NLP

Natural Language Processing

Done by Group-1

Business Problem

Objective:

Our clients want to launch a new product and want to understand the reviews of the competitors to see what are the challenges faced . Hence want to improve the respective features to address the challenges. Thus the newly launched product should have enhanced features.

Our approach consists of steps as following:

- Extraction or collection of the reviews of the competitor's product from different platforms
- Pre- processing the data: Clean and transform the data
- Sentiment analysis , positive and negative feature extraction.

Abstract

The analysis of product review sentiment analysis plays a major role in improving quality ,reducing Complaints about the product and also suggest client that to maintain the product Quality under situation of present market. Advance Machine learning techniques plays a vital role in getting accuracy of product reviews in dealing with Non linear Complex situation Instead of Traditional statistical methods. In the present study we made an attempt to get an accuracy of product reviews through Naïve bayes ,Support vector,K- fold classifications were calculated on training and testing data set(Tf-idF,polarity).The results revealed that SVR gave good accuracy compared to other model to forecast the data in future.This study will provide a policy maker with information to help clients on developing Product quality To assist the company in implementing new product policy that favour customer and business industries.

Product - Air Fryer

- Air Fryer is a small counter top convection oven designed to stimulate deep frying without submerging the food in oil.
- Air Fryer is similar to an oven in the sense that it bakes and roasts, but the difference is its heating elements are only located on top and are accompanied by a large, powerful fan, resulting in food that’s super crispy in no time and most notably, with less oil than deep- fried counterparts.

Extraction of Data

Data is extracted with the help of Amazon Review Exporter.

Brands

Brands included in Analysis

PHILIPS

HAVELLS

AGARÓ

American Micronic

SOLARA
QUALITY DRIVEN, CUSTOMER OBSESSED

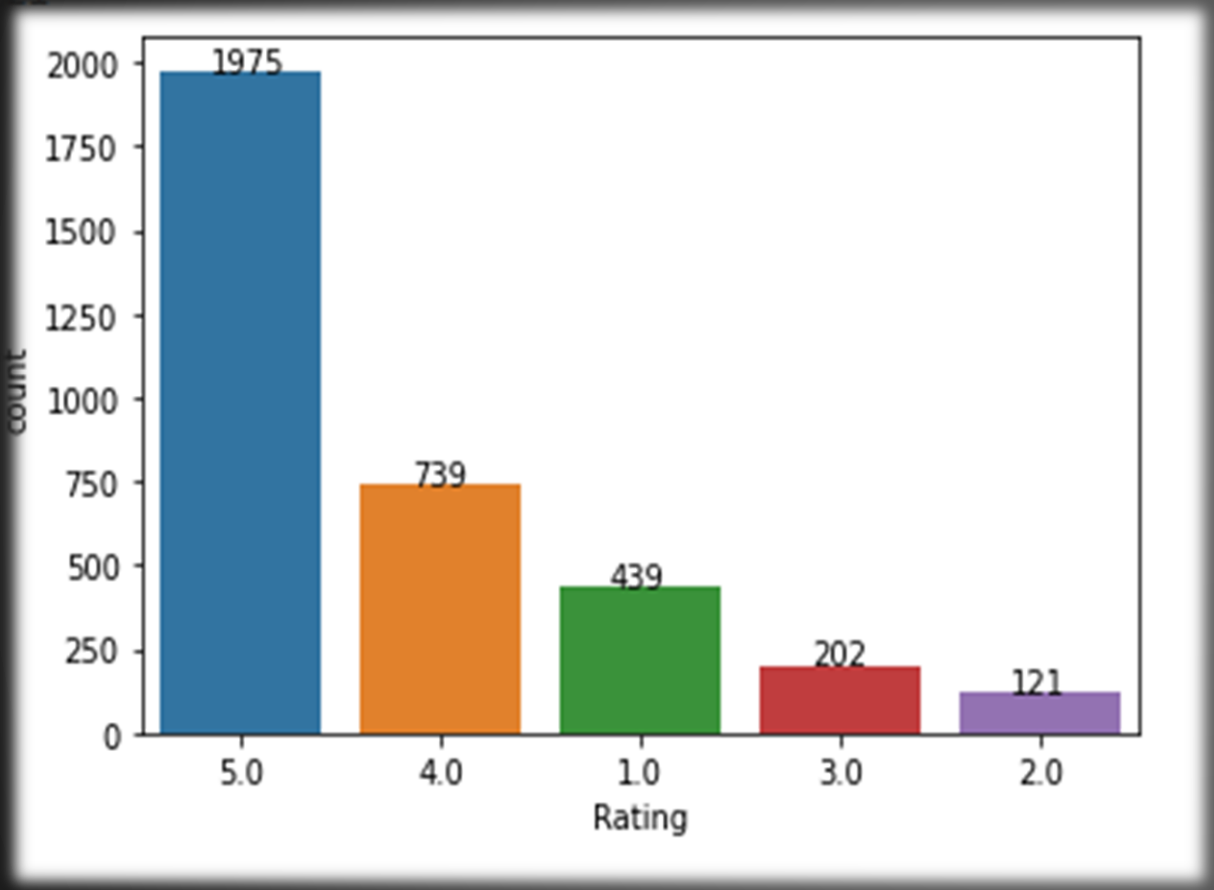
amazon

VARADA

KENSTAR
we love you

INALSA
HOME APPLIANCES

Bar Plot

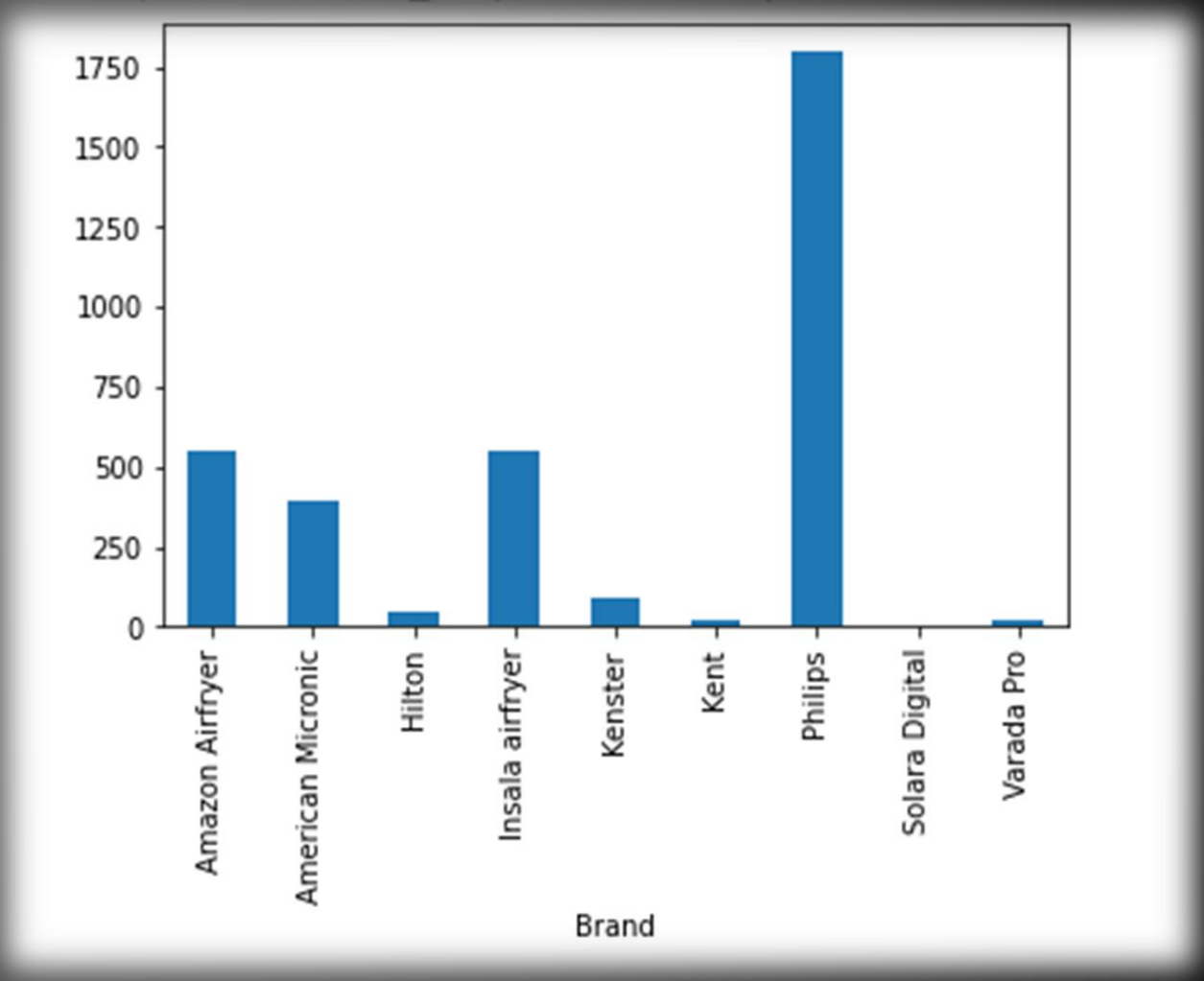


Total Entries - 3478

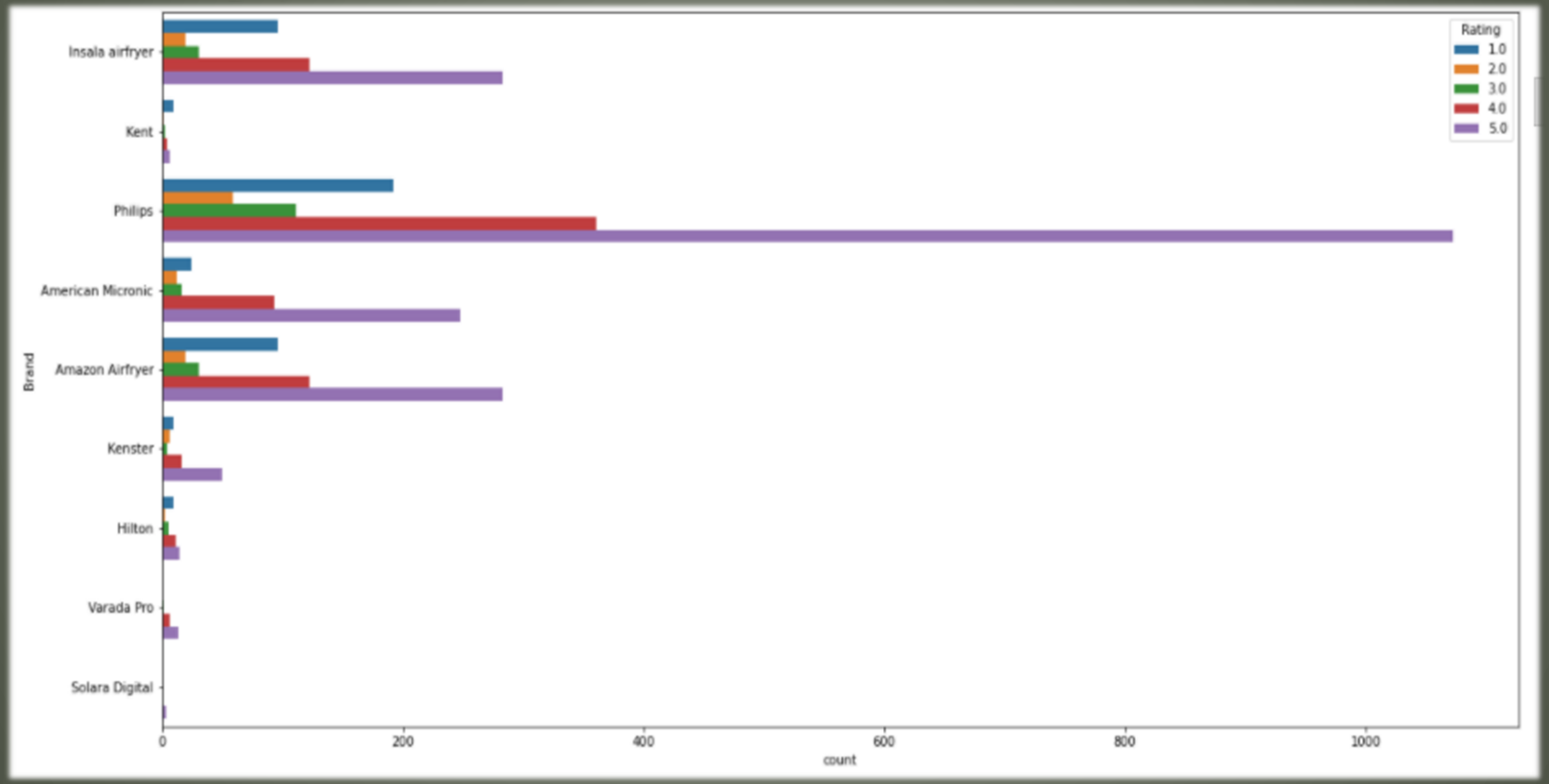
- Highest reviewed rating is 5.0 with 1975 reviews
- Lowest reviewed rating is 2.0 with 121 reviews

Brand Wise Reviewed Bar Plot

- Plot shows Phillips have highest number of reviews.
- Amazon and Insala are equal in its competition.
- Kent and Varada pro are least reviewed.



