Web & Social Media Analytics

Submitted By
Vishwanath Rajasekaran
Harsh Kushwaha
Srishti Sri

Introduction

- With a growing trend towards digitization and prevalence of mobile phones and internet access, more consumers have an online presence and their opinions hold a good value for any product-based company, especially so for the B2C businesses. The industries are trying to fine-tune their strategies to suit the consumer needs, as the consumers leave some hints of their choices during their online presence.

- In the retail e-commerce world of online marketplace, where experiencing products are not feasible. Also, in today's retail marketing world, there are so many new Phones are emerging every day. Therefore, customers need to rely largely on product reviews to make up their minds for better decision making on purchase. However, searching and comparing text reviews can be frustrating for users. Hence we need better numerical ratings system based on the reviews which will make customers purchase decision with ease.

GOAL

- In this capstone project, we will be solving a problem in the mobile phone industry of the US, one of the major smartphones markets in the world.

- By analyzing the sentiment of the reviews, we can find the features of the phones that have resulted in positive/negative sentiments. This will help companies include or improve those particular features while developing a new product.

- Comparing the competitors' pricing and their market shares will help companies decide the price of their products.

- Before purchasing any product, we all look at similar products in various brands. This data will help the companies know their major competitors in the market.

DATA COLLECTION

○ **Phone metadata:** This data contains the product information and is independent of the consumer/reviewer activity and includes description, price, sales-rank, brand info, and co-purchasing links etc. The original data was in json format. The json was imported and decoded to convert json.

　　○ The sample dataset is shown below:

	category	description	title	also_buy	brand	feature	rank	also_view	similar_item	date	price	asin
36	['Cell Phones & Accessories', 'Cell Phones', '...	[ICE CENIOR *SENIOR PHONE*, QUAD BAND Super bi...	UNLOCKED DUAL SIM SLOT *CENIOR PHONE* QUAD BAN...	[]	Ice	[2G NETWORK, Quad Band: GSM 850/900/1800/1900 ...	[>#6,356,920 in Cell Phones & Accessories (See...	[]				8050110508
1274	['Cell Phones & Accessories', 'Cell Phones', '...	[Standard package: 1 x original phone 1 x Qual...	Nokia 3310 Blue Nokia	[B075FL4H89, B00R25GJJW]	Nokia	[GSM 900/1800, Simple and elegant, Classic mod...	[>#32,759 in Electronics (See Top 100 in Elect...	[B075FL4H89, B075FKZMR2, B00TLWTJLO, B0757B64H...	class="a-bordered a-horizontal-stripes a-spa...	October 19, 2014		B00005KBGR
1295	['Cell Phones & Accessories', 'Cell Phones']	[The Nokia 5180i is a handset offered by TracF...	Nokia 5180i TracFone Prepaid Cell Phone with 1...	[]	Nokia	[]	[]	[]				B00005S0M4
1354	['Cell Phones & Accessories', 'Cell Phones', '...	[Based on Motorola’s i90c phone, the i95c...	Motorola i95cl Phone (Nextel)	[]		[]	[>#3,798,962 in Cell Phones & Accessories (See...	[]				B00006J9HH
1361	['Cell Phones & Accessories', 'Cell Phones', '...	[This is a GSM Samsung R225 cellular phone tha...	T-Mobile Dual-Band Wireless Phone	[]	Samsung	[]	[>#5,552,033 in Cell Phones & Accessories (See...	[]				B00006LIQB

DATA COLLECTION

◦ **Phone data** : This contain Contains the consumer activity information.

  ◦ The sample dataset is shown below:

	overall	verified	reviewerID	asin	style	reviewerName	reviewText	summary	unixReviewTime	vote	review_sentiment
0	5.0	True	A24E3SXTC62LJI	7508492919	{'Color': 'Bling'}	Claudia Valdivia	Looks even better in person. Be careful to not...	Can't stop won't stop looking at it	1407110400	NaN	POSITIVE
1	5.0	True	A269FLZCB4GIPV	7508492919	NaN	sarah ponce	When you don't want to spend a whole lot of ca...	1	1392163200	NaN	POSITIVE
2	3.0	True	AB6CHQWHZW4TV	7508492919	NaN	Kai	so the case came on time, i love the design. I...	Its okay	1391817600	NaN	NEGATIVE
3	2.0	True	A1M117A53LEI8	7508492919	NaN	Sharon Williams	DON'T CARE FOR IT. GAVE IT AS A GIFT AND THEY...	CASE	1391472000	NaN	POSITIVE
4	4.0	True	A272DUT8M88ZS8	7508492919	NaN	Bella Rodriguez	I liked it because it was cute, but the studs ...	Cute!	1391385600	NaN	POSITIVE

