MINI PROJECT

ON

AMAZON CELL Phone REVIEWS

USING IBM CLOUD

Submitted by

Eppa Lahari

(18UK1A0573)

**ABSTRACT**

The consumer reviews serve as feedback for businesses in terms of performance, product quality, and consumer service. In this research, we predict consumer opinion based on mobile phone reviews, in addition to providing an analysis of the most important factors behind reviews being classified as either positive, negative, or neutral. This insight could help companies improve their products as well as helping potential buyers to make the right decision. The research presented in this paper was carried out as follows: the data was pre-processed, before being converted from text to vector representation using a range of feature extraction techniques such as bag-of-words, TF-IDF, Glove, and word2vec. We study the performance of different machine learning algorithms, such as logistic regression, stochastic gradient descent, naive Bayes and convolutional neural networks. In addition, we evaluate our model’s using accuracy, F1-score, precision, recall and log loss function. Moreover, we apply Lime technique to provide analytical reasons for the reviews being classified as either positive, negative or neutral. Our experiments revealed that convolutional neural network with word2vec as a feature extraction technique provides the best results for both the unbalanced and balanced versions of dataset.

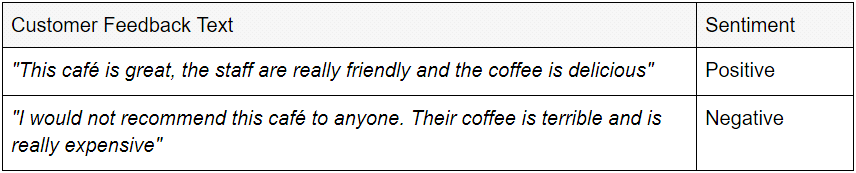
**INTRODUCTION**

OVERVIEW:

The smart phone market has been growing increasingly, not only in the conventional sales but has also been penetrating in the online shop. But not all smart phones have good quality to support the needs of consumers and it is to be noticed by the consumer. Before consumers decide to buy a smart phone, they should know the details of the specifications and functions of the smart phone, it can be learned from results of a review of smart phone users.

PURPOSE:

My project aims at building a model to predict the helpfulness of the review and the rating based on the review text. Corpus-based and knowledge-based methods can be used to determine the semantic similarity of review text. I used Natural language processing to analyze the sentiment (positive or a negative) of the given review.

**For Example:**

I also heard sentiment analysis being referred to (less commonly) as opinion mining and emotion AI.

**LITERATURE SURVEY**

EXISTING PROBLEM:

Amazon’s product review platform shows the average length of the reviews comes close to 230 characters. Sentiment analysis shows that positive sentiment is prevalent among the reviews and in terms of emotions, ‘trust’, ‘anticipation’ and ‘joy’ have highest scores. So, I was assigned to create an analysis system capable of analyzing the reviews based on NLP.

PROPOSED SOLUTION:

I started building a model to predict the helpfulness of the review and the rating based on the review text. Currently consumers who write reviews online are increasing. If the consumers read the whole review, it can consume a lot of time. But if it is read without some evaluation, it will be biased. Sentiment classification aims to overcome this problem by automatically classifying user review by positive or negative opinion.

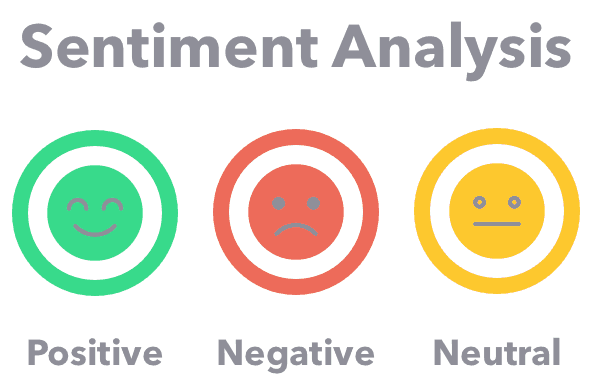
Amazon Phone Cell Reviews

This dataset result from scrapping on Amazon, but I didn’t. I got this dataset from Kaggle. overall, this data regarding reviews cell phone sales, and we got many columns for get some information from that. The reviews from the customer would be show each row with information the brand cell phone is it. So, it would be given us for analysis more deeply about this data. Before we get into the Analysis let’s write some code to build the modelling for sentiment analysis.