Brand Wise Ratings With Count



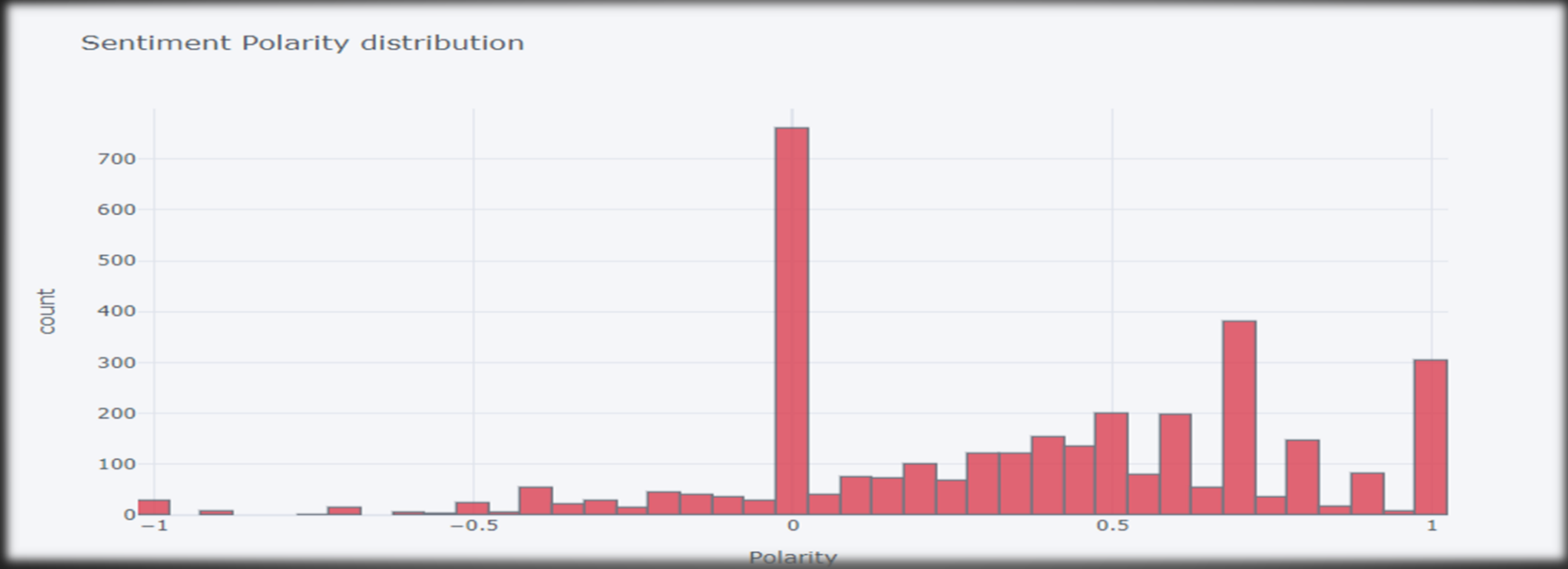
Sentiment Analysis(or opinion mining)

This is a NLP technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

Why Is Sentiment Analysis Important?

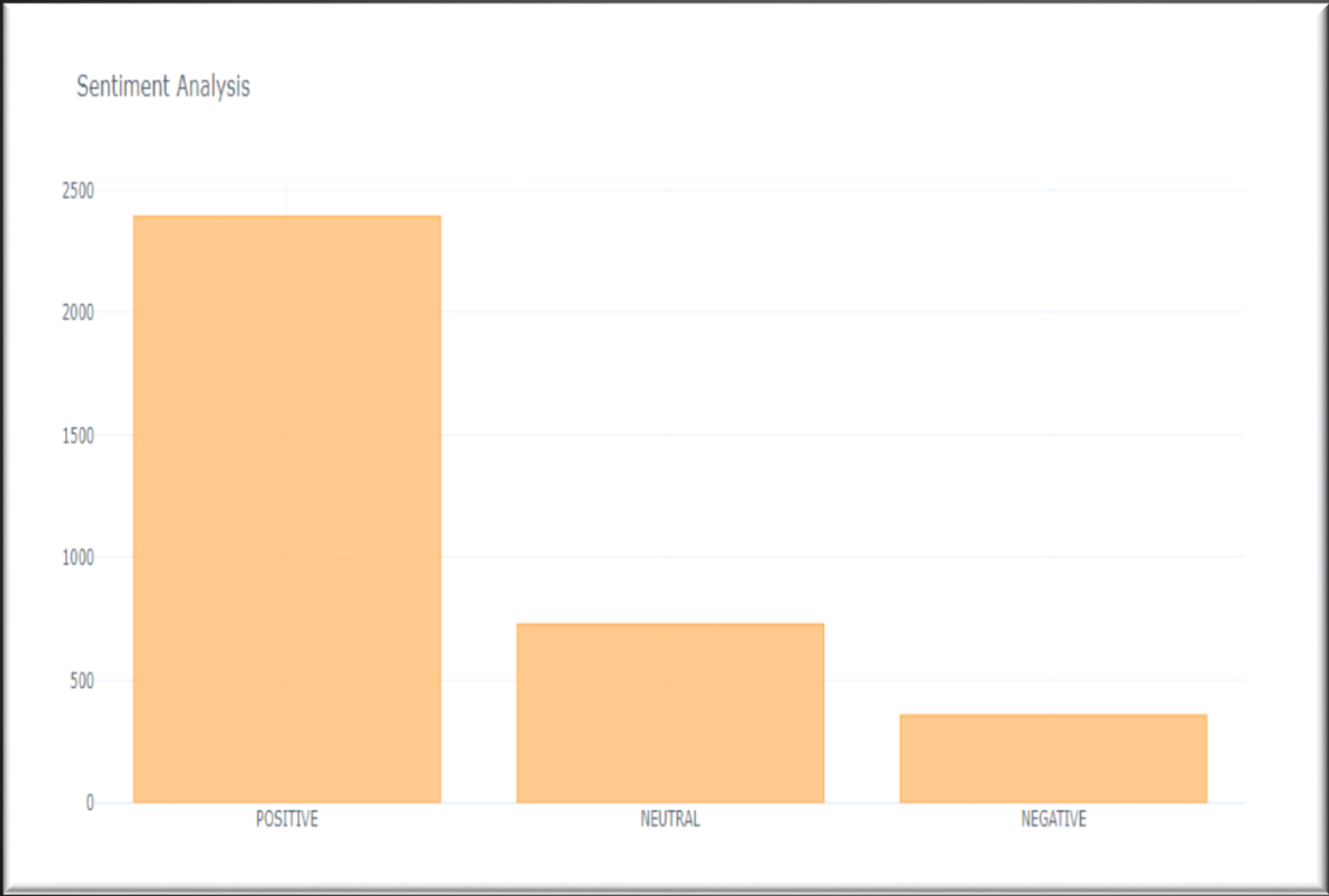
Sentiment analysis is extremely important because it helps businesses quickly understand the overall opinions of their customers. By automatically sorting the sentiment behind reviews, social media conversations, and more, you can make faster and more accurate decisions.

Distribution on Sentiment polarity



- Highest polarity value at 0 is (-0.025 to 0.0249),759 reviews
- -1(-1.025 to -0.9751),27 reviews 1(0.975 to 1.0249),303 reviews
- -0.5(-0.525 to-0.4751),24 reviews 0.5(0.475 to 0.5249),199 reviews