DATA WRANGLING

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1123744 entries, 0 to 1123743
Data columns (total 22 columns):
 #   Column            Non-Null Count      Dtype
---  ------            --------------      -----
 0   Rating            1123744 non-null    int64
 1   verified          1123744 non-null    bool
 2   reviewerID        1123744 non-null    object
 3   asin              1123744 non-null    object
 4   style             1123738 non-null    object
 5   reviewerName      1123744 non-null    object
 6   reviewText        1123744 non-null    object
 7   summary           1123744 non-null    object
 8   unixReviewTime    1123744 non-null    int64
 9   vote              1123504 non-null    object
 10  review_sentiment  1123744 non-null    object
 11  category          63438 non-null      object
 12  description       63438 non-null      object
 13  title             63438 non-null      object
 14  also_buy          63438 non-null      object
 15  brand             63438 non-null      object
 16  feature           63438 non-null      object
 17  rank              63438 non-null      object
 18  also_view         63438 non-null      object
 19  similar_item      63438 non-null      object
 20  date              63438 non-null      object
 21  price             63438 non-null      object
dtypes: bool(1), int64(2), object(19)
memory usage: 189.7+ MB
```

◦ **Merging Dataframes**: Phone reviews and meta datasets in json and csv files were saved in different dataframes and two dataframes were merged together using left join and "asin" was kept as common merger.

◦ Final merged data frame description is shown below:

DATA WRANGLING

- We will be categorizing only Phone data for our analysis

```
[9]  # Categorizing only Cell Phone Category for furture analysis
     df_meta['category'] = df_meta['category'].astype(str)
     df_meta = df_meta[df_meta.category.str.contains("'Cell Phones'")]
```

- Also we will be segregating Phone data from the merged dataset by filtering Phone from title

```
##########################################
## EXTRACTING PHONES FROM TITLE COLUMN
##########################################
product_reviews_p = product_reviews2[product_reviews2["title"].str.contains("Phones|Phone|phones|phone")]
```

- Final phones dataset has 45562 row.

Handling Duplicates, Missing Values

- 34 Duplicate records from meataphone and 3452 from phone are dropped .

- Image column has been dropped from Phone dataset.

- Review , style,Summary,vote column having null values was filled by bfill method.

- Summary and reviewText columns was merged and created a new column review_text. After that both the columns were dropped.

- unixReviewTime was converted to datetime '%m %d %Y format and created a new column Date&Time.

- By splitting Rating feature into good and bad rating we have creating new feature rating class.

Rating Distribution
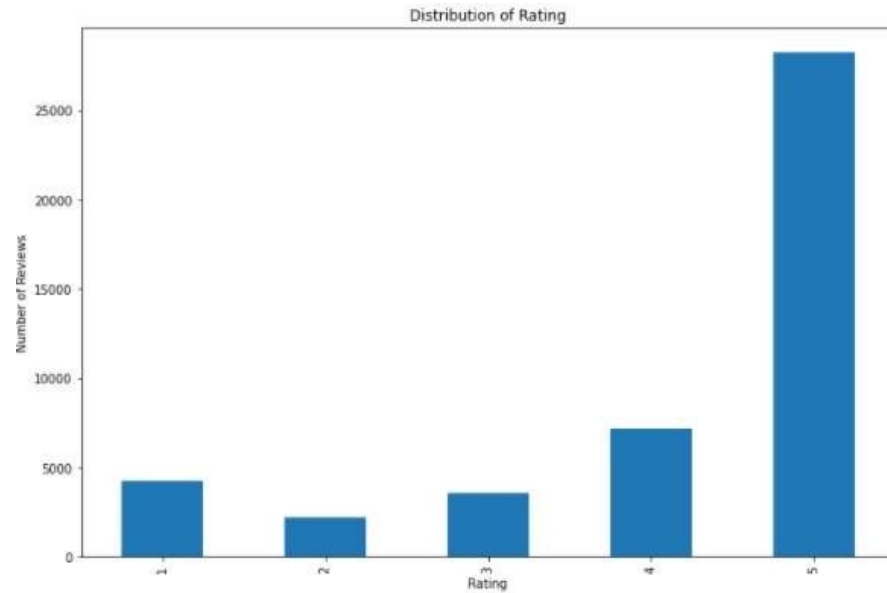
