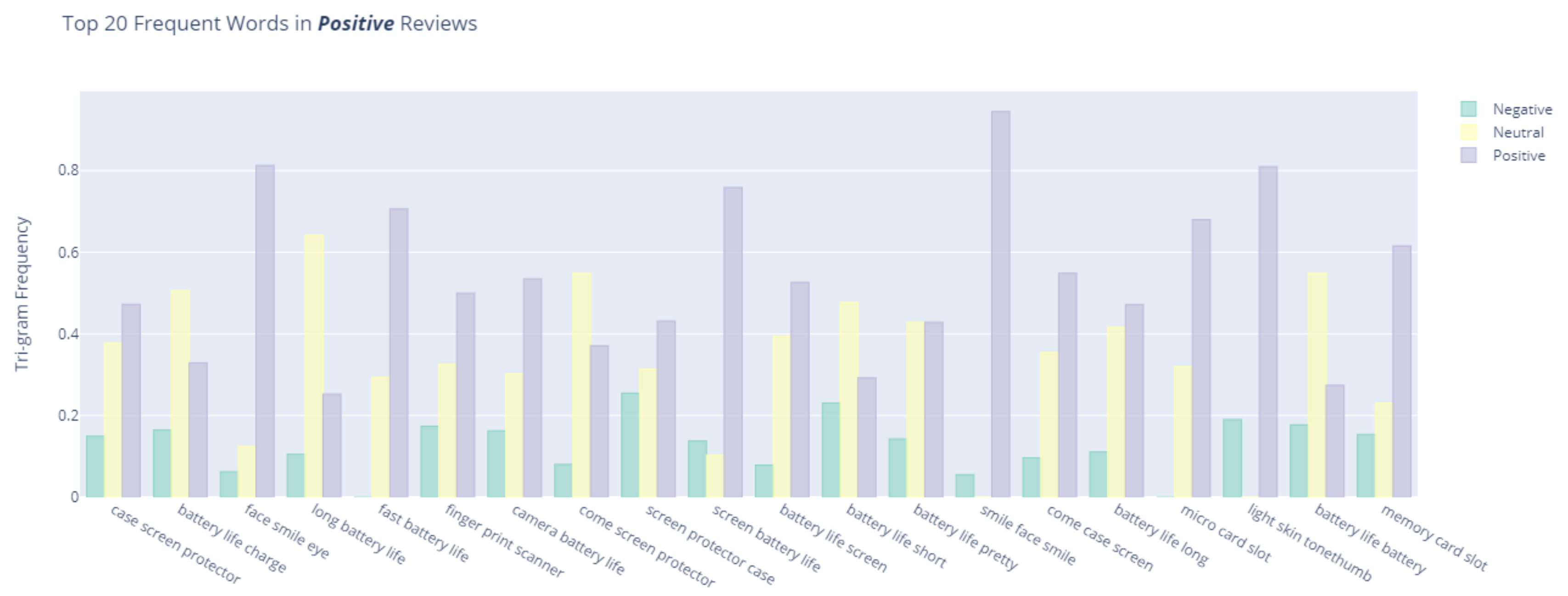
Bi-gram Frequency (Neutral Reviews)



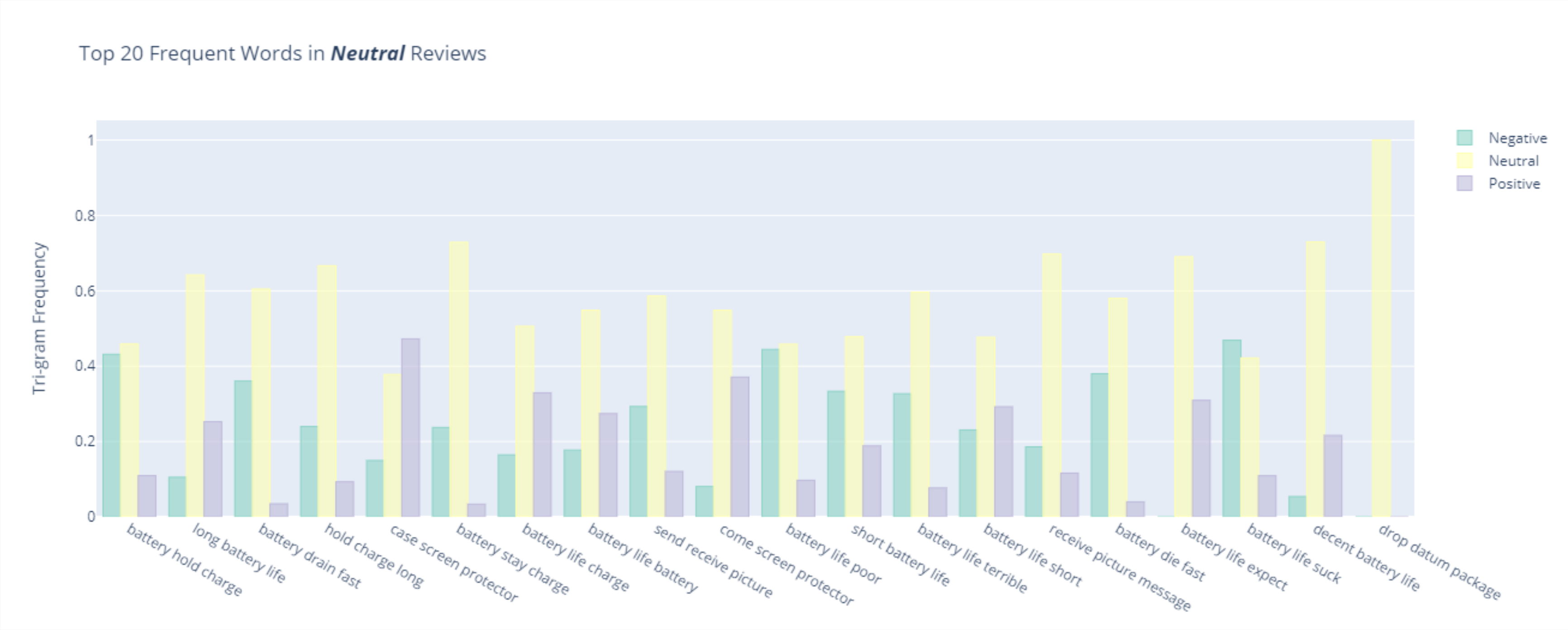
Bi-grams Frequency (Negative Reviews)