Sentimental Analysis

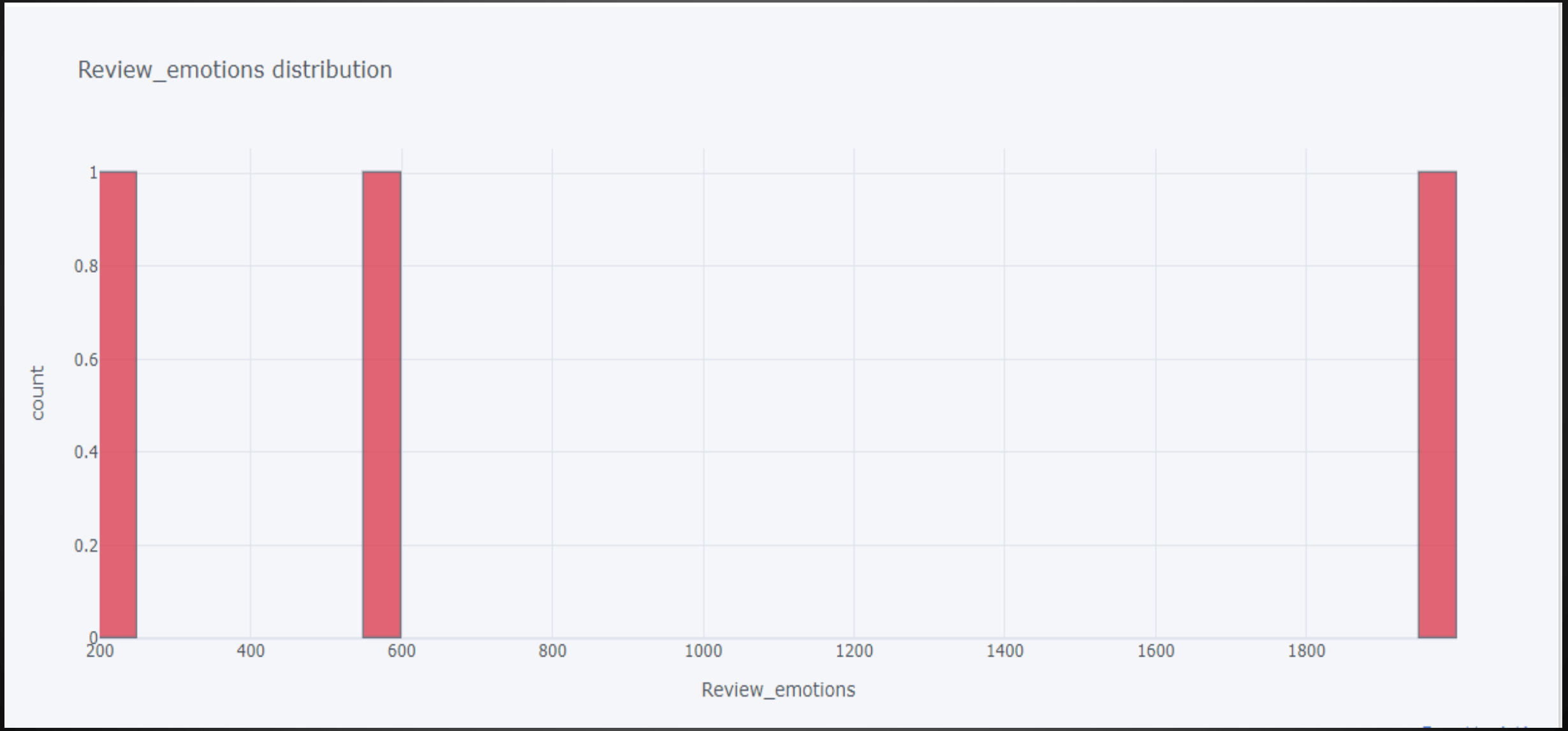


- Positive reviews-2391
- Neutral reviews-728
- Negative reviews-357

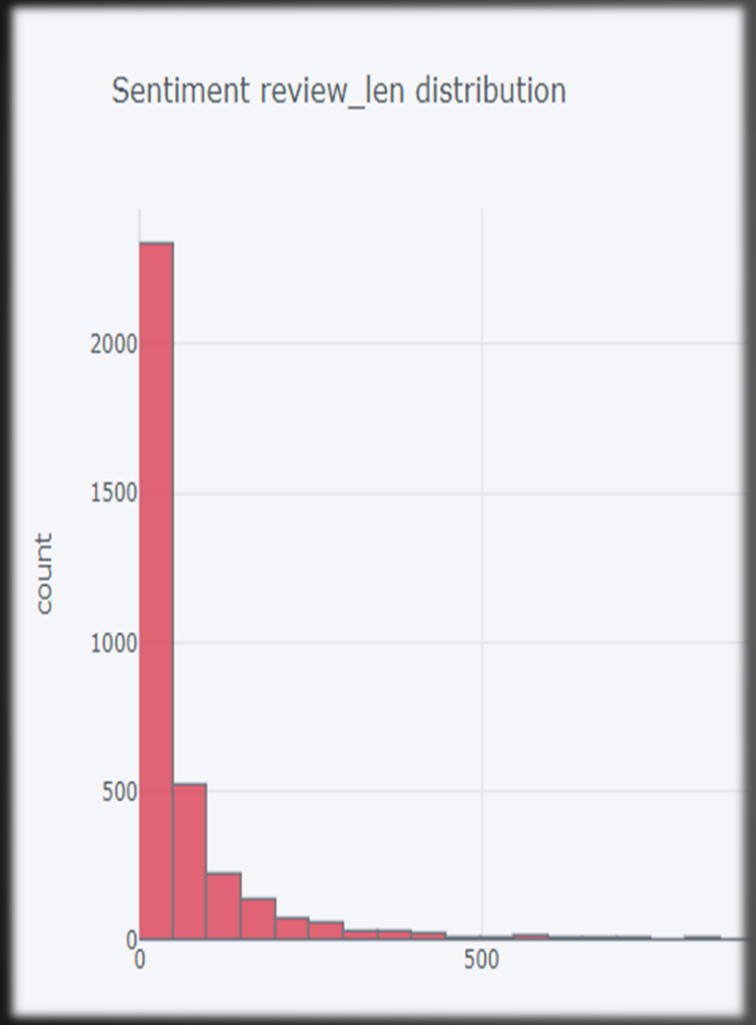
- Using sentiment analysis we can assume that majority of review is positive about the product on brands.

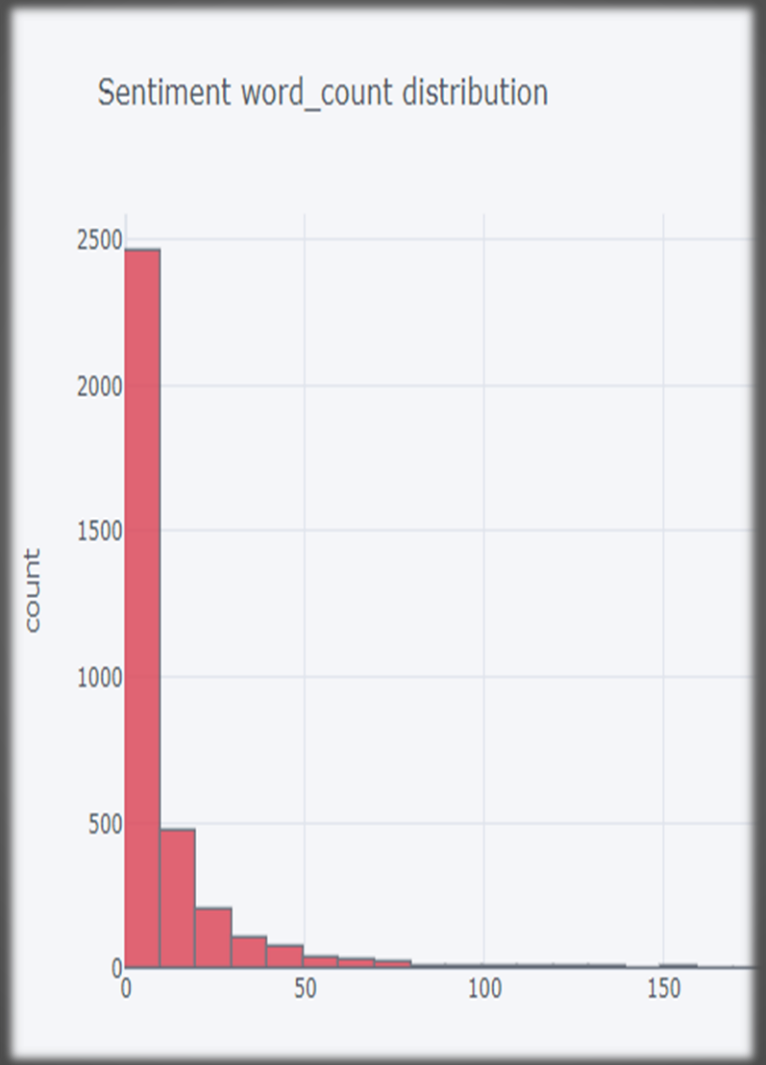
- 439 review on rating 1
- 121 reviews on rating 2
- 202 reviews on rating 3
- 739 reviews on rating 4
- 1975 reviews on rating 5

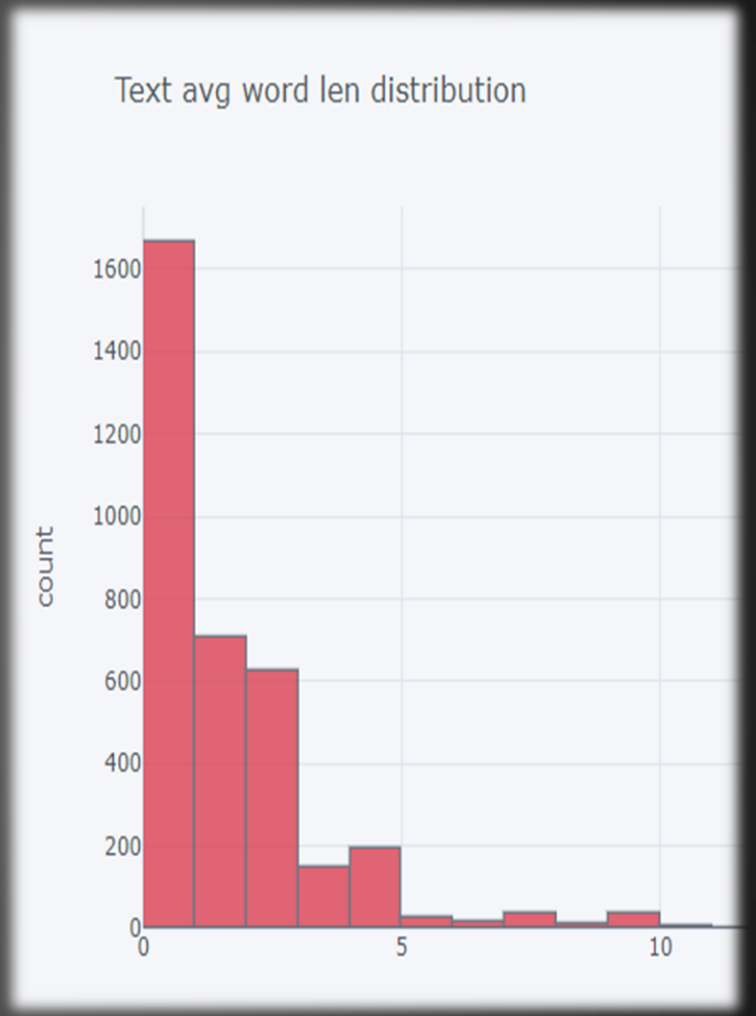
Emotion Distribution



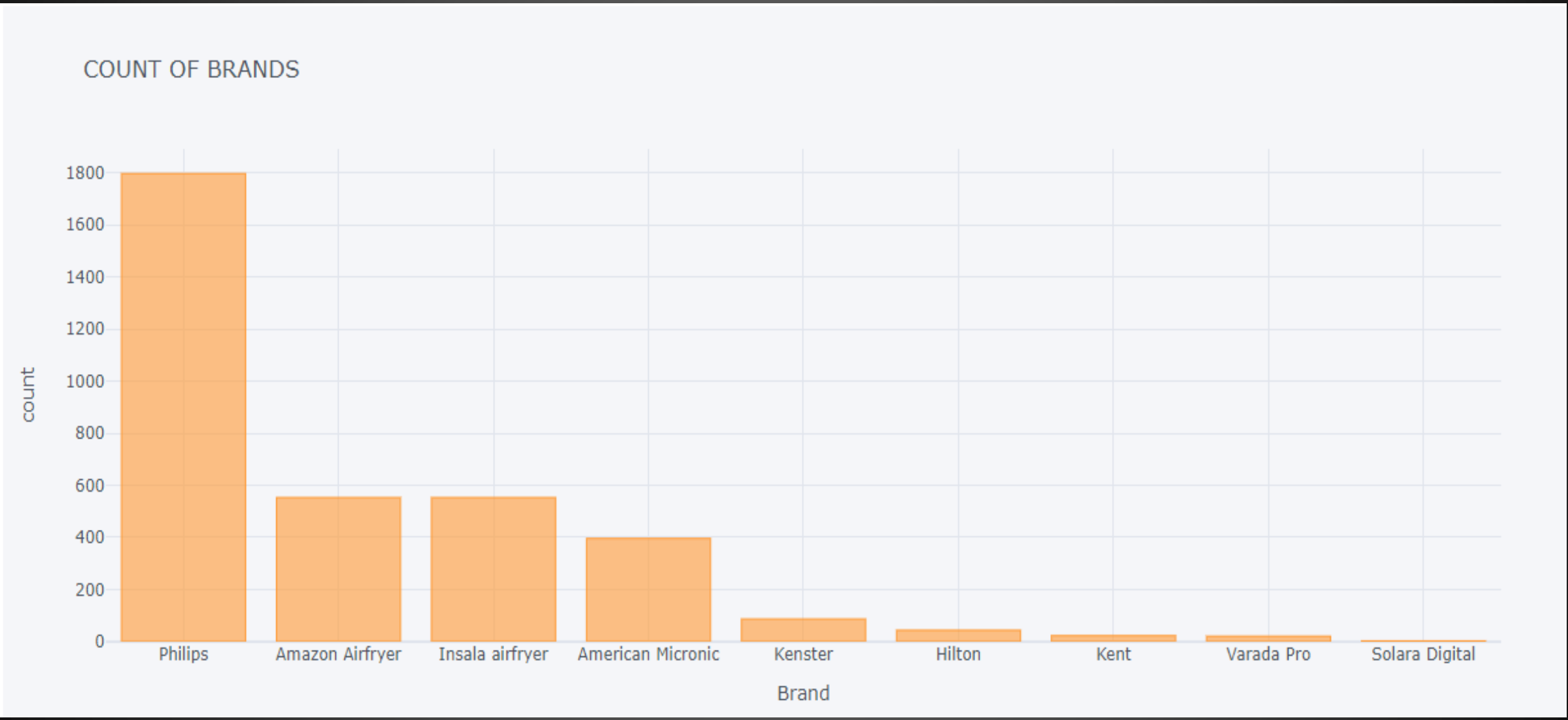
Distribution plot on Review length,Word count, Average word length