- We have seen 25K+ people have given 5 rating which is overall highest.

Descriptive Statistic Summary

- 25213 customer gives ratings and mean of the ratings is 4.1, which means that customers prefer to give high ratings for products. To be able to predict the ratings reasonably, we classified them as 'good' and 'bad' above.

According to the statistics on rating stars:

- 4246 customers give 1 star

- 2178 customers give 2 stars

- 3565 customers give 3 stars

- 7202 customers give 4 stars

- 28229 customers give 5 stars

- 6424 customers give bad ratings

- 38996 customers give good ratings

Preprocessing Text: TEXT ANALYTICS

Since, text is the most unstructured form of all the available data, various types of noise are present in it and the data is not readily analyzable without any pre-processing. The entire process of cleaning and standardization of text, making it noise-free and ready for analysis is known as text preprocessing. In this section, the following text preprocessing were applied.

Removing accented characters Accented characters/letters were converted and standardized into ASCII characters.

Expanding Contractions Contractions are shortened version of words or syllables. They exist in either written or spoken forms. Shortened versions of existing words are created by removing specific letters and sounds.
In case of English contractions, they are often created by removing one of the vowels from the word.

Removing Special Characters :One important task in text normalization involves removing unnecessary and special characters.

Lemmatization:The process of lemmatization is to remove word affixes to get to a base form of the word.
The base form is also known as the root word, or the lemma, will always be present in the dictionary.

Removing stopwords: Stopwords are words that have little or no significance. They are usually removed from text during processing so as to retain words having maximum significance and context. Here we have used given text file to remove sop words.

Building a Text Normalizer Based on the functions which we have written used and with additional text correction techniques (such as lowercase the text, and remove the extra newlines, white spaces, apostrophes), we built a text normalizer in order to help us to preprocess the new_text document.
After applying text normalizer to 'the review_text' document, we applied tokenizer to create tokens for the clean text.

Preprocessing Text: TEXT ANALYTICS

- As a result of that, we had 36488 words in total. After completing all data wrangling and preprocessing phases, we save the dataframe to csv file as a 'clean_text.csv. After cleaning, we have 45420 observations.

- We have used wordcloud package to visualize the cleaned data.

Preprocessing Text: TEXT ANALYTICS

◦ As identified Phone has been the most popular word in the dataset.
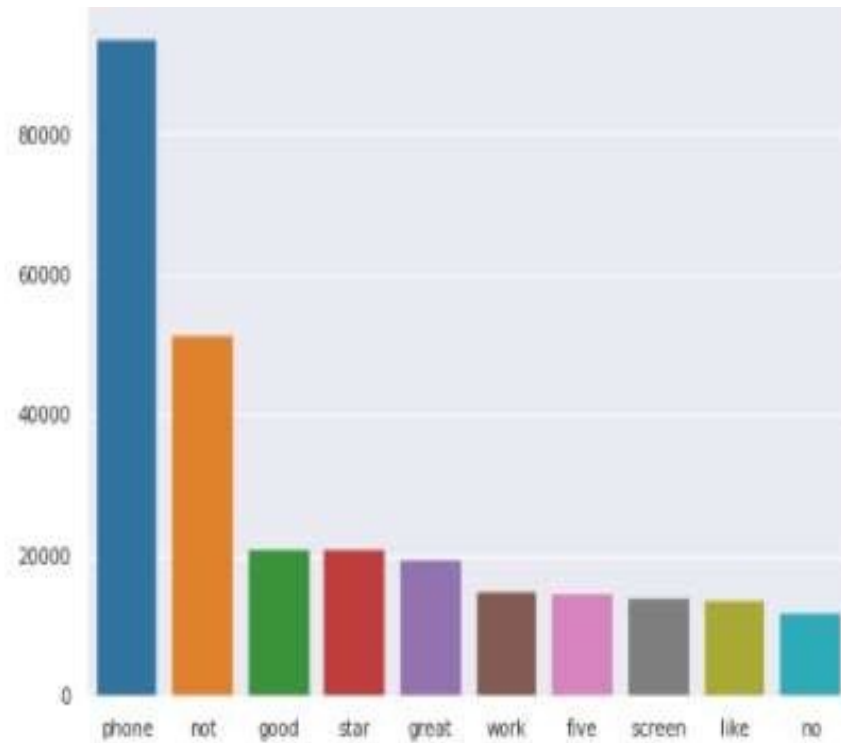
Sentimental Analysis

For Sentimenal Analysis I have used only sentiment review and clean_text column