**Why is sentimental analysis needed?**

In today’s environment where we’re suffering from data overload (although this does not mean better or deeper insights), companies might have mountains of customer feedback collected. Yet for mere humans, it’s still impossible to analyze it manually without any sort of error or bias. Sentiment analysis provides answers into what the most important issues are. Because sentiment analysis can be automated, decisions can be made based on a significant amount of data rather than plain intuition that isn’t always right.

**Sentiment Analysis**



Sentiment Analysis is process using text analytics to obtain various data sources from the internet and various social media platforms. The objective of sentiment analysis to get some insight from the text, ex. reviews, tweets, news description and many more.

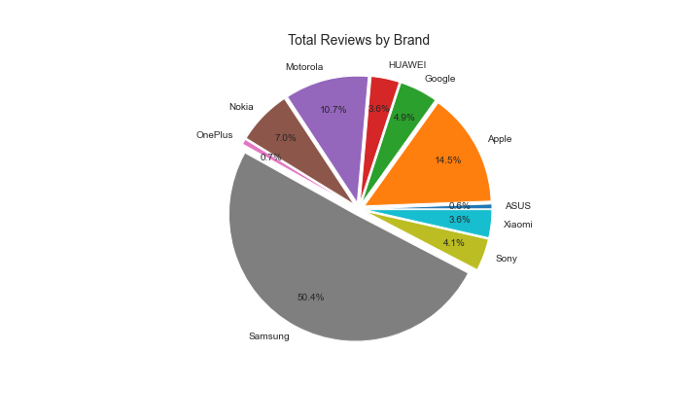
Every single day, the internet overwhelmed with million data from various sources. The sentiment analysis is very important tools for connected whole entire data. Which means, the companies would have many inputs from user or consumer efficiently.

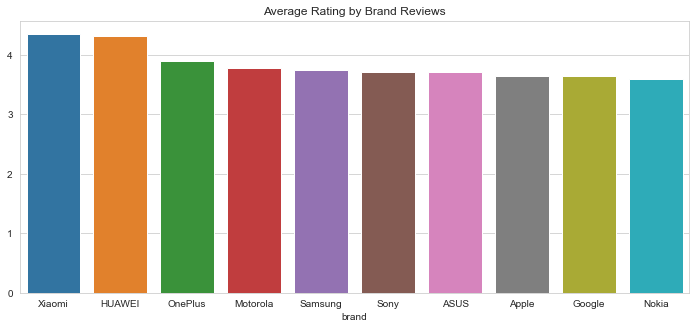
The majority of sentiment analysis approaches take one of two forms: polarity based, where pieces of texts are classified as either positive or negative, or valence-based, where the intensity of the sentiment is taken into account. For example, the words ‘good’ and ‘excellent’ would be treated the same in a polarity-based approach, whereas ‘excellent’ would be treated as more positive than ‘good’ in a valence-based approach.

Sentiment Analysis is one of the fields of Natural Language Processing (NLP) which builds a system for recognizing and extracting opinions in the text form. When we using the sentiment analysis, the information was previously is unstructured would be transformed into more structured data. After we know about the generally sentiment analysis, we get into Vader. If your big fans of Star Wars, Vader isn’t novelty for you, that is Darth Vader. But we don’t talk about that.

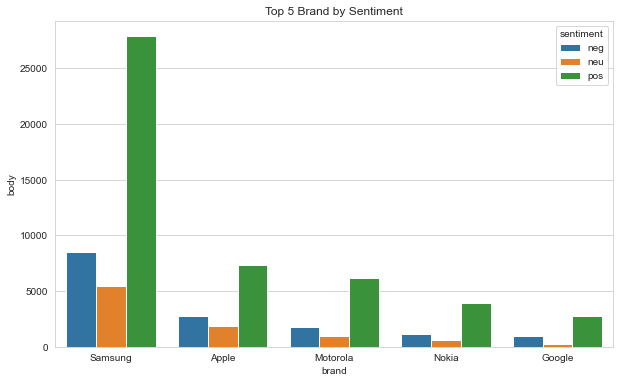
**Exploratory Data Analysis**

In the last section we try to analyze the result of Vader analysis. let’s get into the EDA for this dataset