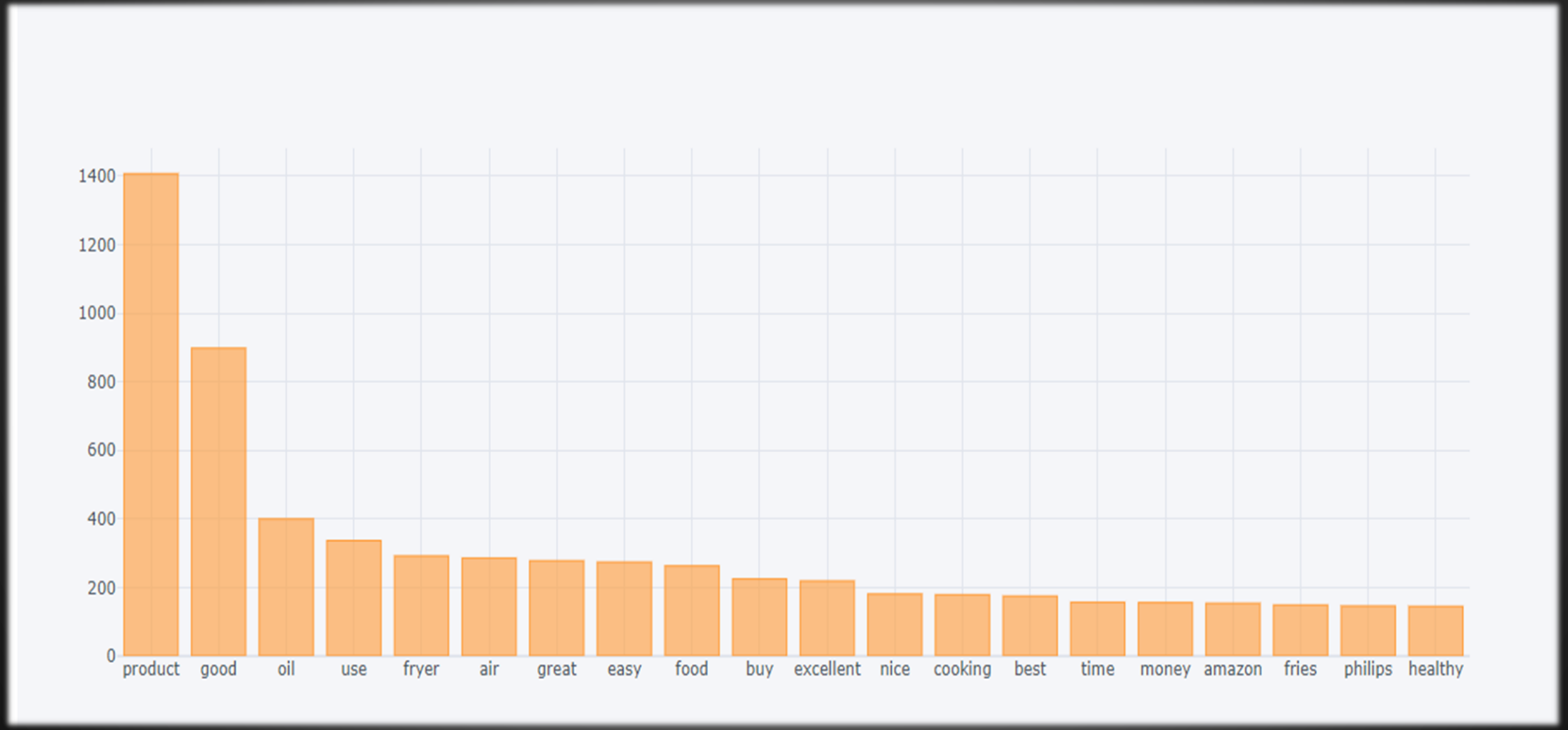


Observing the plot we can say Phillips stands the highest count, Amazon and Insala are equal, and the lowest count is by Solara digital.



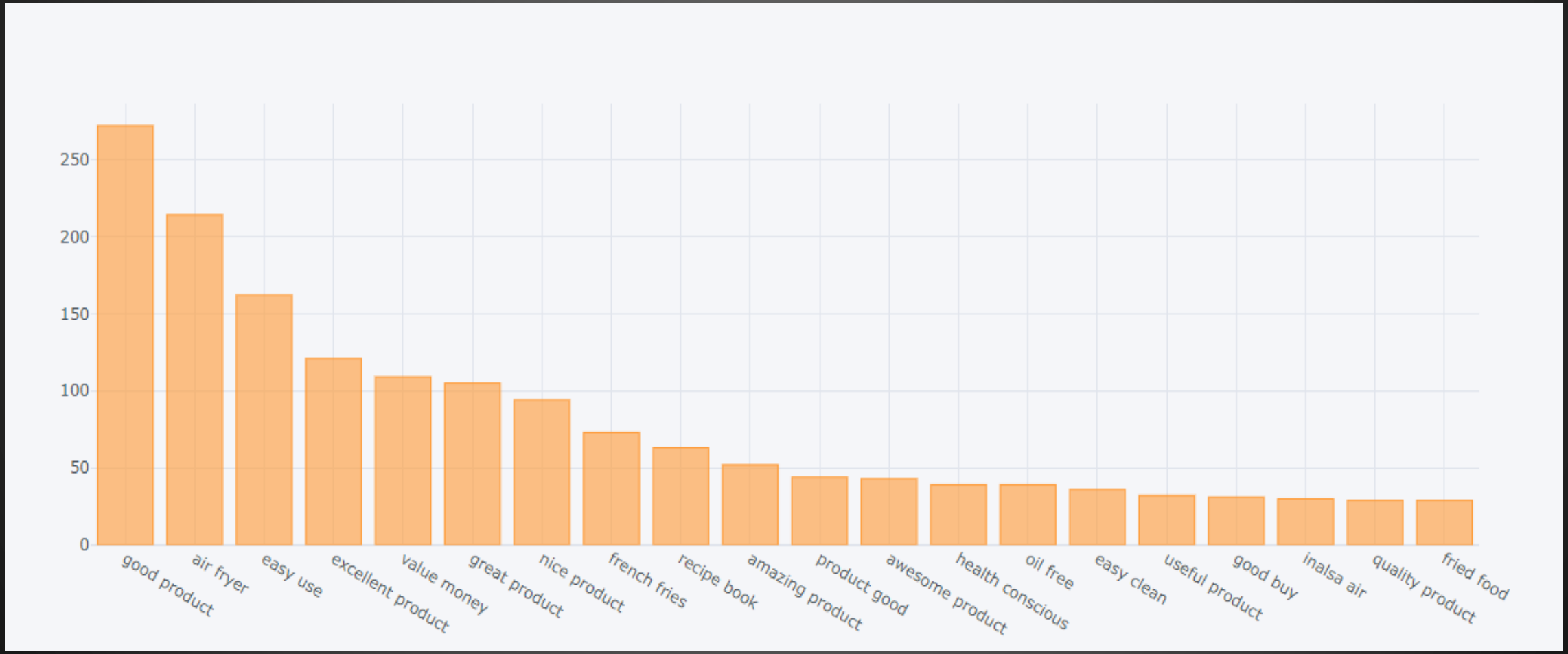
Unigram

The distribution of unigram after removing stop words



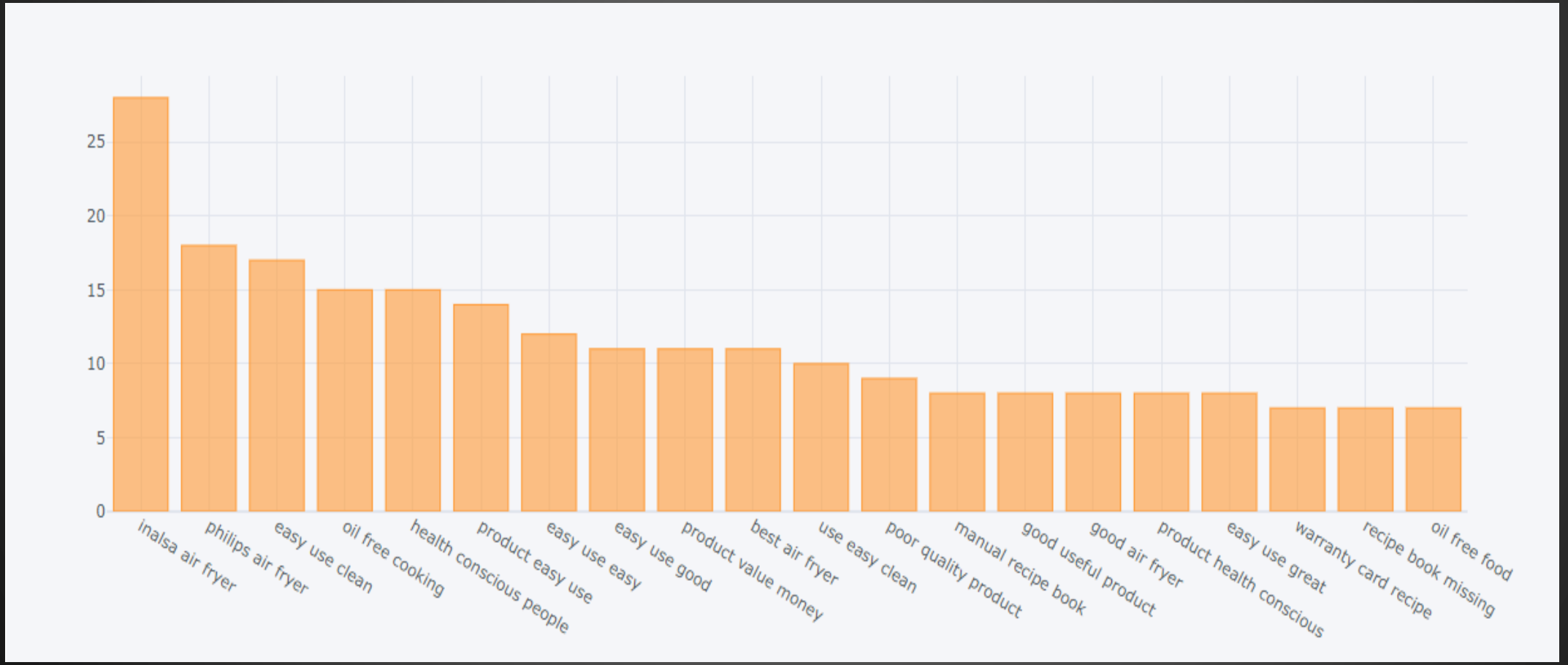
Bigram

Bigrams in review after removing stop words.



Trigram

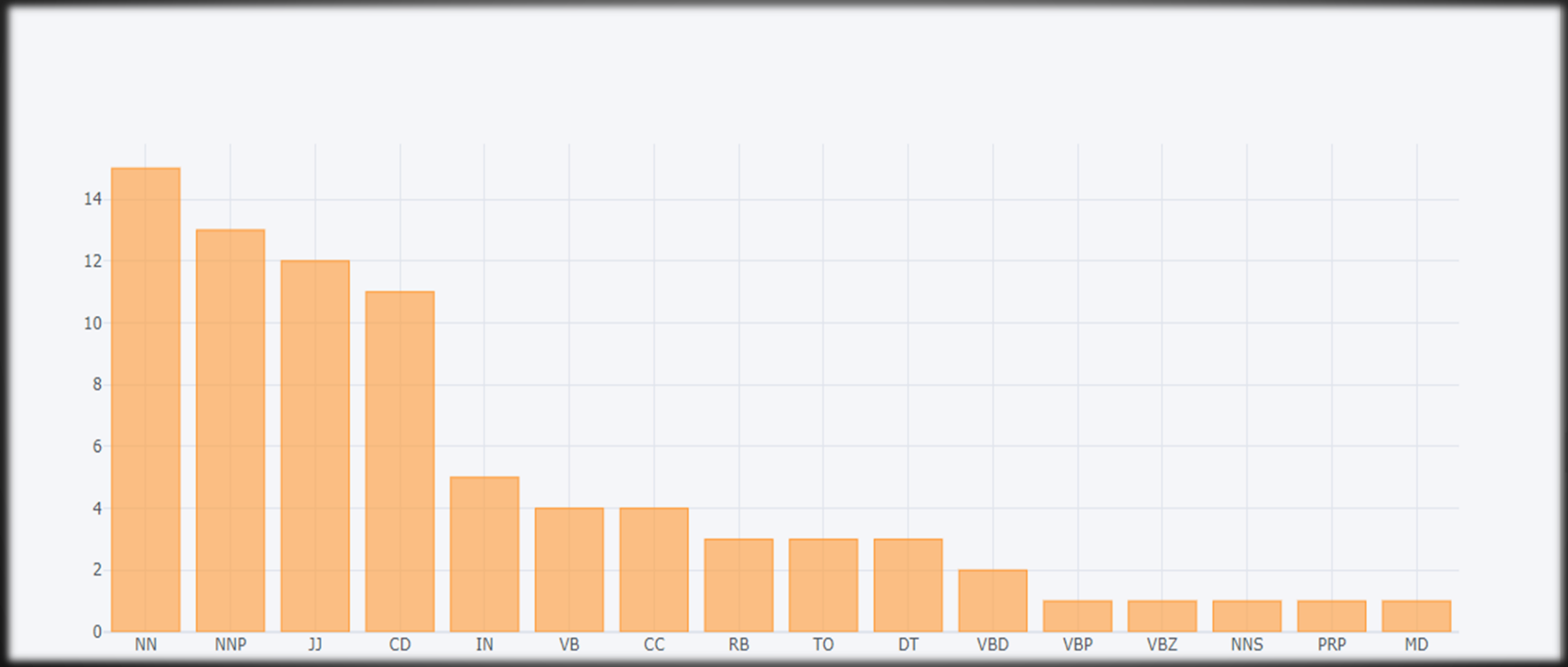
The distribution of top trigrams without stop words.