```
Cleandata = df_review[['clean_text','review_sentiment']]
```

Now we have a cleandata dataframe as below

	clean_text	review_sentiment
45410	five star excellent product good seller	POSITIVE
45411	five star good	POSITIVE
45412	five star best price quality product great seller	POSITIVE
45413	phone like wear car radiator gradually heat sh...	NEGATIVE
45414	five star excellent product	POSITIVE
45415	shelle belle like funny rarely like fb reboot ...	POSITIVE
45416	five star not bad phone	POSITIVE
45417	bad iphone phone no good freeze touch power ho...	POSITIVE
45418	range price	NEGATIVE
45419	five star awesome thank love	POSITIVE

Sentimental Analysis

Spliting the dataset into train and test set

```python
# Splitting the dataset into train and test set
train, test = train_test_split(Cleandata,test_size = 0.1)
# Removing neutral sentiments
train = train[train.review_sentiment != "Neutral"]
```

Train_pos and train_neg as used for checking positive and negative review

```python
train_pos = train[ train['review_sentiment'] == 'POSITIVE']
train_pos = train_pos['clean_text']
train_neg = train[ train['review_sentiment'] == 'NEGATIVE']
train_neg = train_neg['clean_text']
```

As a next step I have separated the Positive and Negative comments of the training set in order to easily visualize their contained words. After that I cleaned the text from stopwords if any. Now they were ready for a WordCloud visualization which shows only the most emphatic words of the Positive and Negative review.

Sentimental Analysis

- Positive Word set: good, star, great, look, love

Sentimental Analysis

- Negative Word set: bad, problem, not, return, issue, replace, phone

Types Of Cell Phone Category And Their Percentage Of Total
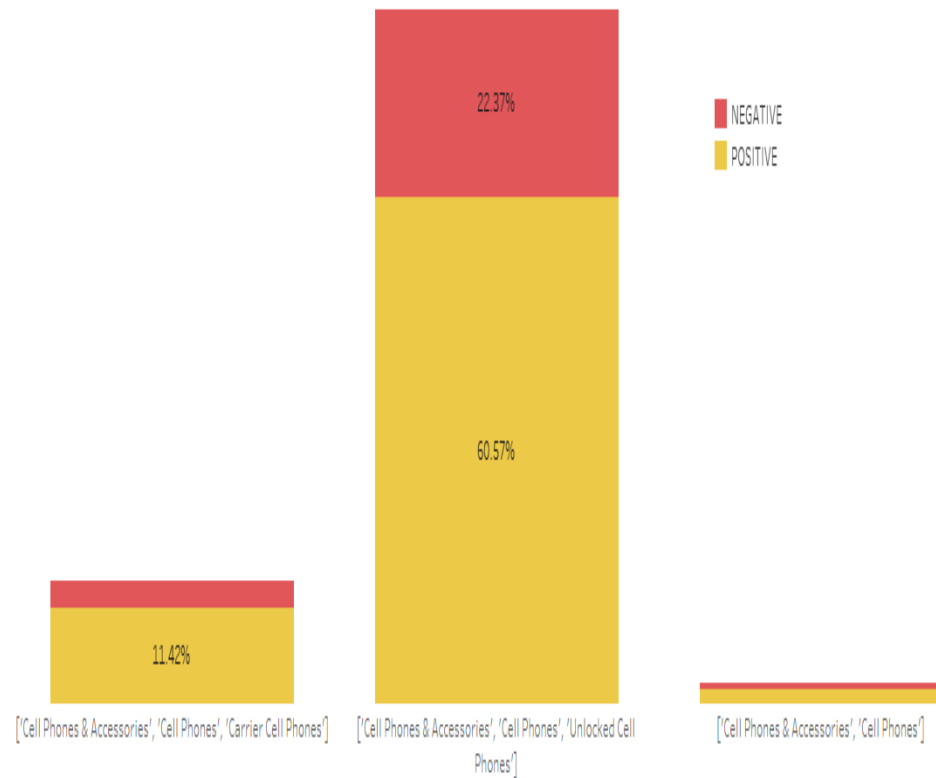
EXPLORATORY DATA ANALYSIS (EDA)

- After collecting data, wrangling data then exploratory analyses were carried out. The following insights were explored through exploratory analyses.

- Unlocked Cell Phones are seeming to be most likely used by customer having 60.57% positive reviews of the total number of reviews.

EXPLORATORY DATA ANALYSIS (EDA)
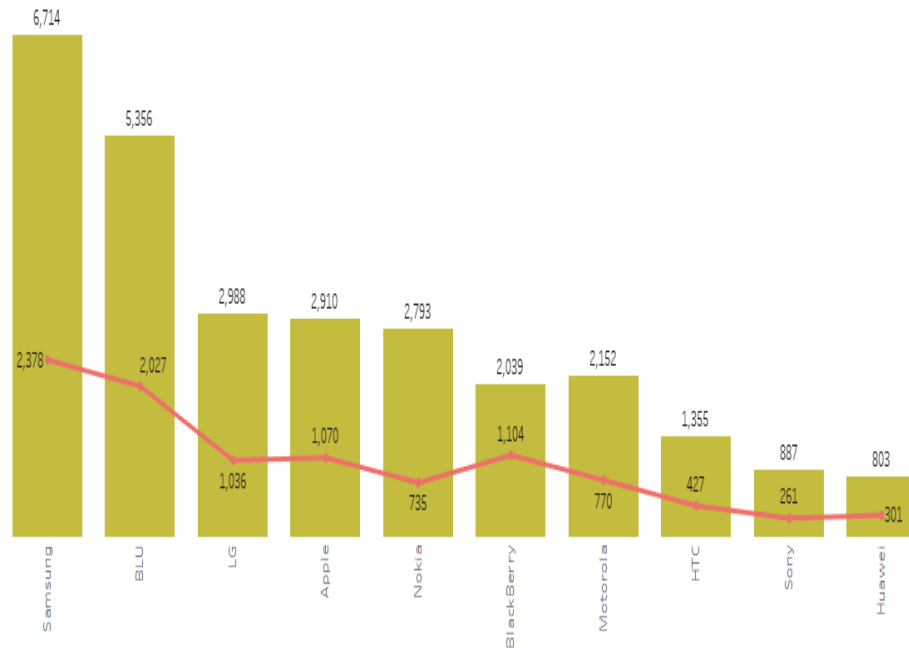
- Samsung is the topmost brand likely used by customer having a total of 6714 positive reviews and also its competitor is BLU having not much difference in terms of positive review.