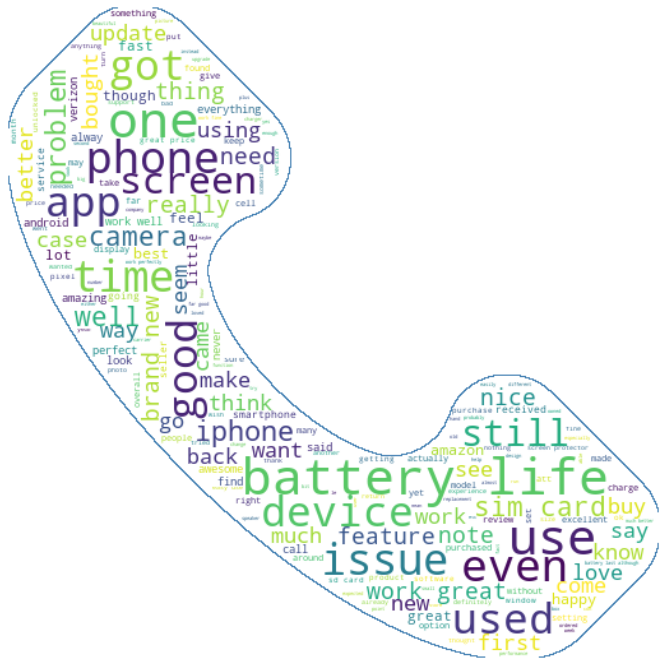




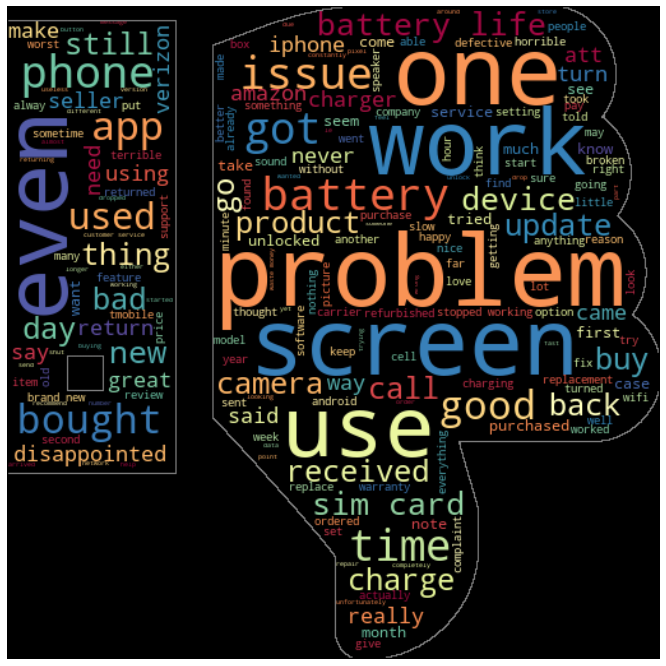
We have half data reviews dominated by **Samsung brand (50,4%)**, I assume so many users in amazon website buying Samsung brand, because you can write a review when your account buying the product. From second graph we know that top 2 on rating is **Xiamoi** and **Huawei**brand in **average rating above 4**. But they don’t have many reviews when we compare with 2 big brands on this data (looks from total reviews) **Samsung**and **Apple** brand just have average rating above 3. Let’s see how the reviews on top 5 total reviews (by frequent reviews) for each brand.



As we see on the graph Samsung brand have highest positive review than the competitor because they have a huge reviews on the amazon. Generally, we have average compound score **0.51 on negative review** and **0.68 on positive review**. Which means the model separated into good labeling for negative, neutral or positive reviews.



