**Sentiment Analysis using NLP**

**Project Final Report**

Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral (Source: The Oxford English dictionary).

Sentiment analysis (also known as opinion mining or emotion AI) is the use of [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing), [text analysis](https://en.wikipedia.org/wiki/Text_analytics), [computational linguistics](https://en.wikipedia.org/wiki/Computational_linguistics), and [biometrics](https://en.wikipedia.org/wiki/Biometrics) to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to [voice of the customer](https://en.wikipedia.org/wiki/Voice_of_the_customer) materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from [marketing](https://en.wikipedia.org/wiki/Marketing) to [customer service](https://en.wikipedia.org/wiki/Customer_relationship_management) to clinical medicine(Source: Wikipedia).

**Benefits/Uses of Sentiment Analysis:**

1. Helps improve customer service - Sentiment analysis provides businesses an idea of the customer opinion about their services. It helps identify problems and persuades the organization to look for a solution. The management can make use of sentiment analysis to improve their customer services.
2. Helps develop quality products- The responses from the customers can be used as a guideline to improve the service quality, better future product development, reduce customer churn or improve how the product is presented.
3. Develop new marketing strategies- With more data and information gathered through sentiment analysis, the organizations could develop an effective marketing strategy. By observing the customers’ conversations on their social media and detect the specific key messages related to your brand, specific marketing campaigns can be designed for the target consumers.
4. Improves crisis management- Frequent monitoring of the customers’ responses or opinions towards a brand would help to identify any issues quickly.

**Problem Statement**

The task here is to analyze customer reviews for cell phones and predict customer rating and categorize the reviews as 'positive', 'negative' or 'neutral' based on the review sentiment.

**Source of data**

The data was collected from Amazon website on Dec 26,2019. The data contains 2 csv files – items.csv and reviews.csv.

items.csv contains retrieved (read: scraped) items from Amazon.com search results using generated URL and specific query string to search only specific brands and has minimal 1 star review.

reviews.csv contains reviews for previously retrieved items at items.csv but not with columns from items.csv.

Here’s a link to the data

<https://www.kaggle.com/grikomsn/amazon-cell-phones-reviews>

**Data Cleaning and Wrangling**

Items.csv contained 720 rows and 10 columns.

reviews.csv contained 67,986 rows and 8 columns.

Items.csv contains columns such as product asin, cellphone brand, phone title, total reviews, producturl, product rating, reviewurl, price, image.

Reviews.csv contains columns such as product asin, customer name, rating (overall rating by customer), review title, review body, helpful votes.

Items.csv had 4 null values for brand column which was dropped using dropna() function.

reviews.csv had NaN values for review title and body columns. Dropped those columns as well because they are of not any use if they do not review text.

Once the data free from null values, the 2 data frames were merged into one dataframe using the common column ‘asin’. The resultant dataframe item\_reviews has 67,781 rows and 17 columns.

**Exploratory data analysis**

Exploratory data analysis revealed some very interesting facts about our data. I first explored items.csv.

The below graph shows a visual representation of the number of reviews for each brand.

Chart, histogram

Description automatically generated

Samsung is the most reviewed cellphone - at almost 37,000 reviews. The second highest is Motorola at 9419, and Apple stands third with 6315 reviews.

Here’s a screenshot showing the average rating of each phone brand

Table

Description automatically generated

We can see that Xiaomi tops the list. HUAWEI is second and ASUS third. It’s interesting that none of the top brands/highly reviewed brands have the highest rating.

**Text Cleaning/Pre-processing/Training:**

The goal of the preprocessing work is to prepare our data for Modeling. I performed pre-processing on item\_reviews.csv dataframe. The pre-processing for NLP problems is different from the pre-processing we perform for normal datasets.

First, I cleaned the text data by removing html tags, special characters, converted all text to lowercase, removed stop words (words which have no significance, words like a, an, the, and etc.), performed stemming (obtaining base form (root stem) of a word from its inflected form) and lemmatization (where we remove word affixes to get to the base form of a word (root word)). The difference between root stem and root word - root word is always a lexicographically correct word (present in the dictionary).