- Also, Apple , LG and Nokia are the next Rivals companies of Samsung having almost same number of positive comments from the customer

EXPLORATORY DATA ANALYSIS (EDA)
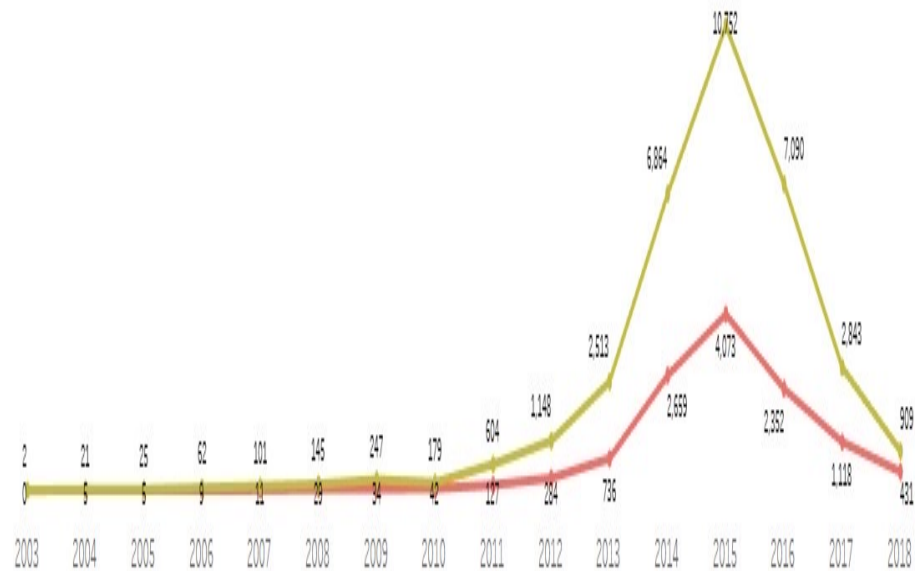
- If we have a look on the yearly stats in terms of reviews, we can observe after 2012 people are likely most interested in giving reviews(bad or good) to their purchased items and seems there is sudden increase in the number of comments in 2015 around 14k + in total.

- This might be because people are more aware of power and benefits of reviews.

EXPLORATORY DATA ANALYSIS (EDA)
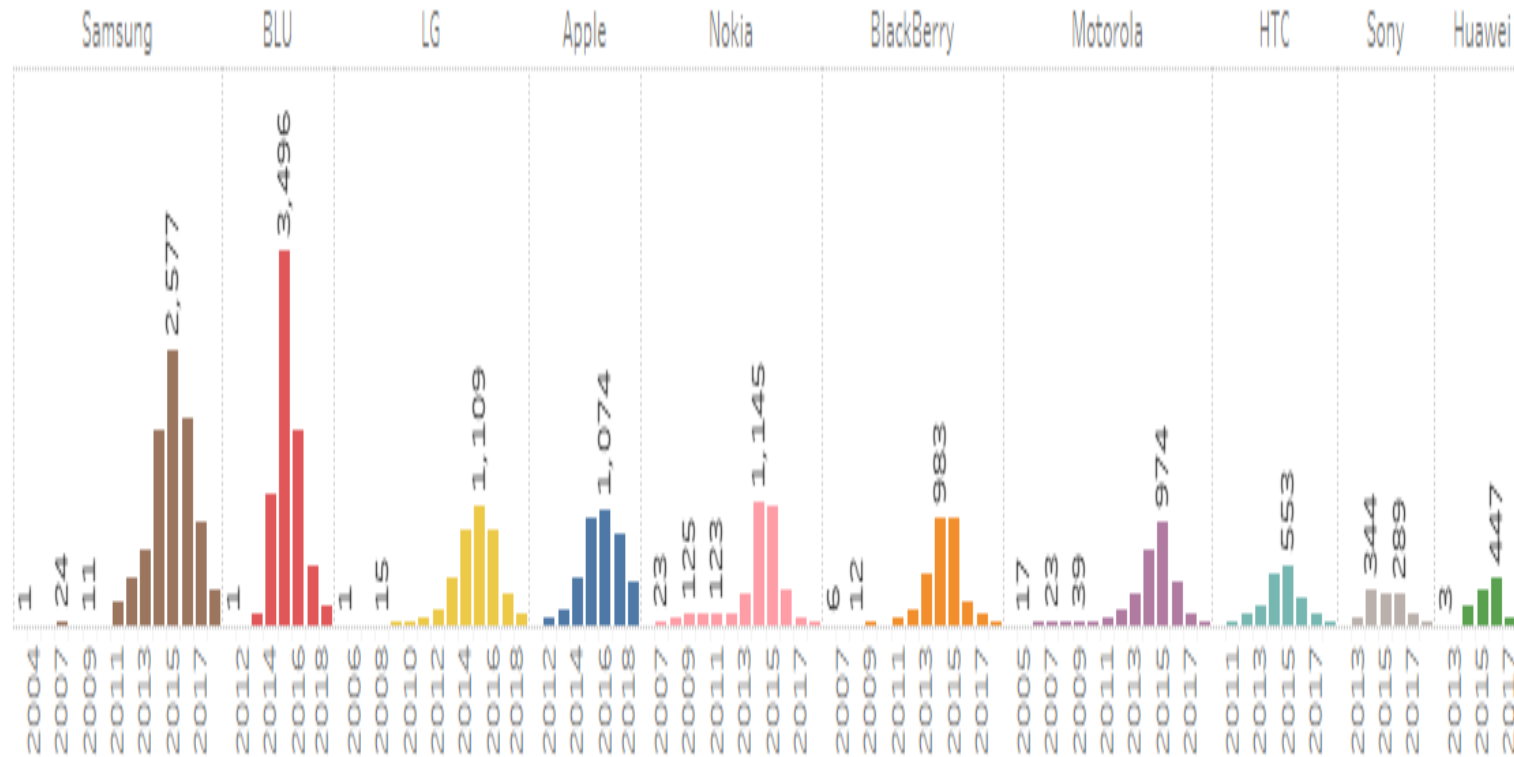
- Top 10 brand Yearly Review Stats
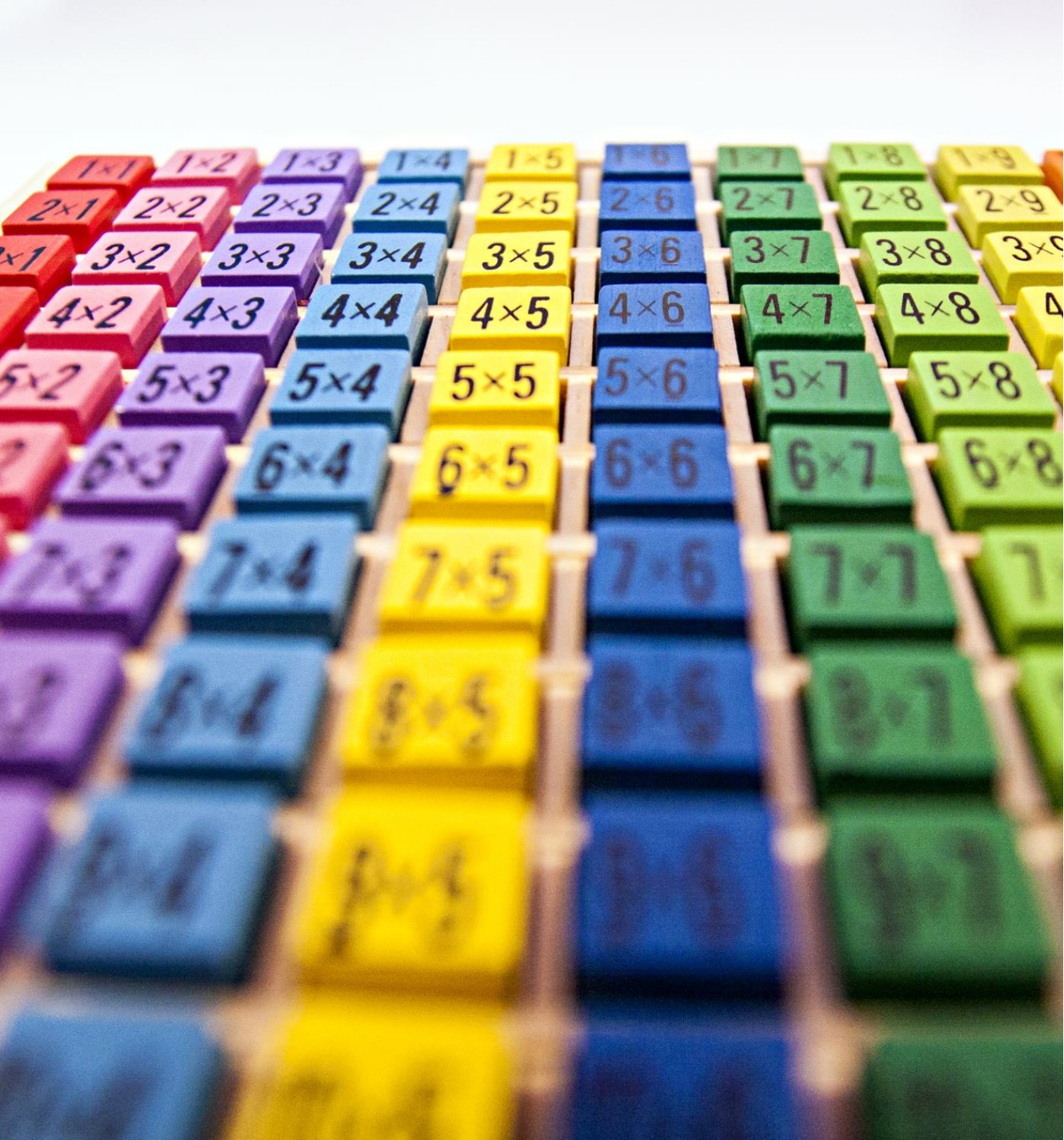
MACHINE LEARNING MODEL

○ In this project, the model needs to predict sentiment based on the reviews written by customers who bought phones. This is a supervised binary classification problem. Python's Scikit libraries was used to solve this problem. We have used **Naive Bayes** machine learning algorithms for modelling.

○ Naive Bayes implements the naive Bayes algorithm for multinomial distributed data, and is one of the two classic naive Bayes variants used in text classification (where the data are typically represented as word vector counts). This algorithm is a special case of the popular naïve Bayes algorithm, which is used specifically for prediction and classification tasks where we have more than two classes.