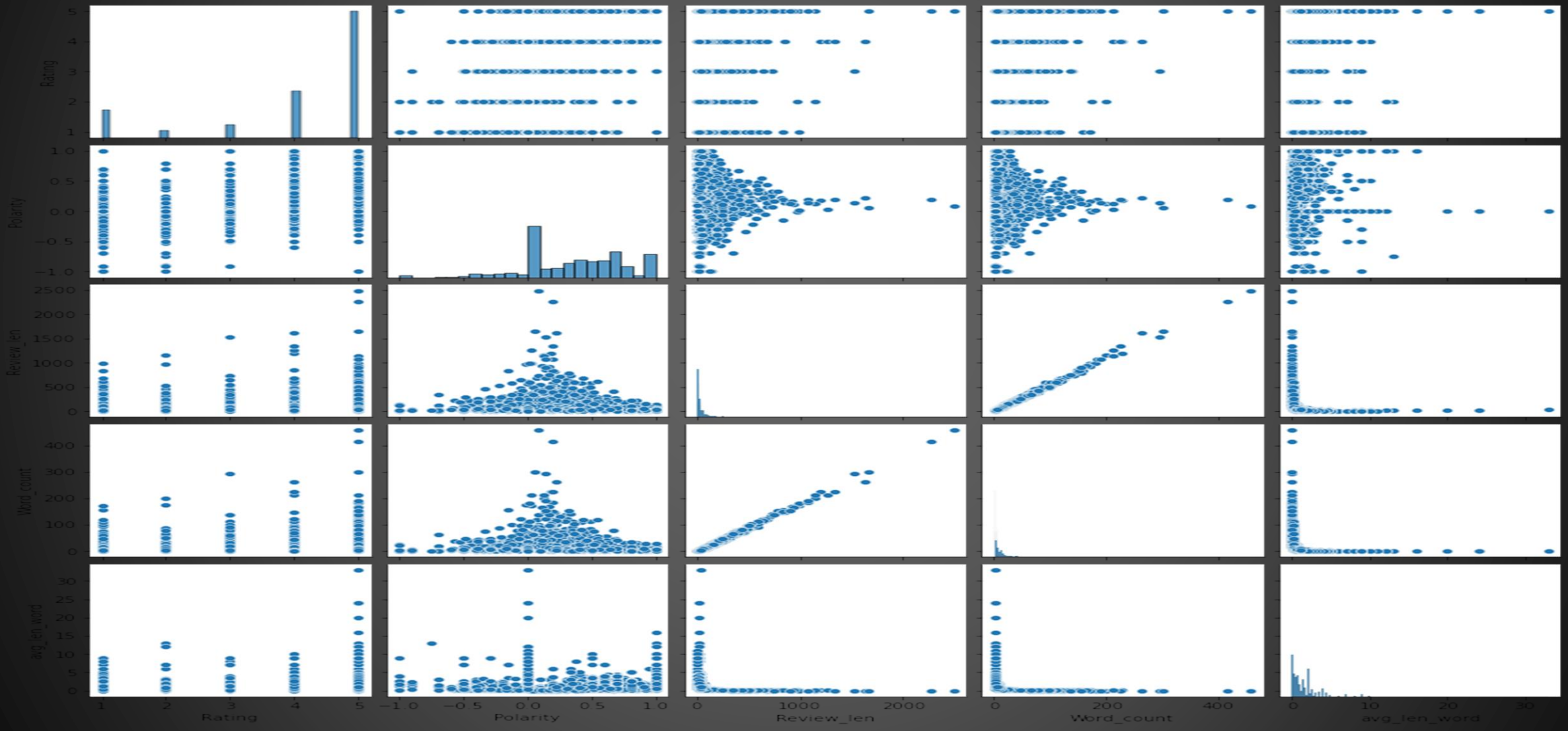
Part Of Speech Tagging

POS is a process of assigning parts of speech to each word, such as noun, verb, adjective, etc

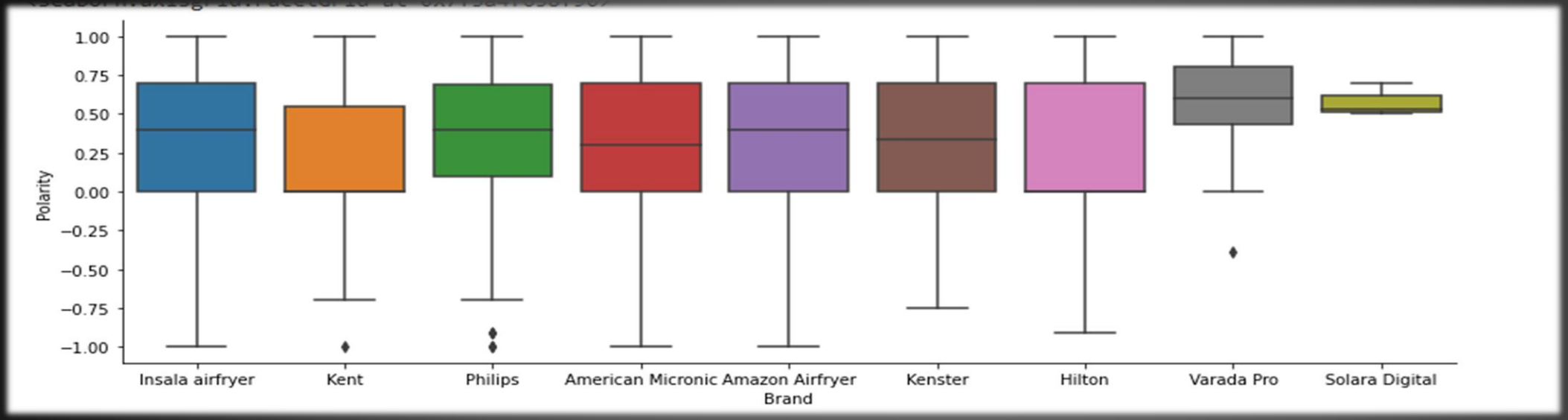
Distribution plot on POS from reviewed text.

