Product Rating Generation Based On Public Opinion Using Sentiment Analysis

Thesis submitted in partial fulfilment of the requirement for the degree of

Bachelor of Computer Science and Engineering

Under the Supervision of

Dr. Jia Uddin

By

Reshan Ahmed Rabbi (13301078)

Md. Walid Amin Khan (13301034)

Md. Shaown (13301133)

Md. Shajidul Alam (13301075)

School of Engineering and Computer Science

24th December 2017

BRAC University, Dhaka, Bangladesh

Declaration

We hereby declare that this thesis has been done based on the results we obtained from our work. Due acknowledgement has been made on text. This thesis neither in parts nor as a whole have been submitted previously by anyone of any institute or university for the award of any degree.

Signature of Supervisor

_____
Dr. Jia Uddin

Signature of the Authors

_____
Reshan Ahmed Rabbi
(13301078)

_____
Md. Walid Amin Khan
(13301034)

_____
Md. Shaown (13301133)

_____
Md. Shajidul Alam
(13301075)

Acknowledgement

First and foremost, we would like to thank Almighty Allah for enabling us to initiate the research, to put our best efforts and successfully conclude it.

Secondly, we would like to thank our supervisor Dr. Jia Uddin. Without his support, guidance and contribution we could not have done it. He motivated us to do this research thoroughly and always was there to help us with any aid he could offer. This paper have been completed due to his great supervision and inspiration. We are truly grateful to him.

We revere the patronage and moral support extended with love, by our parents as well as our friends. They helped us with their direct or indirect suggestions which aided in achieving our goal. We would also like to acknowledge the assistance we received from numerous resources over the Internet especially from fellow researchers' work.

Last but not the least, we thank BRAC University for providing us the opportunity of conducting this research and for giving us the chance to complete our Bachelor degree.

Contents

List of Figure

List of Tables

List of Abbreviations

API – Application programming interface

CSV – Comma separated value

IDE – Integrated development environment\

JSON – Java script object notation

MatPlotLib – Mathematical plotting library

NLP – Natural Language Processing

NLTK - Natural language tool kit

Abstract

In modern days, people tend to check reviews and opinions on a product before buying. The main goal of our system is according to the public opinion of a product, give a rating to the product on a scale of 0 to 10. Also, we are plotting this rating for better understanding of how every attribute of a product stands against time. For public opinion, we collected data from twitter. The reason behind using twitter tweets is because it is one of the most popular social sites. The data is pre -processed and then filtered for irrelevant characters. The data is then clustered based on different attributes of product. After that using the Naïve Bayes classifier we do a sentiment analysis of the data to calculate the polarity. This polarity is then converted to a scale of 0 to 10 (where 5 is average) and thus the rating of an individual product is obtained. This polarity for each product for every attribute is plotted in a graph where axis-x represents time and axis-y represents polarity.

CHAPTER 01

Introduction

We live in an era in which people are more likely to use internet and other technologies for help. People are more dependent on modern technologies rather than conventional ways of things. This is even applicable in terms of their decision making. People tend to rely on internet or other services to help reduce their confusion. Big confusions arise while purchasing a product and people tend to search for reviews of a certain product before purchasing it. Our main target is to build a system which generates a rating of any product. This rating is solely dependent on public opinion of a product. Depending on public opinion, a products rating would be given.

1.1 Motivation

Now a days before buying any product people are more likely to see a review of a product. There are a large number of review sites on internet. There are even some pages on social networking sites like Facebook which give reviews on various product. Every type of product has review starting from phones, pc, tablets even home appliances like fans, electric irons etc.

1.2 Contribution Summary

The main purpose of doing this project is, this project will enable companies to check what attributes of the product needs to be improved. As this system will provide rating based on mass user opinion, the rating reflects the state of mind of users about that specific products and its attributes. Also there have been many researches in this sentiment analysis field but none have been done in context of business. We are trying to create a system using sentiment analysis which will be helpful in practical life. This system will also help analyze how the attributes of a product stands against the test of time. Also this system will save user a huge amount of time.

1.3 Thesis Outline

Over all in this paper, Chapter 2 provides the Background study in details including the algorithms and techniques used in the system, Chapter 3 discusses the proposed model including the algorithms and techniques, Chapter 4 presents the results and analysis and lastly Chapter 5 gives the conclusion and future work.

CHAPTER 02

Background Analysis

Sentiment analysis is a process that finds opinions, emotions from texts, tweets and other sources of natural languages [6]. All the opinion/emotion is captured using natural language processing [1]. As now a days the amount of data is getting bigger and bigger 'Natural Language Processing' is becoming more and more popular. Using natural language processing we can find out trends, popularity etc. This is big field for research thus we have done our research on generating product reviews from public sentiment. Collected from social networking site. In our research we have selected mobile phones as our product and tweets regarding phones as our experimental datasets.