Below is review of new dataset formed having only rating_class and clean text.

	rating_class	clean_text
0	0	bad reception phone ugly heavy terrible user i...
1	1	pretty good phone improvement samsung decide t...
2	1	not user friendly motorola samsung phone not m...
3	1	best phone own be europe phone network better ...
4	1	love phone real problems phone amazingly light...

Later rating_class was re-named to Positivity

	Positivity	clean_text
0	0	bad reception phone ugly heavy terrible user i...
1	1	pretty good phone improvement samsung decide t...
2	1	not user friendly motorola samsung phone not m...
3	1	best phone own be europe phone network better ...
4	1	love phone real problems phone amazingly light...

MACHINE LEARNING MODEL

- Before splitting data into train and test dataset we have dropped all unnecessary column which are not going to be used in our algo.
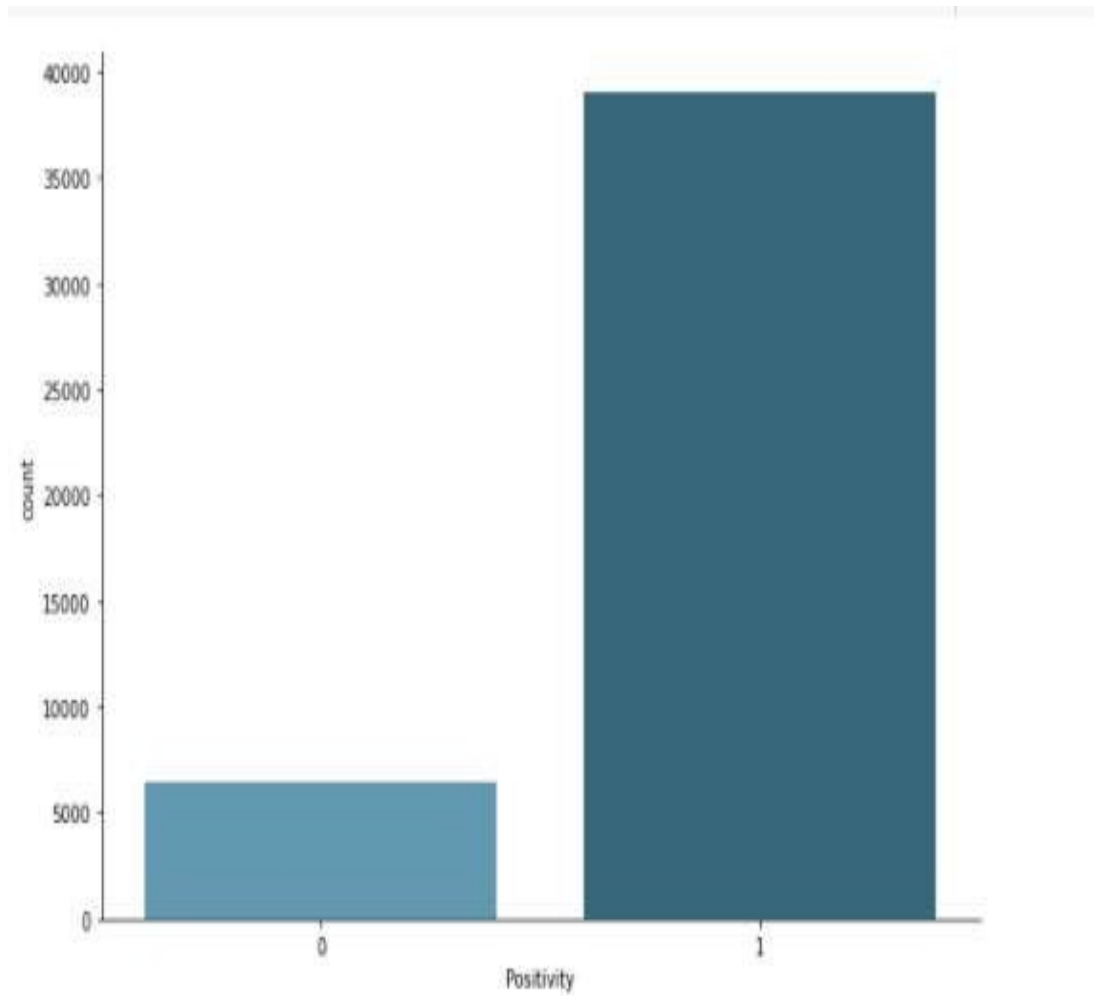
MACHINE LEARNING MODEL

- *The bar chat below showing a comparison between positive and negative reviews using phone dataset*