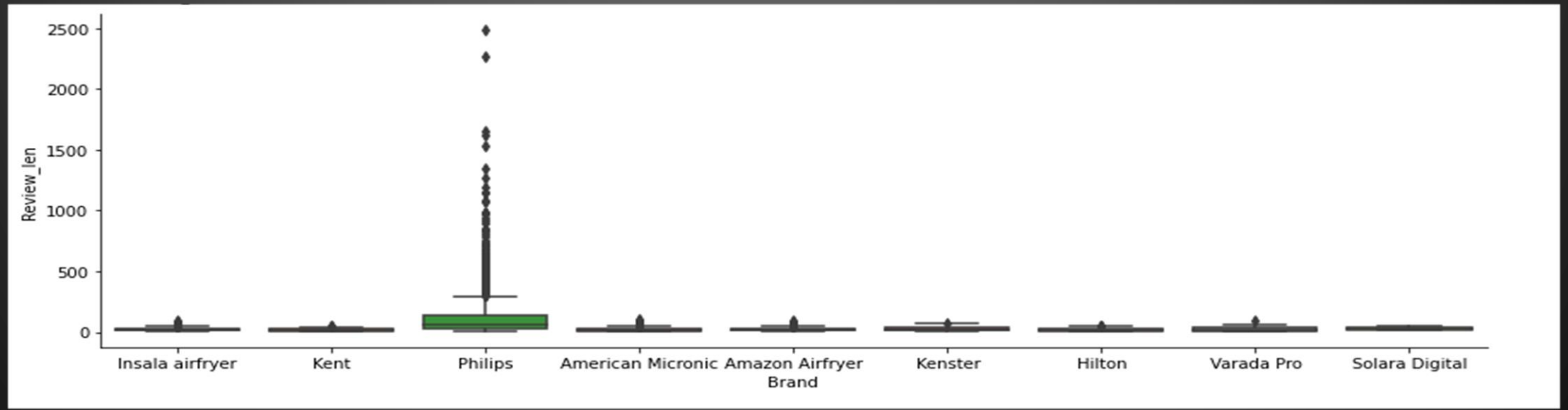


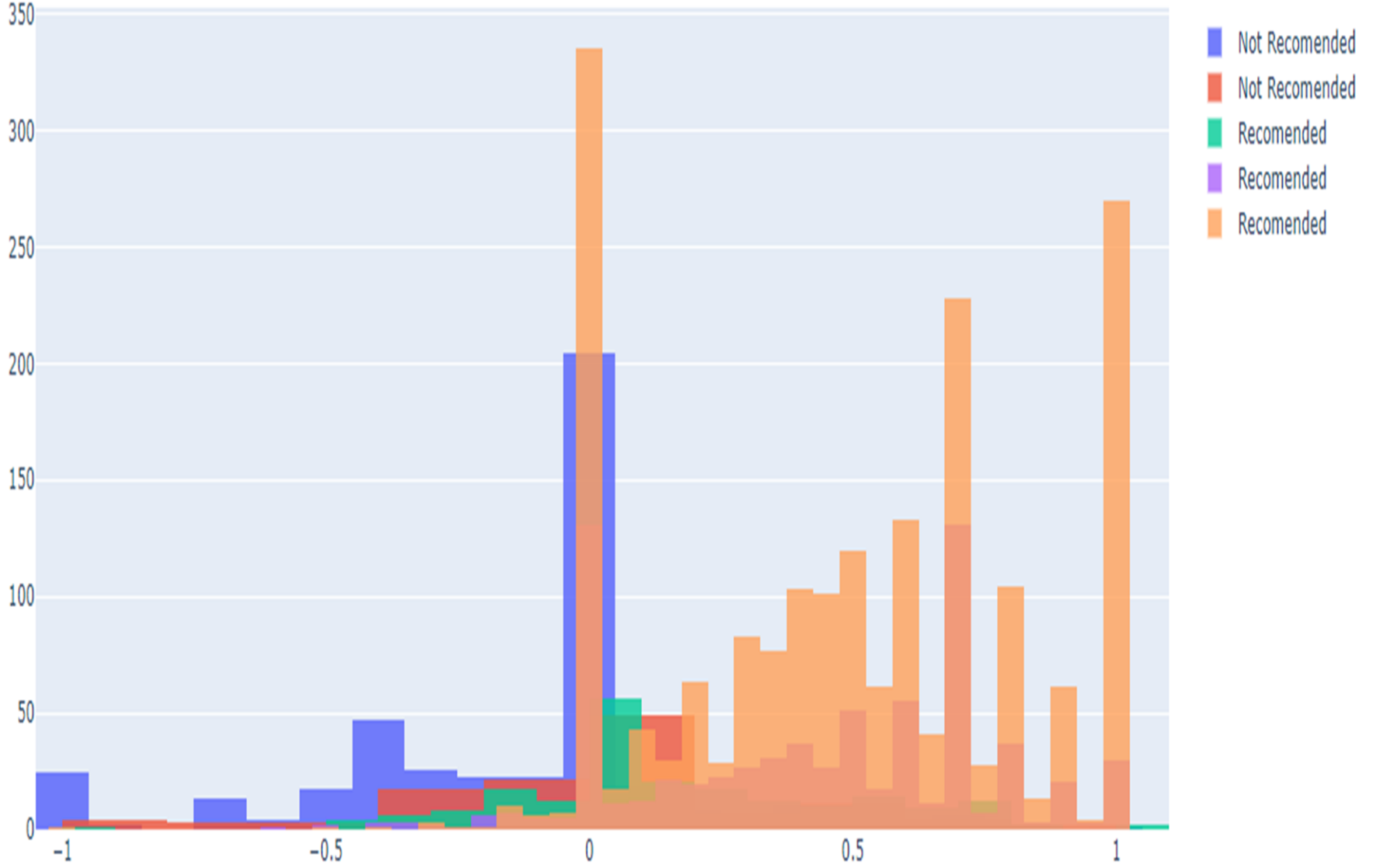
Pair Plot



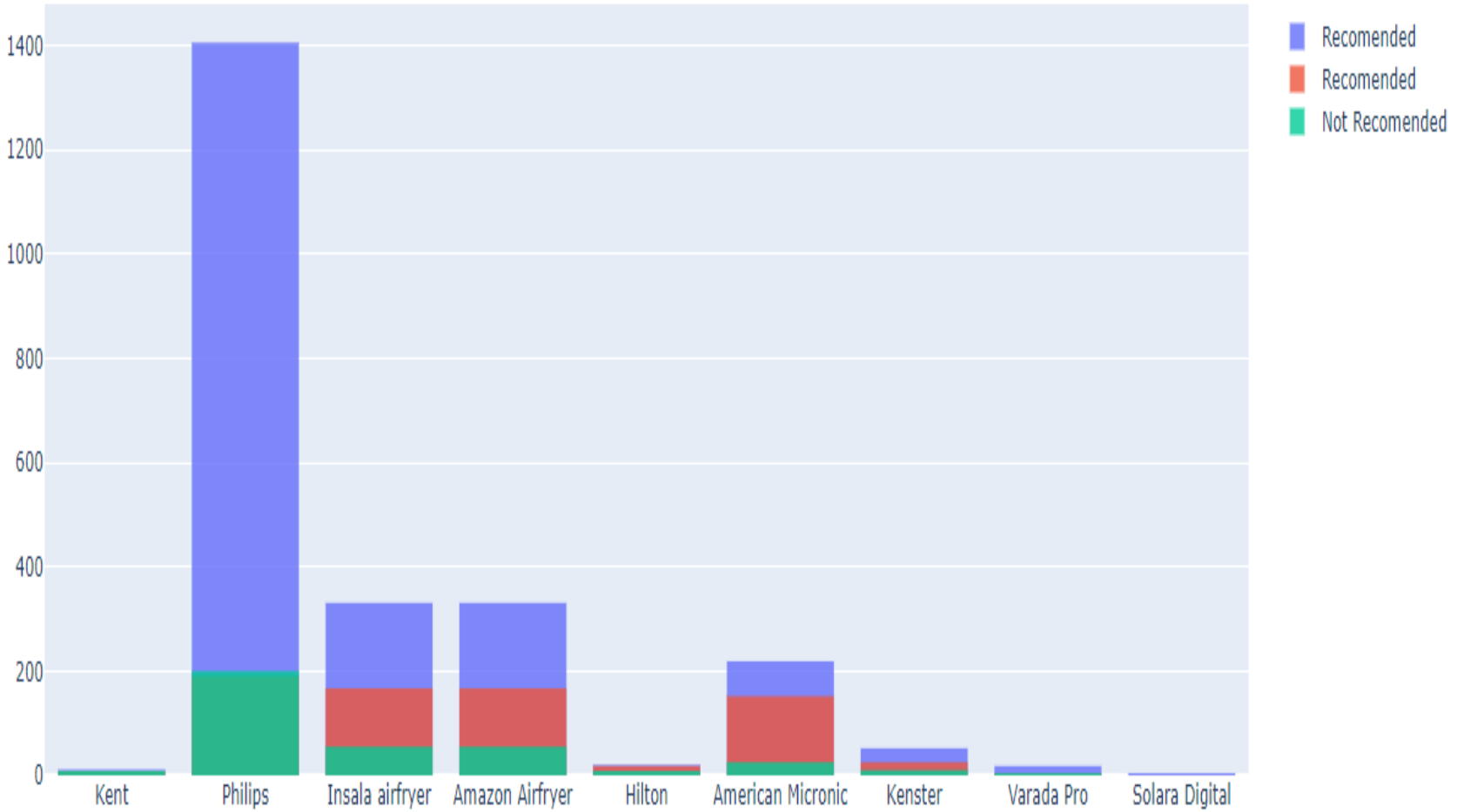
Box Plot


Polarity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Insala airfryer
Kent
Philips
American Micronic
Amazon Airfryer
Kenster
Hilton
Varada Pro
Solara Digital
Brand


Review_len
2500
2000
1500
1000
500
0
Insala airfryer
Kent
Philips
American Micronic
Amazon Airfryer
Kenster
Hilton
Varada Pro
Solara Digital
Brand



Based on rating and polarity
When rating is greater than or equal to 0 taken as recommended, less than 0 is considered as not recommended for further process.



Philips - 190,1406

Amazon - 55,167,331

Insala - 55,167,331

AmericanMicronic- 25,152,219

Kenster - 52,10,25

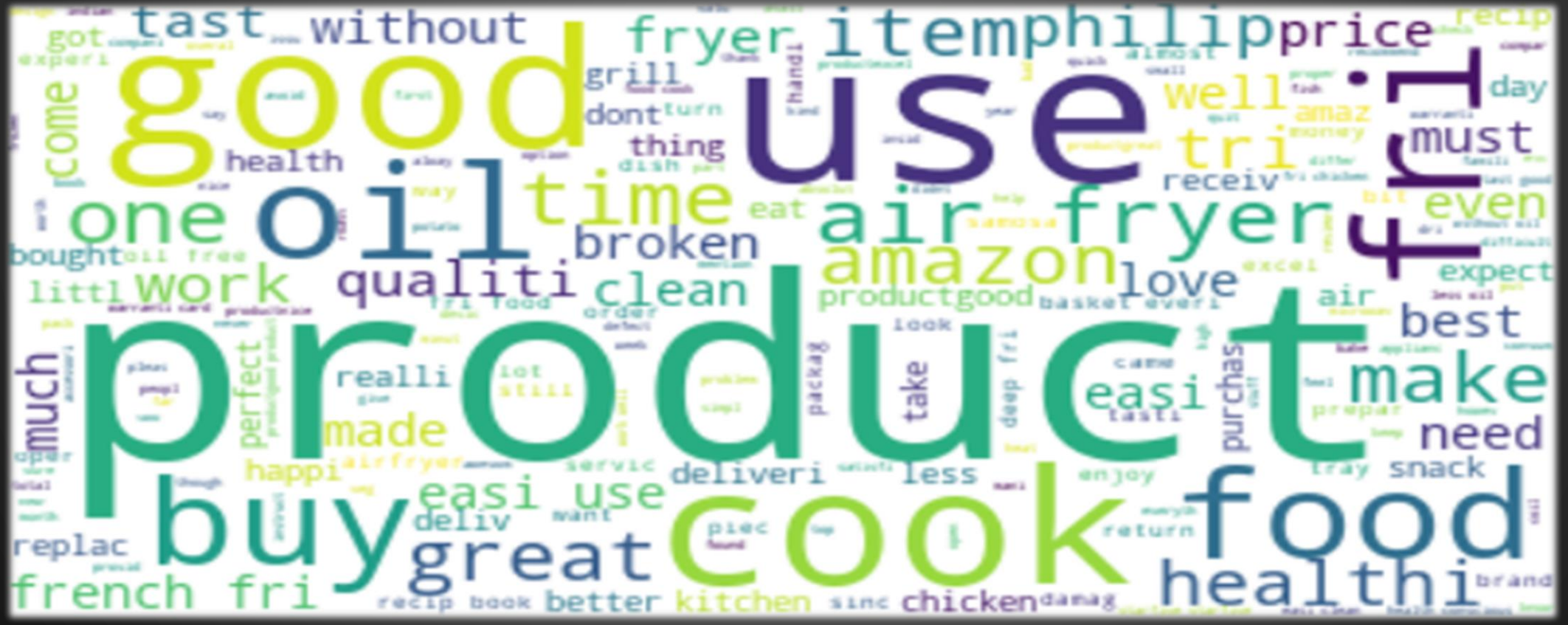
Hilton- 18,6

Kent- 11,5

Varada Pro- 18,1

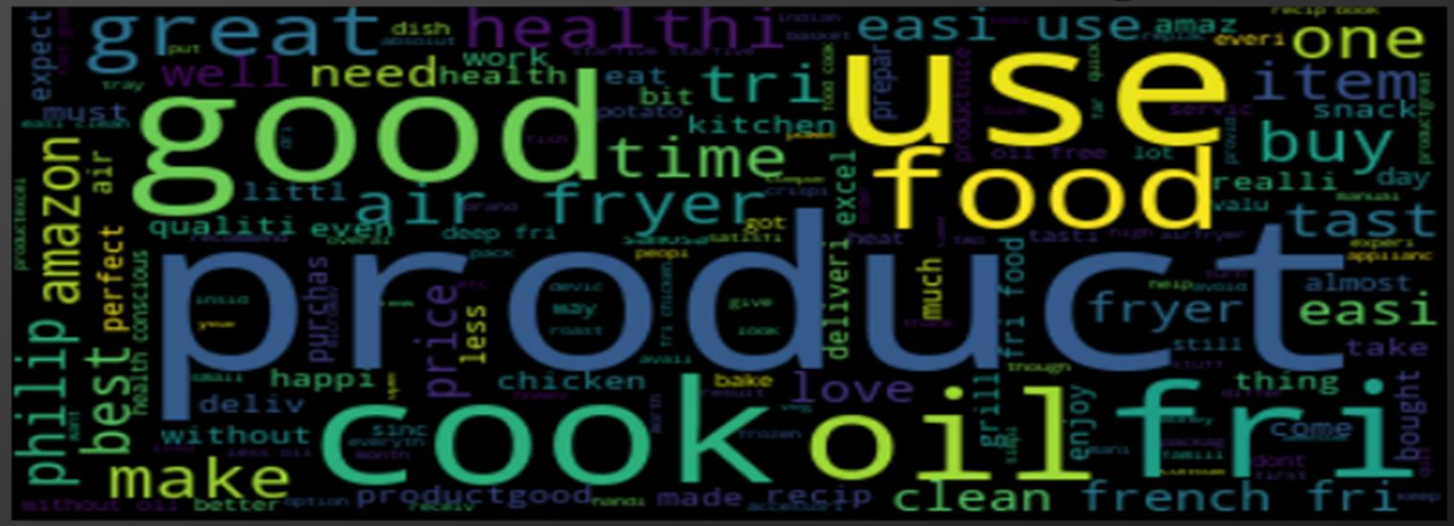
Word cloud

A **word cloud** (also known as a tag **cloud** or text **cloud**) is a visual representation of a text, in which the **words** appear bigger the more often they are mentioned. **Word clouds** are great for visualizing unstructured text data and getting insights on trends and patterns.