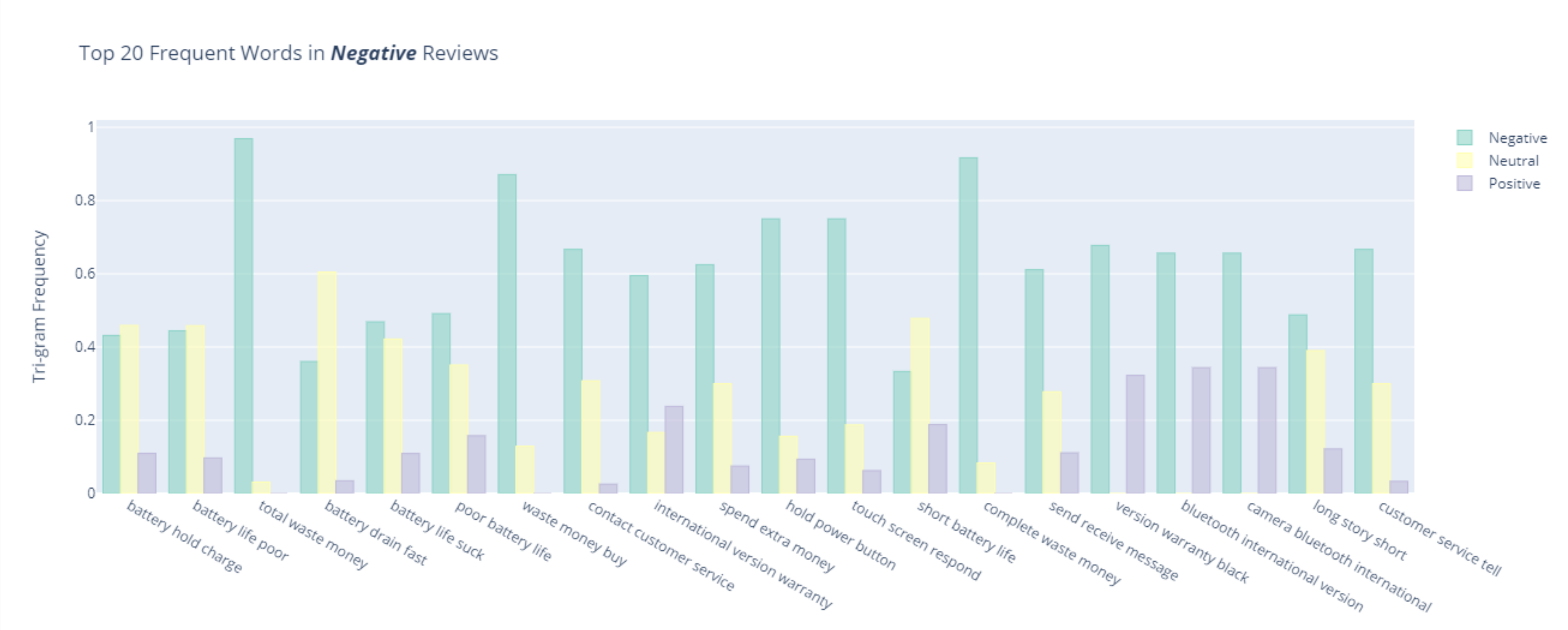
Tri-gram Frequency (Positive Reviews)