In most of the review sites available over the Internet, the reviews of products are mostly given by technical wizards. People who have immense knowledge about technologies are bound to know much about any given product as they know every little details that is needed to be known in order to understand the efficiency and performance of a product. These type of people do not even need to use a product to judge, just a specification list works fine for them to give positive or negative reaction on the product [22]. In the review sites, most the reviews are given by these sort of people which means a much more technical review. This is not actually a bad thing as their judgment is far more accurate than any average person but the problem is most of the people are not well aware of a lot of attributes mentioned on the reviews [10]. Most of the people do not understand the reason behind a review being positive or negative.

Then again it is not like general reviews are not posted on these review sites but as we know the best reviews are prioritized by the service. That is why general reviews are found by facing a lot of difficulty but people tend to look over 2-3 reviews and make their decision.

Mostly non-technical people who have less knowledge of technologies tend to share their reviews and opinion on social networking sites. The reviews are actually their personal opinions about a product which are more informal and less technical in most cases. Their posts reflect the reasons of their liking or disliking a product. There is a problem though. These posts are really hard to keep track of. Even though using hashtags and other methods posts regarding

any specific topic or things can be seen, keeping track of these posts are difficult. Also most of these posts are not even taken into account due to less followers, audience type [21].

Due to all this reasons mentioned above we thought of collecting all these posts and analyzing it to create a rating system. This system will take all the tweets from twitter and then process them and using Naïve Bayes algorithm [3], do a sentiment analysis on this tweets to get the polarity [7] which further leads us to the rating of a product. Thus pushing people towards a rating system.

The product rating generation system using sentiment analysis is a user friendly system. This system is based on the public opinion. As mentioned in the background study, this system is a generalized form of reviews given by users in social networking site. For public opinion we scrapped tweets using a fake user agent script. Then we trimmed data and did a sentiment analysis on the data to get the polarity. From that polarity, the rating is generated.

2.1 Natural Language Processing

Natural language processing is focuses on machines and human language. It basically enables machines to understand human languages. Its main challenges are recognizing speech, understanding human language etc.[11]. The main syntaxes of natural language processing are parts of speech tagging, parsing, word segmentation, sentence breaking and terminology extraction. The main processes that we used will be described in the naïve bays classifier section [6].

2.2 Supervised Learning

Supervised learning is basically training the algorithm manually so that it can predict relatively correctly for future datasets [2]. In supervised learning training datasets given with the desired outputs. From all the information received by the algorithm it then predicts the probability for unknown attributes.

2.3 Bayes Theorem

For probability theory or to find out the probability of a event Bayes theorem/ Bayes rule is used. Bayes theorem follows the prior knowledge of the conditions for a specific event and then calculates the probability of a certain event occurrence.[7]

The mathematical equation of Bayes theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \quad \textbf{(1)}$$

Where P(A) and P(B) are probabilities of A and B disregarding each other. P(B|A) is the probability of B occurring depending on the occurrence of A. Finally, the answer, P(A|B) is the conditional probability of A occurring given B is true.

2.4 Naïve Bayes Classifier

While working on a dataset with millions of records with different attributes, Naive Bayes approach is recommended.

The Naïve Bayes algorithm uses Bayes Theorem with strong independent assumptions [3]. Bayes theorem works on conditional probability. Conditional probability is the probability is that something will happen given that something else has already happened. It predicts probabilities for each class such as the probability that given record or data point belongs to a particular class [7]. If there are m possible classes $A=\{a_1,a_2,\ldots\ldots a_m\}$ for tweets $T=\{t_1,t_2\ldots\ldots t_n\}$then using bayes rule we can predict the probability tweet t to be in class a

$$P(a|t) = \frac{P(a)P(t|a)}{P(t)} \quad \textbf{(2)}$$

A naïve bayes occures independently thus it assumes each term or word w_k, t_k is the frequency of each word w_k, n_d is the number of unique words then the equation becomes [7]

$$P(a|t) \propto P(a) \prod_{k=1}^{nd} [P(wk|a)]\text{^}tk \quad \textbf{(3)}$$

In Figure 1, the basic workflow of Naïve Bayes Classifier is shown. For each attribute it traverses each node and finds the probability to be in a specific class. If it finds all the values of an attribute, it goes to the next attribute. If it does not get the values, then it goes to the different nodes and checks again.