I used nltk (natural language toolkit) library to perform text cleaning and pre-processing. I defined individual functions to remove special characters, perform stemming, lemmatization and tokenization (removing stop words). Then I defined a function called normalize\_corpus which in turn called the individual functions. Then I passed the entire review text to the function ‘normalize\_corpus’, the output of this function is a clean text on which we can perform Modeling.

I decided to filter the item\_reviews dataframe based on ‘cellphone brand’. I chose only 4 brands because performing sentiment analysis on all 10 brands would be very lengthy and exhausting. I chose the top 3 most reviewed brands – Samsung, Motorola and Apple and chose Xiaomi as the 4th brand because it has the highest rating.

Now there are 4 dataframes namely apple, motorola, samsung and xiaomi. The next step is to perform Vectorizing- I used Count Vectorizer to transform a given text into a vector based on the frequency (count) of each word that occurs in the entire text. I split each of these dataframes into train sets and test sets.

**Text Visualization**

After splitting data into 4 dataframes, I used WordCloud generator to visualize the text in each of these 4 dataframes. Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.

Here’s a text visualization for Apple

A picture containing text

Description automatically generated

We can see the words 'phone', 'iPhone, 'good', 'new', 'battery', 'screen', 'work' is the most frequently used words in the apple review text.

Text Visualization for Samsung

Text

Description automatically generated

Phone is the most frequently used word. ‘Use’, ‘Samsung’, ‘love’, ‘Five Star’, ‘one’, ‘screen’ all these words also have high frequencies/importance.

Text Visualization for Motorola:

Text

Description automatically generated

‘phone’ is almost always the most frequently used word. We ca see that ‘Motorola’, ‘use’, ‘screen’, ‘good’, ‘work’, ‘one’, ‘great’, ‘great phone’ ‘battery life’ al these words are frequently used by customers.

Text Visualization for Xiaomi

Text

Description automatically generated

‘phone’ is the most frequent word. We can also see that customers have used words like ‘good price’, ‘love’, ‘Great phone’, ‘work’, great’, ‘screen’, ‘camera’, ‘fast’ etc. to describe their phone. Looks like Xiaomi users are very happy with the price of the phone and it’s camera.

**Modeling**

The goal of our project is to find out the customer rating based on the reviews provided by the customer. Based on the reviews we must classify the customer rating into 5 classes – 1,2,3,4 or 5 and classify the customer sentiment as either ‘positive, ‘negative’, ‘neutral’.

To predict the rating, we could use either Regression or classification models. I chose to use a classification model because there are 5 classes, and the task is to identify to which class the review belongs. The next task is to predict customer sentiment, for which we could use algorithms like TextBlob or VADER.