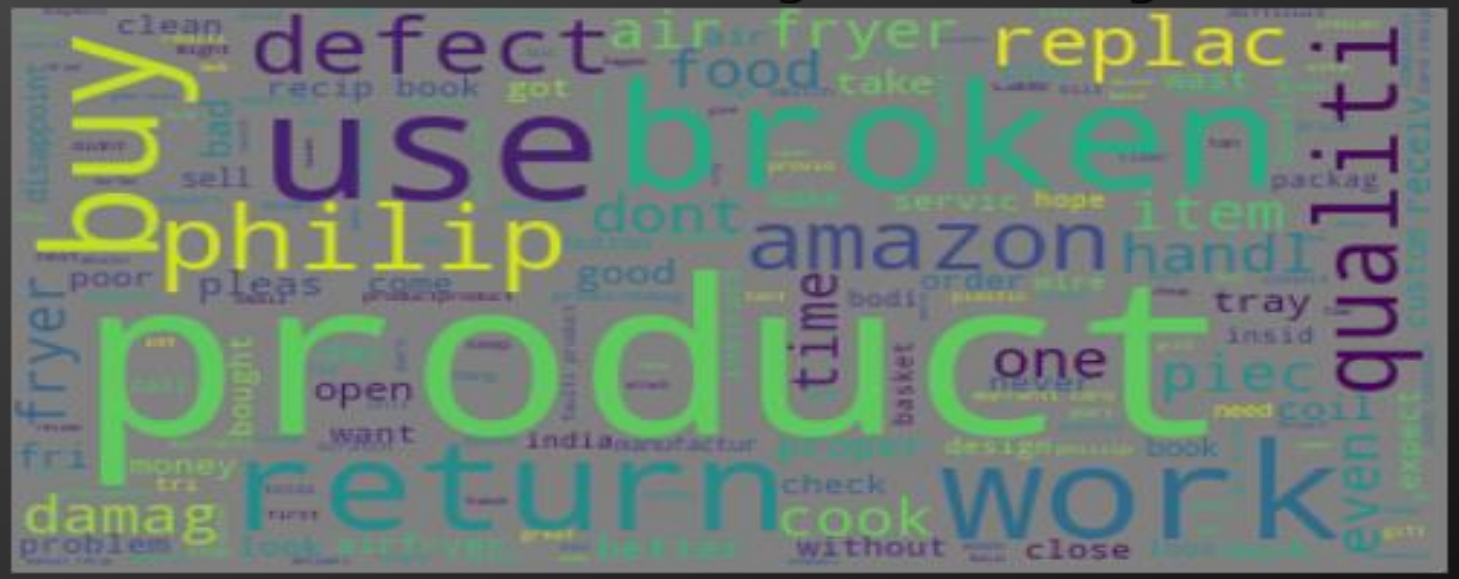


Positive & Negative Rating

Reviews with Positive Rating



Reviews with Negative Rating



```
# highest positive sentiment reviews
only_pos=product_data[product_data["Polarity"] >0].sort_values("Sentiment_Type", ascending = Fals
e)[[ "CleanedText","Sentiment_Type","Brand","Polarity"]].head(5)
```

	CleanedText	Sentiment_Type	Brand	Polarity
1	air fryer get hot outsid	POSITIVE	Kent	0.125000
2447	great product amaz price	POSITIVE	Insala airfryer	0.700000
2452	good product reason price	POSITIVE	Insala airfryer	0.450000
2453	excel product	POSITIVE	Insala airfryer	1.000000
2455	easi oper	POSITIVE	Insala airfryer	0.433333

The most positive reviews indeed correspond to some good feedbacks.

```
# highest Neutral sentiment reviews
only_Neutral=product_data[product_data["Polarity"] ==0].sort_values("Sentiment_Type", ascending = F
alse)[["CleanedText", "Sentiment_Type","Brand","Polarity"]].head(10)
```

	CleanedText	Sentiment_Type	Brand	Polarity
0	push button work	NEUTRAL	Insala airfryer	0.0
2314	five star	NEUTRAL	American Micronic	0.0
2316	five star	NEUTRAL	American Micronic	0.0
2319	five star	NEUTRAL	American Micronic	0.0
2320	five star	NEUTRAL	American Micronic	0.0

```
# highest Negative sentiment reviews
only_Neg=product_data[product_data["Polarity"] <0].sort_values("Sentiment_Type", ascending = False)[[
"CleanedText", "Sentiment_Type","Brand","Polarity"]].head(5)
```

	CleanedText	Sentiment_Type	Brand	Polarity
6	worst experi	NEGATIVE	Insala airfryer	-1.00
517	total wast moneytak lot time cooktoo much unne...	NEGATIVE	Philips	-0.30
603	satisfi	NEGATIVE	Insala airfryer	-0.25
592	expect upto	NEGATIVE	Philips	-0.10
587	take long time cook	NEGATIVE	Insala airfryer	-0.05

Accuracy score of Naives Bayes is 82.83

```
classifier.fit(X_train, y_train)
pred_NB = classifier.predict(X_test)
from sklearn.metrics import accuracy_score, precision_score, recall_score
print('Accuracy score of NB: ', accuracy_score(y_test, pred_NB)*100)
cm = metrics.confusion_matrix(y_test, pred_NB)
cm
```

```
Accuracy score of NB:  82.8397212543554
array([[ 48,  15,  52],
       [ 11, 144,  77],
       [ 20,  22, 759]])
```

Accuracy score of SVC is 87.64

```
print("Accuracy of SVC: ",accuracy_score(Y_test,y_pred)*100)
print("Classification Report\n",classification_report(Y_test,y_pred))
```

```
Accuracy of SVC:  87.64367816091954
Classification Report
              precision    recall  f1-score   support

          -1       0.81      0.57      0.67        37
           0       0.78      0.86      0.82        72
           1       0.92      0.93      0.92       239

    accuracy                           0.88       348
   macro avg       0.83      0.79      0.80       348
weighted avg       0.88      0.88      0.87       348
```

Application of TFIDF and bag of words

In TFIDF there is huge data loss compared to bag of words.

Here,

X = cv.fit_transform(corpus).toarray()

y=product_data['pol_num']

Accuracy

Accuracy value is high in SVM using k-fold compared to Naives Bayes.so the data is proceeded with svm

LDA:

Latent Dirichlet Allocation (**LDA**) is an example of topic model and is used to classify **text** in a document to a particular topic. It builds a topic per document model and **words** per topic model, modeled as Dirichlet distributions. Here we are going to apply **LDA** to a set of documents and split them into topics.

From data,

No.of topics taken =3

>3 the data is getting overlapped on each other

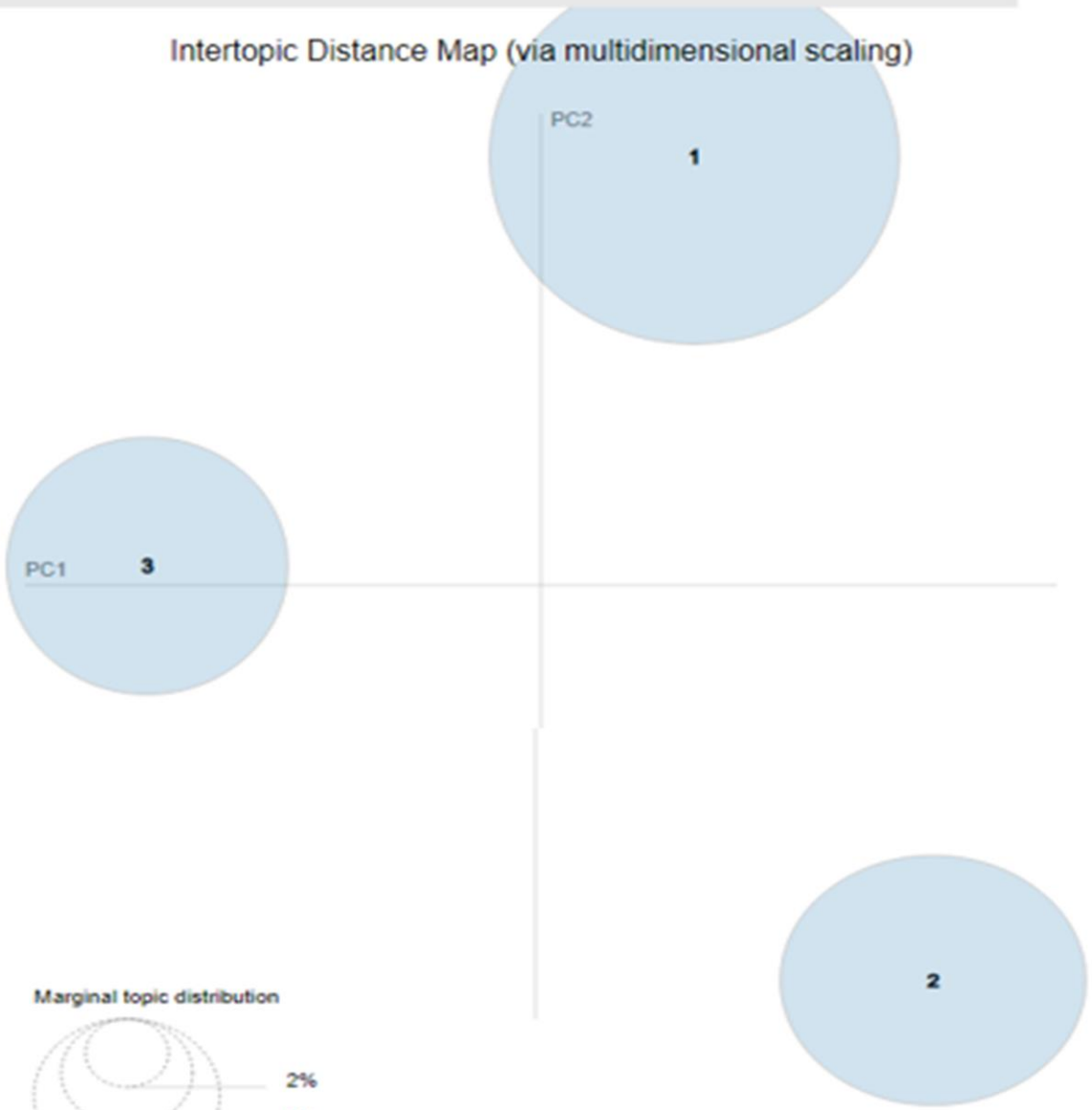
Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