```
#Split data into train and test
x = df['clean_text']
y = df['Positivity']

X_train, X_test, y_train, y_test = train_test_split(x,y,test_size=0.33,random_state=42)

text_clf = Pipeline([('tfidf',TfidfVectorizer()),('clf',MultinomialNB())])

text_clf.fit(X_train,y_train)

predictions = text_clf.predict(X_test)

print(confusion_matrix(y_test,predictions))
cm = confusion_matrix(y_test,predictions)
print(classification_report(y_test,predictions))
```
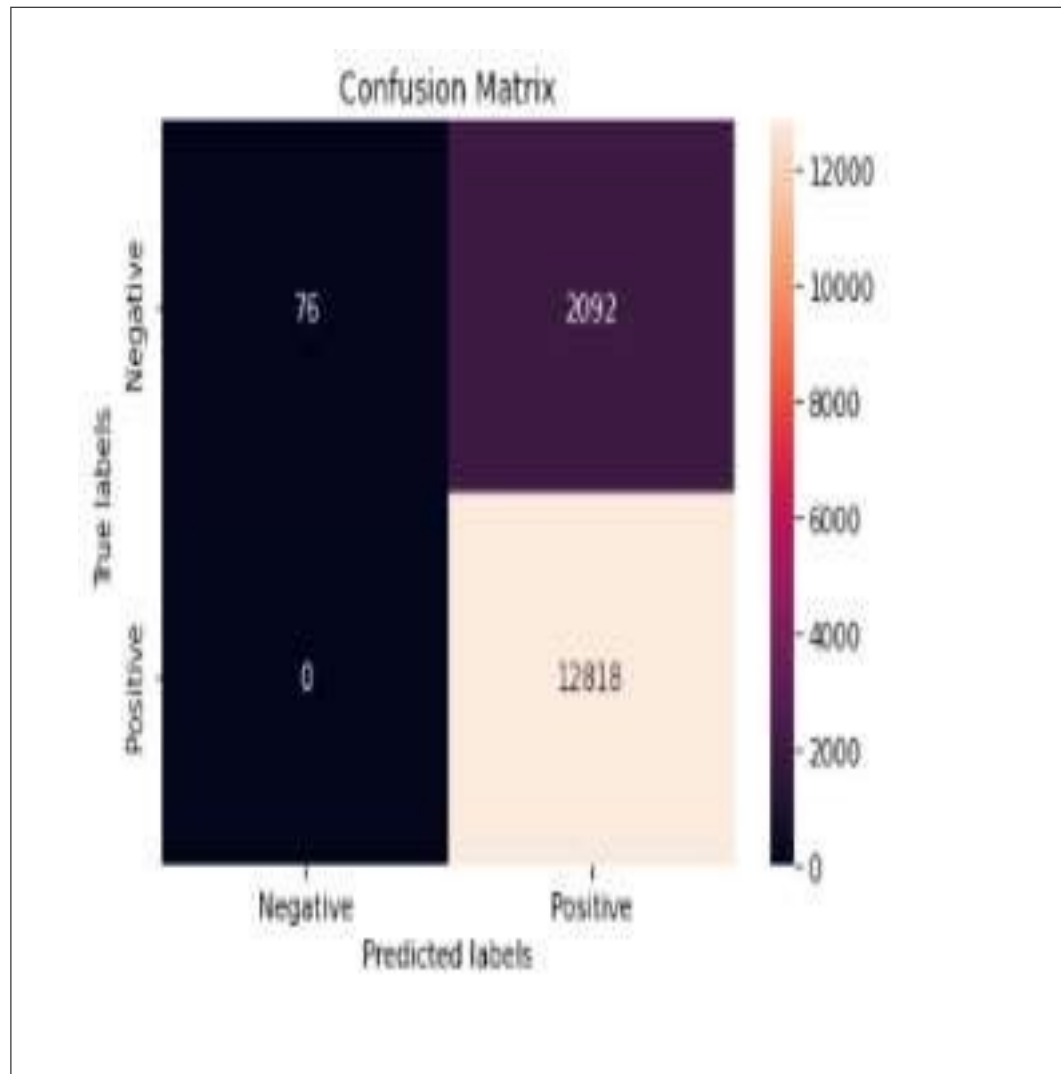
MACHINE LEARNING MODEL

○ Split dataset set into train and test and using MultinomialNB. Naive Bayes classifier for multinomial models.

Confusion matrix

Confusion Matrix

- True Positives (TP) In our dataset 12818 are the correctly predicted positive review which means that the value of actual class is yes and the value of predicted class is also yes.

- False Positives (FP) :When actual class is no and predicted class is yes. Here we have a total of 2092 FP review

- True Negatives (TN) :These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. We have 76 reviews which are correctly predicted as negative.

- False Negatives (FN) – When actual class is yes but predicted class in no. We don't have any review which predicted as negative though in actual it is positive.

Confusion matrix

- We got an accuracy of ˜86% on the test set.

- Accuracy :Accuracy is the most intuitive performance measure and it is simply a ratio of co rrectly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. In our case being a review is positive has an accuracy of 86%.

- Accuracy = TP+TN/TP+FP+FN+TN

  - Precision : Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. In our case being a review is positive has an precision of 86%.

- Precision = TP/TP+FP

  - Recall (Sensitivity) Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. In our dataset we have a recall for positive review as 1.

- Recall = TP/TP+FN

  - F1 score:F1 Score is the weighted average of Precision and Recall. Therefore, this score ta kes both false positives and false negatives into account. We have an f1 score of 92% for positive review.

- F1 Score = 2*(*Recall* * Precision) / (Recall + Precision)

Confusion matrix

Weighted average scores:
- The sum of the scores of all classes after multiplying their respective class proportions.
- It is the simple mean of scores of all classes. So, macro- average recall is the mean of the recalls of all the classes.

```
               precision    recall  f1-score   support

           0       1.00      0.04      0.07      2168
           1       0.86      1.00      0.92     12818

    accuracy                           0.86     14986
   macro avg       0.93      0.52      0.50     14986
weighted avg       0.88      0.86      0.80     14986
```

THANK YOU