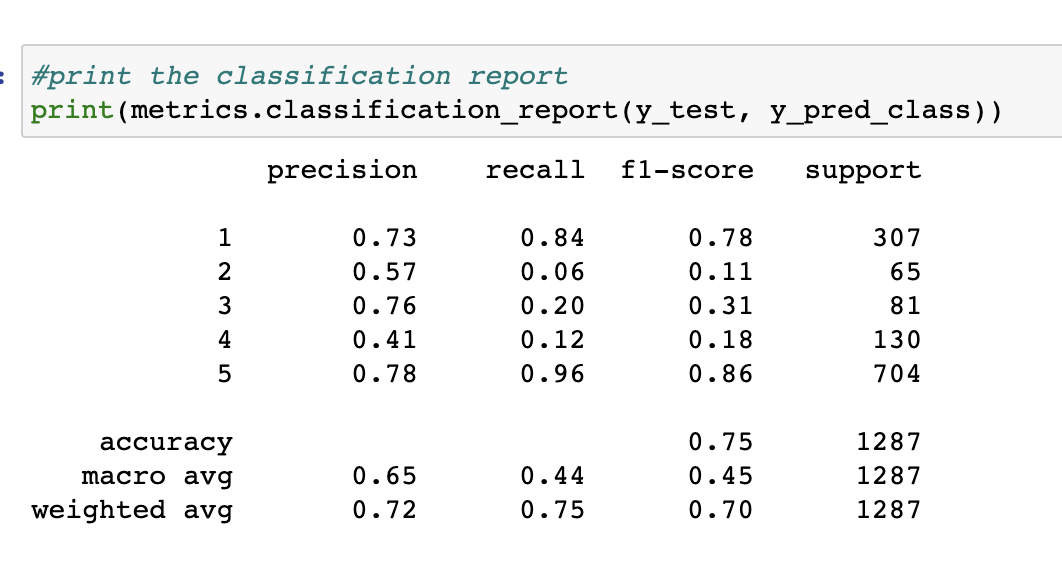
**Model 1: Multinomial Naïve Bayes**

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text. It is completely used for text documents where the text belongs to a class. The attributes required for this classification are basically the frequency of the words that are converted from the text document.

**Multinomial Naïve Bayes performance analysis for top 4 brands**

Apple: It produced an accuracy of 0.754 which is pretty good



Samsung: The classifier has an accuracy of 0.75 for samsung as well.

Table

Description automatically generated

The classifier does a good job at predicting rating 5. It works ok for rating 1 as well.

Motorola: accuracy of 0.70, less compared to Apple and Samsung

Classification report for Motorola:

Table

Description automatically generated

Good precision and recall for class 1. High precision and recall for class 5.

Xiaomi: The classifier has an accuracy of 0.72

Classification report for Xiaomi

Table

Description automatically generated

Class 1 and class 5 have good precision and recall. Class 2 has a perfect precision score!! and class 5 has a perfect recall score.

In general, most of the time, the Multinomial Naïve Bayes classifier does a good job at predicting rating 5 and rating 1. The accuracy of around 0.70 – 0.75 is good considering even humans can predict the rating of a product, based on the reviews only 70% percent of the time. And it is easier to predict 1- and 5-star rating reviews mostly because of the highly polar language used in those reviews.

**Model 2: Random Forest classifier**

The random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

**Random Forest Classifier performance analysis for top 4 cellphone-brands**

**Apple:**

The Random Forest Classifier has an accuracy of 0.73

Table

Description automatically generated

Interesting! Class 1 and has average precision. Class 5 have high precision and recall. Class 3 has high precision, low recall.

**Samsung:**

For Samsung, the Random Classifier has an accuracy of 0.739

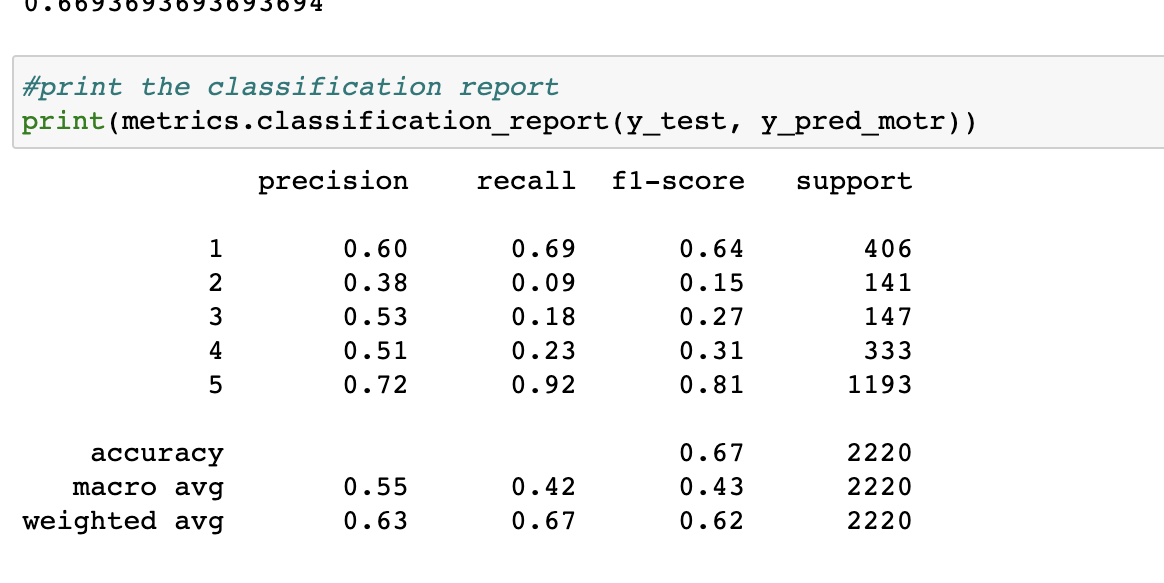
Table

Description automatically generated

Class 5 has high precision and recall; Class 1 has high recall and average precision. The other classes have average precision and low recall.