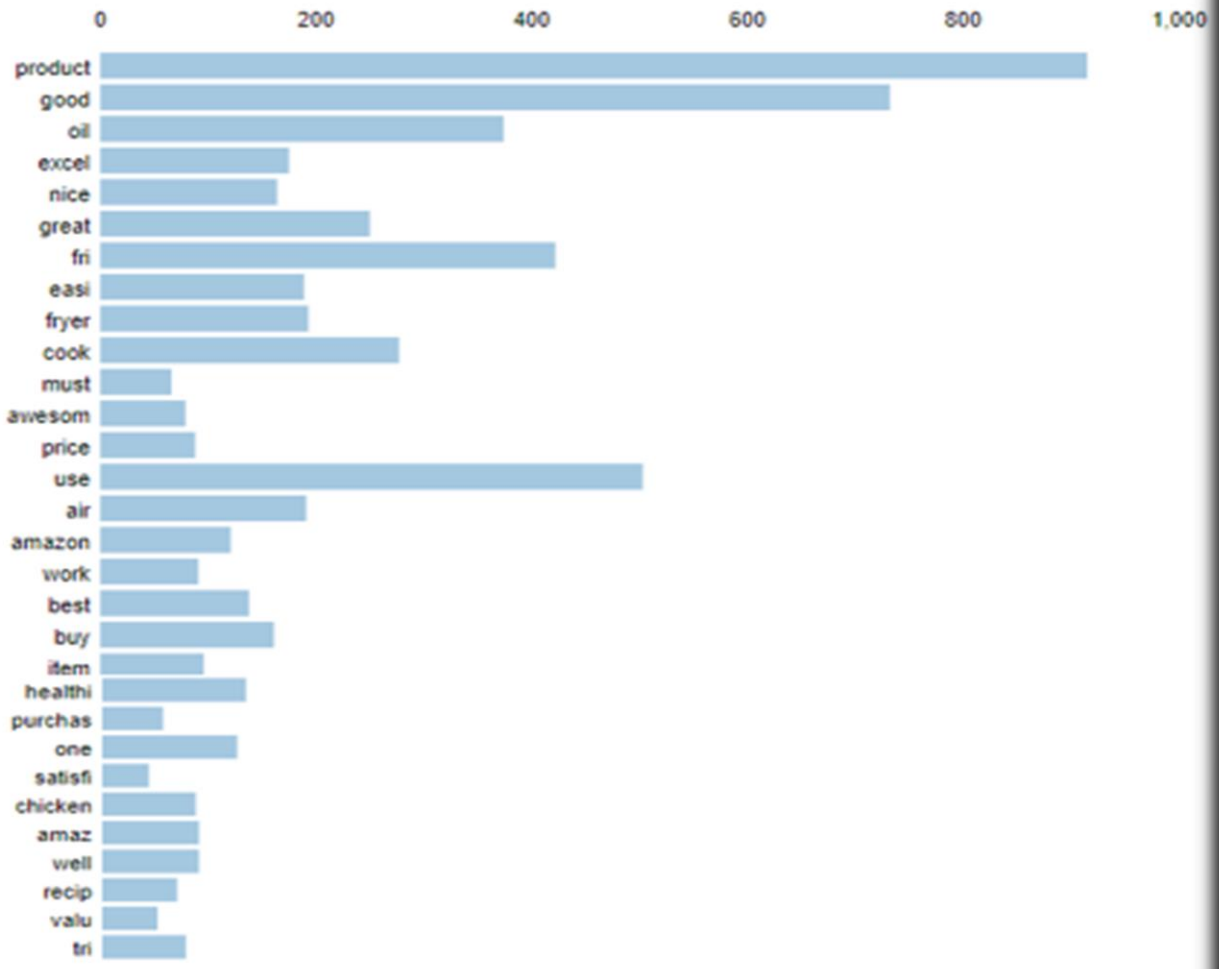
λ = 1

0.0 0.2 0.4 0.6 0.8 1.0

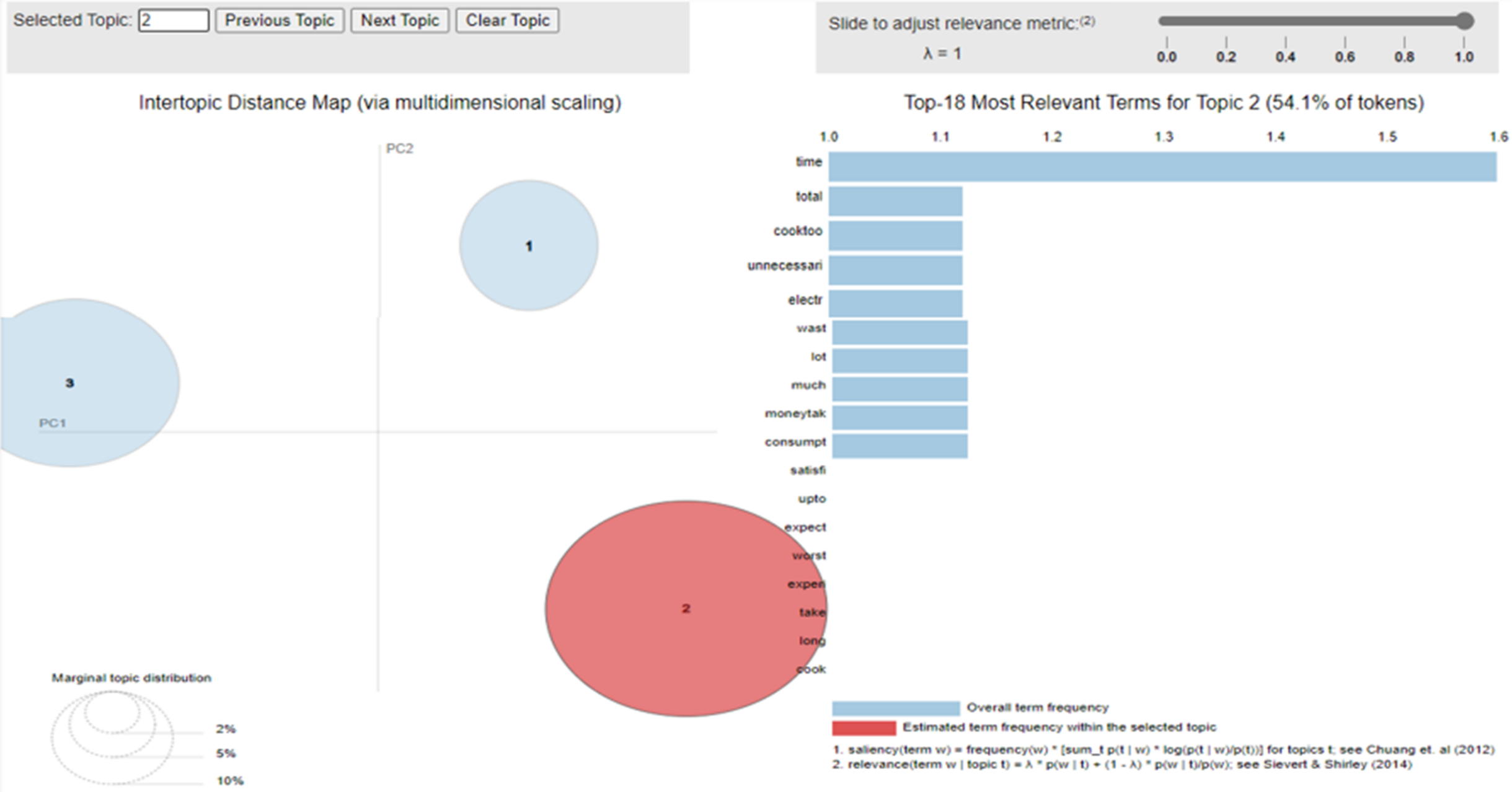
Intertopic Distance Map (via multidimensional scaling)

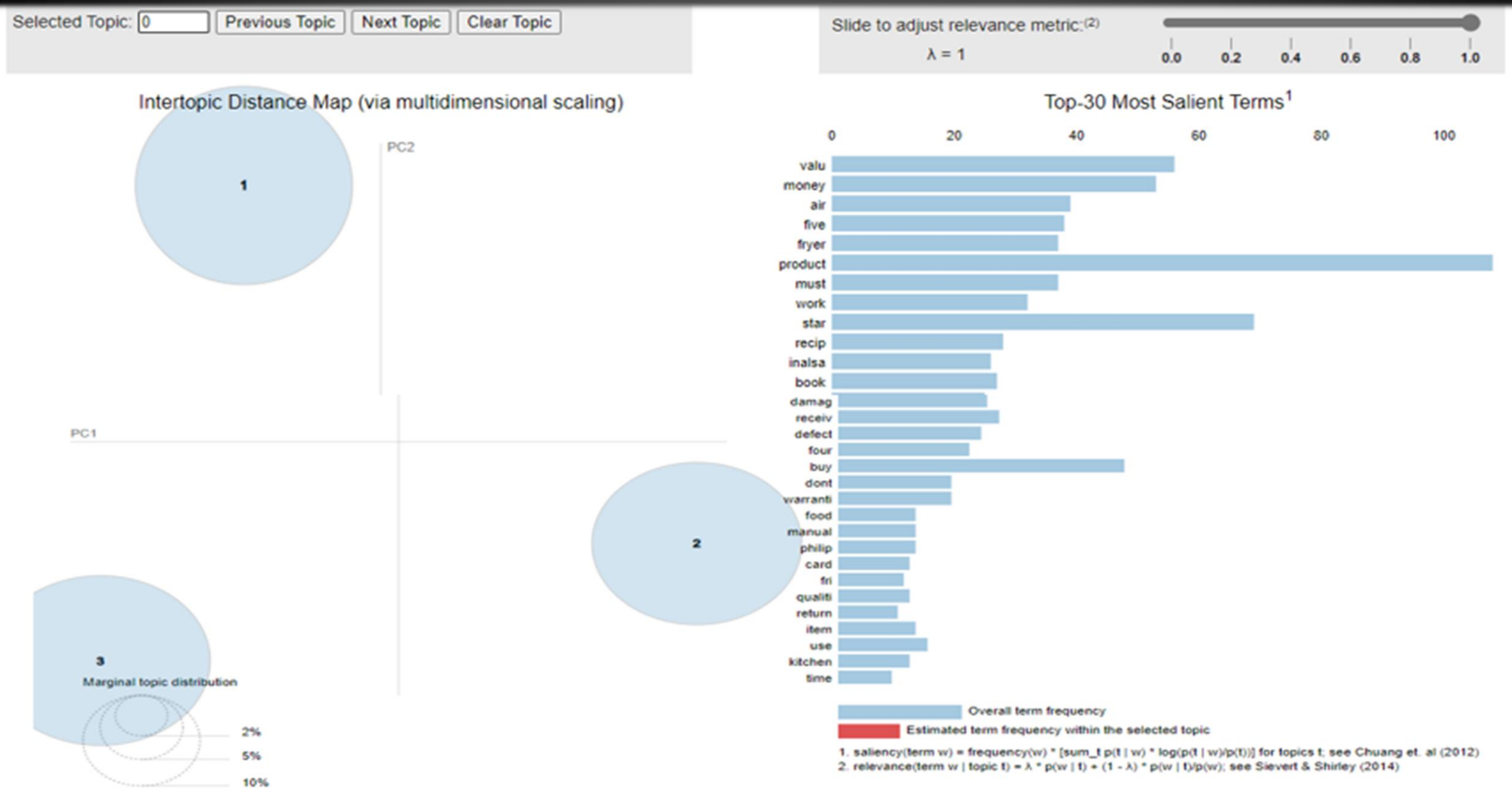


Top-30 Most Salient Terms[1]



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
10%
Top-18 Most Relevant Terms for Topic 2 (54.1% of tokens)
1.0
1.1
1.2
1.3
1.4
1.5
1.6
time
total
cooktoo
unnecessari
electr
wast
lot
much
moneytak
consumpt
satisfi
upto
expect
worst
experi
take
long
cook
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
0
20
40
60
80
100
valu
money
air
five
fryer
product
must
work
star
recip
inalsa
book
damag
receiv
defect
four
buy
dont
warranti
food
manual
philip
card
fri
qualiti
return
item
use
kitchen
time
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Positive reviews based on brands

Phillips

1. Working is good
2. Less oil Consumption,tasty
3. Product easy use and easy clean
4. Best use for deep fry stuff
5. Worth for money and early delivery

Amazon

#Healthy option to buy
#Less oil Consumption,tasty
#Product easy use and easy clean
#Fantastic product at most competitive in price
#From overall Data Amazon AirFrier Stands best in its cost

American Micronic

#Positive reviews on American Micronic
#1.A healthy way to perfect cooking
2.A must-have for an electric kitchen
#3.DAMM GOOD FOR THE PRICE POINT!!!!

Negative reviews based on brands

Amazon

#NEGATIVE REVIEWS ON AMAZON AIR FRYER**
#1.Poor paint quality
#2.Improper packing
#3.Missing of manual and warranty

Kenster

#Negative reviews on kenster
#1.Broken product
#2.Product defective

Kent

#Negative reviews on kent
#1. Bad quality.
#2.Malfunctioning within 2 weeks of purchase
#3. Didn't work after 1 time use

Insala

#Negative reviews on insala air fryer
#1.Low Quality/ Pathetic Customer Support
#2. Product has manufacturing defect.
#3.Poor design and Quality. Look for something Better.
#4. Loose body. Top cover not closing tightly
#5.Air frier became nonfunctional within gaurentee period
#6.No warranty card no recipe book received
#These are the issues faced by the customers on this product.

American Micronic

#Negative reviews on American Micronic
#1.Worst Warranty Service
#2.Not as expected!!
#3.Not upto the Mark
#4.Not a versatile fryer!

Challenges faced :

Phillips:

1.Defective item received.
2.Damaged product.
3.Used product delivered.
4.No proper packing.
5.Malfunctioning fryer.

Amazon:

1.Defective Timmer knob ,not working.
2.Feel better options available in market compared to brough ones.
3. Received fault product or used products.
4.Cheap non stick coating.
5.Instructions book is missing.
6.Facing difficulties in cleaning.
7. Refrain from buying. No service.No Reliability.Failed within 1 year of use.

Insala:

1.Rust inside air fryer.
2.Poor Quality
3.Problem in functioning.
4.Not heating well

American Micronic:

1.Dissatisifactory company response.
2.Didn’t serve the purpose.
3.Worst warranty service.
4.Bad transportation.

Kenster:

1.Poor quality.
2.No recipe book.

Kent:

1.Getting heat outside while working.
2.Bad quality.

How to Over come:

1.Need improvement in Quality of Product.
2.Provided service centres near to customers location.
3.Improvement in length and quality of wire.
4.Service support.
5. Proper Product guidance is required.
6.Quality of non stick coating and plastic quality need to be improved.
7.Customer satisifaction should also be improved.

Reference

Analytics vidhya

Medium

Towards Data science

NLP: Gaining insights from text reviews | by Fredrik Olsson | Towards Data Science

https://medium.com/artefact-engineering-and-data-science/customer-reviews-use-nlp-to-gain-insights-fr

https://www.geeksforgeeks.org/python-nlp-analysis-of-restaurant-reviews/

How to use Natural Language Processing to analyze product reviews? | by Gunnvant Saini | Towards Data Science

Thank you