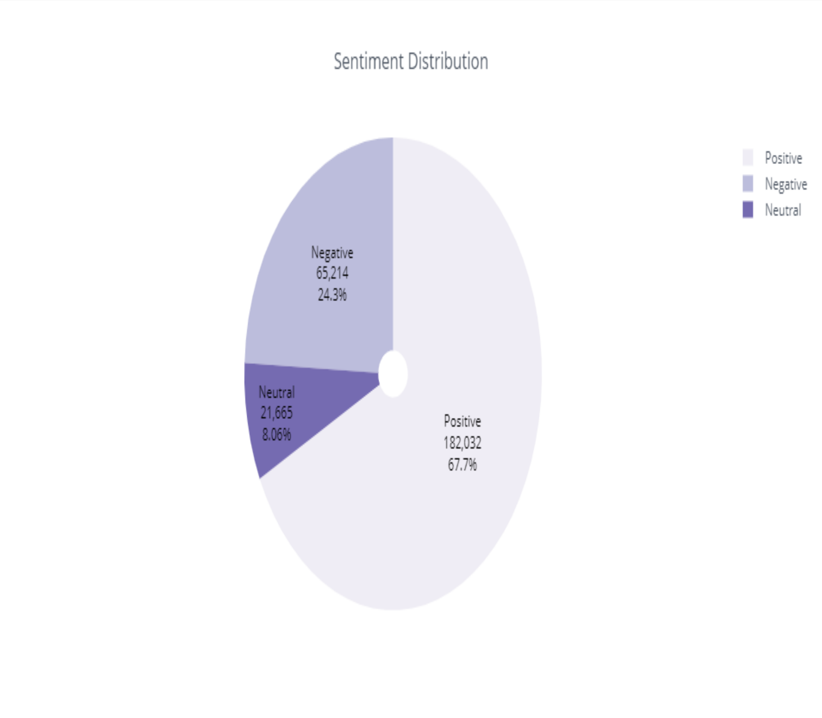
Tri-gram Frequency (Neutral Reviews)



Tri-grams Frequency (Negative Reviews)



Data before and after sampling



Classification

Logistic regression	Accuracy	precision	Recall	F1-score
Train	.76	.80	.79	.79
Val	.69	.69	.69	.68
Train-no noise	.76	.77	.76	.76
Val –no noise	.68	.68	.68	.67
Train-ngrams	.85	.86	.85	.85
Val-ngrams	.71	.72	.71	.71
Train-Trigrams	.86	.87	.86	.86
Val-Trigrams	.72	.72	.72	.71

Classification

SGB Classifier	Accuracy	precision	Recall	F1-score
Train	.74	.75	.74	.73
Val	.67	.67	.67	.66
Train-no noise	.71	.73	.71	.70
Val-no noise	.64	.65	.64	.63
Train -ngrams	.69	.71	.69	.68
Val-ngrams	.69	.64	.64	.63
Train-unigrams	.79	.81	.79	.79
Val-unigrams	.69	.70	.69	.68

Classification

Niave Bayes	Accuracy	precision	Recall	F1-score
Train	0.64	.66	0.64	0.64
Val	0.59	.61	0.59	0.59
Train-no noise	0.61	0.63	0.61	0.61
Val-no noise	0.58	0.59	0.58	0.57
Train -ngrams	0.67	0.70	0.67	0.67
Val-ngrams	0.61	0.63	0.61	0.60
Train-unigrams	0.65	0.70	0.65	0.65
Val-unigrams	0.59	0.63	0.59	0.58

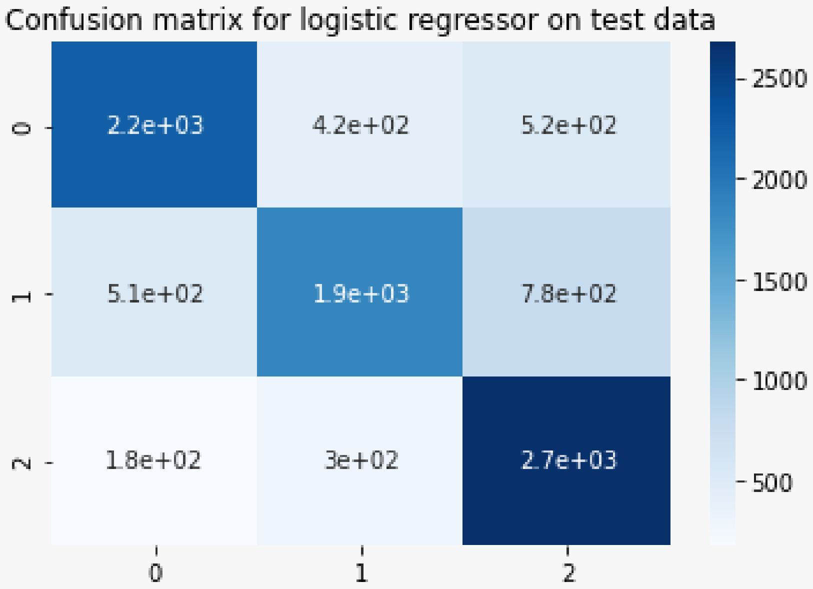
- The best Classifier is Logistic Regressor with Trigram:

Train Accuracy :

.86

Test Accuracy:

.71



Future work

- Enhancing machine learning pipeline for better NLP text processing.
- Trying stacking classifiers and ensemble classifiers on reduced NLP dataset.

Thank you for listening ☺