**Motorola**

The Random Forest Classifier has an accuracy of 0.66 for Motorola, quite low compared to other brands. Lets’ look at its classification report



Random Forest Classifier has produced average results for Motorola. It has high precision and recall for class 5, average precision and recall for class 1, but it is not that good in classifying the other classes.

**Xiaomi**:

Random Forest classifier has an accuracy of 0.71

Table

Description automatically generated

Class 5 has good precision and high recall. Class 1 has average precision and low recall. Other classes have average precision and low recall.

*Comparing all 4 brands, it’s evident that Multinomial Naïve Bayes Classifier has better performance and is better at classifying reviews into its respective classes than Random Classifier.*

In general, it looks like both classifiers are good at predicting 1-star and 5-star rating. It must be easier to predict 1- and 5-star rating mostly because of the highly polar language used in those reviews

I used another classifier called **SVM (Support Vector Machine)** which produced almost the same results as the Random Forest Classifier. *So, comparing the 3 classifiers, I think Multinomial Naïve Bayes is the most suitable algorithm for our data.*

**Sentiment Analysis**

There are many packages available in python which use different methods to do sentiment analysis.

TextBlob and VADER are the most popular among them. And among the two, VADER is the most popular for online data.

It uses a list of lexical features (e.g., word) which are labeled as positive or negative according to their semantic orientation to calculate the text sentiment. Vader sentiment returns the probability of a given input sentence to be positive, negative, and neutral.

To perform sentiment analysis on our data, I downloaded vader\_lexicon from nltk library and imported SentimentIntensityAnalyzer.

Let’s visualize the customer sentiment for each brand of cellphone using VADER

Apple

Apple:

Positive review percentage 3387/6315 = 53.6%,

Negative review percentage 1227/6315 = 19.4%,

Neutral review percentage

531/6315 = 8.4%

Chart, bar chart

Description automatically generated

Samsung

Chart, bar chart

Description automatically generated

Samsung:

Positive sentiment percentage 23476/37701 = 62.2%,

Negative sentiment percentage 6989/37701 = 18.5%,

Neutral sentiment percentage 3159/37701 = 8.3%

Motorola

Motorola:

positive sentiment percentage 6501/9419 = 69%,

negative sentiment percentage 1732/9419 = 18.3%,

neutral sentiment percentage 647/9419 = 6.8%.

Chart, bar chart

Description automatically generated

Xiaomi

Xiaomi:

positive sentiment percentage 3032/5574 = 54.3%,

negative sentiment percentage 627/5574 = 11.2%,

neutral sentiment percentage = 752/5574 = 13.4%

Chart, bar chart

Description automatically generated

**Conclusion**

Comparing the customer sentiments of Apple, Samsung, Motorola and Xiaomi - Apple customers are either very happy with their phones or are quite unhappy. Compared to the other 3 brands, Apple has the highest percentage of negative reviews- 19.4 percent. It has 53.6 percent positive reviews.

Samsung has lot of reviewers- 37,701. Among them, 62 percent are positive reviews, 18.5 percent are negative. Among the 4 brands, Samsung has the second highest negative sentiment percentage.

Motorola has 9400 reviews. Among them, 69 percentage are positive reviews, 18.3 percent negative reviews. We can see that Motorola has the highest positive sentiment percentage compared to the other 3 brands.

However, Xiaomi has 4.4/5 rating. It has the highest average rating compared to all other brands. However, we can see that it has only 54.3 percentage positive sentiment, a little more than Apple, and has more neutral sentiment (13.4 percent) than negative sentiment (11.2 percent).

**Future Work**

I have applied Classification models to classify product rating. Some experts suggest that we can use Regression models as well to predict the product rating based on reviews. So, in my future work I would try to use Regression models and see how it performs on our data.