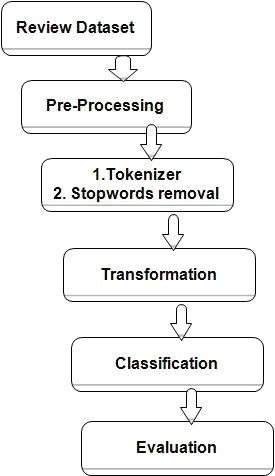
Word cloud for positive reviews



Word cloud for negative reviews

**THEORITICAL ANALYSIS**

BLOCK DIAGRAM

****

SOFTWARE DESIGNING

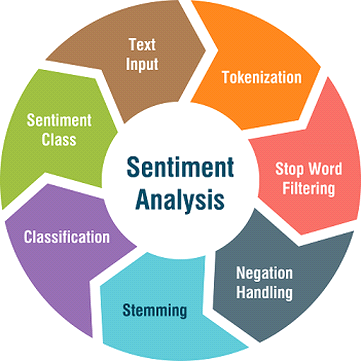
For software we would need a compatible operating system for python, java script and HTML. Software needed are:

1) TensorFlow

2) Kera’s

3) Flask

**EXPERIMENTAL INVESTIGATION**



1) Data Collection: Data for this experiment was collected from online sources. It is a dataset of size 24KB.

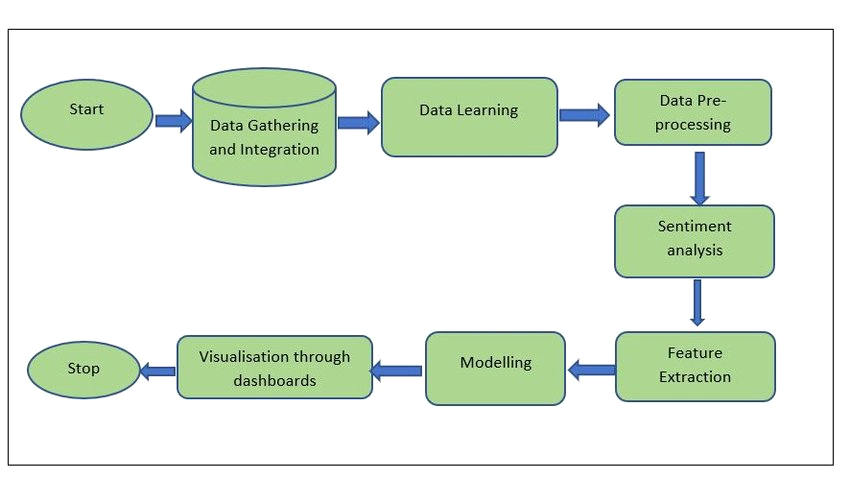
2) Initial Processing Data: I imported the dataset and using libraries like pandas and NumPy & removed the punctuations and numbers. Also, I removed the unwanted like name, date etc. I removed the null values and regular expressions. I converted each word into lower case of its own and applied stemming to remove the stop words.

3)Using rating we convert them into binary: If rating is greater than or equal to 3, then the data is “Positive”. If the rating is less than 3, then the data is “Negative”.

4)Model making: I initialized the model and added input, hidden and output layer. I configured the learning process, trained and tested the model, and optimized the model. I predicted the model by giving inputs and finally saved it.

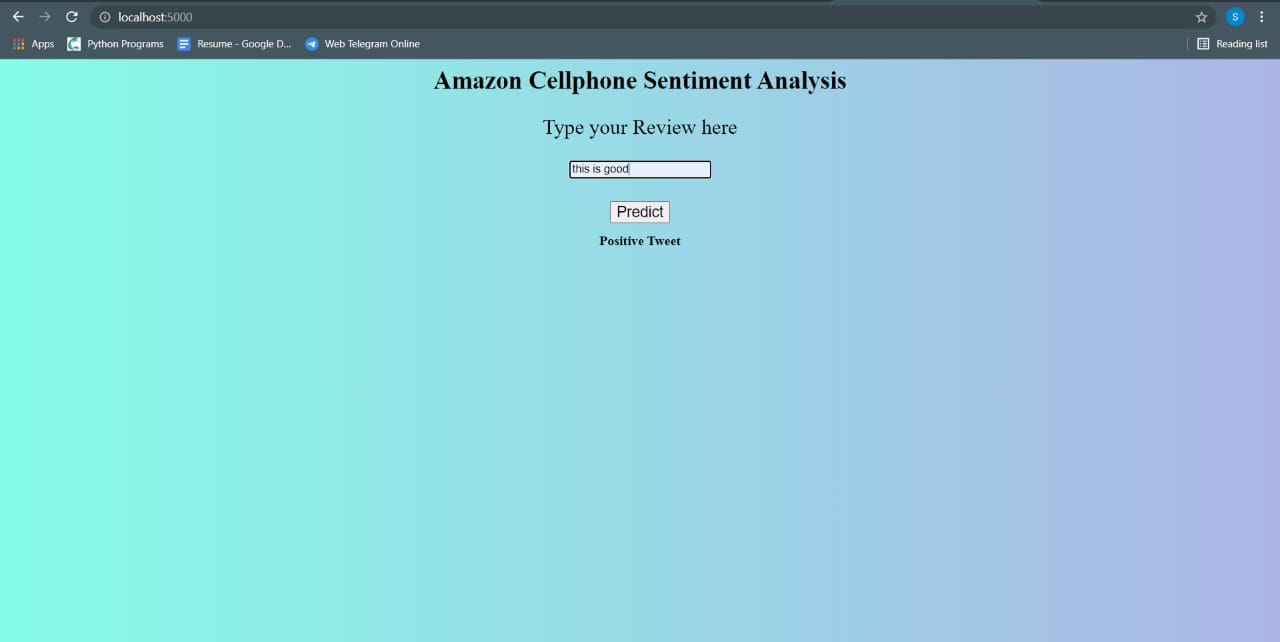
5)Preparing HTML file**:** Finally, I prepared the HTML file and ran my model. This model will predict positive and negative reviews from the input given.

**FLOWCHART**



**RESULT**

Output:



**ADVANTAGES AND DISADVANTAGES**

Advantages:

1.By using sentiment analysis, you gauge how customers feel about different areas of your business without having to read thousands of customer comments at once.

2. Increases confidence in new customers and will save their time.

3.Brings credibility to products and the company.

4.Knowing which product works best.

Disadvantages:

1. One negative review of a product or business can skew a potential customer’s view of them

2.We need to keep reviews current and up to date. Otherwise, they will seem out of date and irrelevant.

3.Disgruntled customers have the freedom to say whatever they like. This could lead to malicious or damaging information being posted.

4.Lack of touch or feel of products during online shopping is a drawback.

**APPLICATIONS**

63% of customers are more likely to purchase from a site which has user reviews. Also, to make them user friendly and easy for customers, my project analyses whether the review is a positive or a negative review and allows the customer to make good decisions.

**Conclusion**

From this word cloud, we can conclude some interesting reviews for negative or positive reviews. They have the same word in both word cloud is ‘battery life’ and ‘screen’. I assume, these two words can mean negative and positive. Let’s try an example, in positive reviews ‘battery life’ word give us information about the battery life on this product is good or vice versa in negative reviews.

**Future Research Questions**

Check VarImp() (variable importance) of the term good in cheap vs expensive phones dataset.

Check VarImp() or Frequency of terms in reviews that get 1/2/3/4/5-star ratings in order to see which terms relate to which rating. This can provide business insight on which things those specific companies should improve or leverage.

Check the term "camera" VarImp () to see how important a phone’s camera is in its customer satisfaction.

Check 1/2/3/4/5-star ratings have the greatest distribution across all observations.

For example, if observations with 4 stars have the greatest distribution, then create a model to predict 4 stars. Then, if the accuracy is good, then use terms and 4-star rating or not 4-star rating to predict whether observation is 5-star rating or 1/2/3. And so on. This is an assembly model.

Recode dependent variable: e.g., 1-2 stars = BAD. 3 stars = NEUTRAL. 4-5 stars = GOOD**.**

**BIBLIOGRAPHY**

\*Kaggle dataset link:

[https://www.kaggle.com/grikomsn/amazon-cell-phones-reviews](https://www.kaggle.com/grikomsn/amazon-cell-phones-reviews?fbclid=IwAR29SIzAkEYpaDDZUHaW-eFKOHPMuXIe2FuqEuNxgaBrjijZasOSuzTg5Ms)

**APPENDIX**