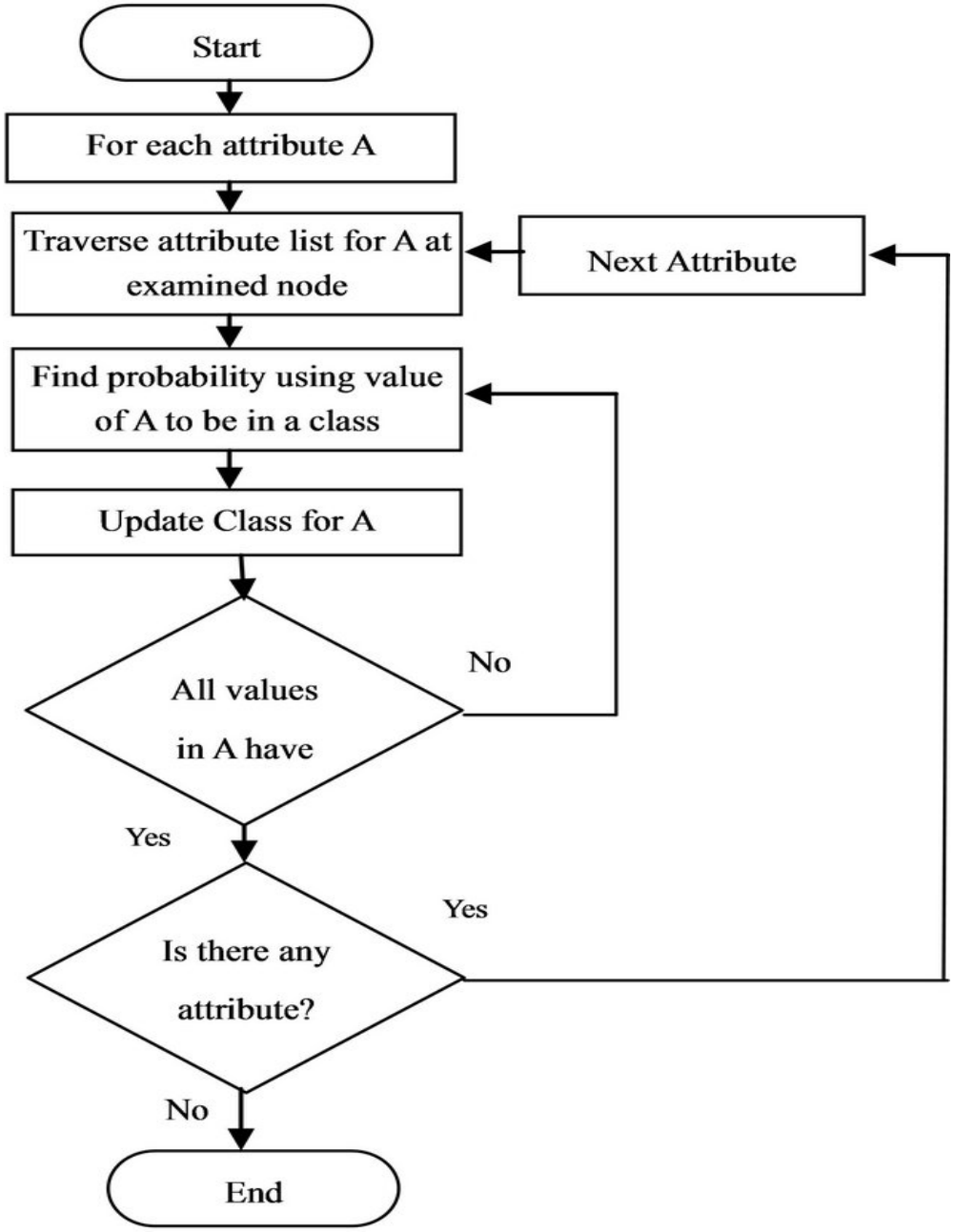


Figure 1: Naïve bayes classifier workflow

In Figure 2, Naïve Bayes pseudo code is given. First it extracts the vocabulary and has attribute counter. For each attribute, go to the node and check if the attribute belongs to the class. For each text, word is tokenized. Then the tokenized word's probability is measured. Then each word is scored. Then it returns the score.

```
TRAINMULTINOMIALNB(C, D)
 1  V ← EXTRACTVOCABULARY(D)
 2  N ← COUNTDOCS(D)
 3  for each c ∈ C
 4  do N_c ← COUNTDOCSINCLASS(D, c)
 5     prior[c] ← N_c/N
 6     text_c ← CONCATENATETEXTOFALLDOCSINCLASS(D, c)
 7     for each t ∈ V
 8     do T_ct ← COUNTTOKENSOFTERM(text_c, t)
 9     for each t ∈ V
10     do condprob[t][c] ← (T_ct+1) / Σ_t'(T_ct'+1)
11  return V, prior, condprob

APPLYMULTINOMIALNB(C, V, prior, condprob, d)
 1  W ← EXTRACTTOKENSFROMDOC(V, d)
 2  for each c ∈ C
 3  do score[c] ← log prior[c]
 4     for each t ∈ W
 5     do score[c] += log condprob[t][c]
 6  return arg max_c∈C score[c]
```

Figure 2: Naïve Bayes pseudo code

CHAPTER 03

Proposed Model

The basic idea of our project is to gather data from twitter and run a sentiment analysis on the gathered data. Then get the calculated polarity of each attributes of a product. After that the average of the attributes of the product is calculated and converted to 10 which is the rating of the product.

3.1 Process

There are 7 steps in the proposed model. Those are:

- The first step is the collection of data from twitter using a script to scrape tweets.
- The second step is preprocessing the gathered data to a supervised form.
- The third step is clustering the data according to their attributes.
- The fourth step is building a list of positive and negative words is added.
- The fifth step is adding external attributes of products.
- The sixth step is to do tokenizing and parts of speech tagging.
- The seventh step is to do a sentiment analysis on the data to get the polarity.
- The eighth step is to show graphical representations for each attribute.
- The final step is to generate a rating based on the polarity.

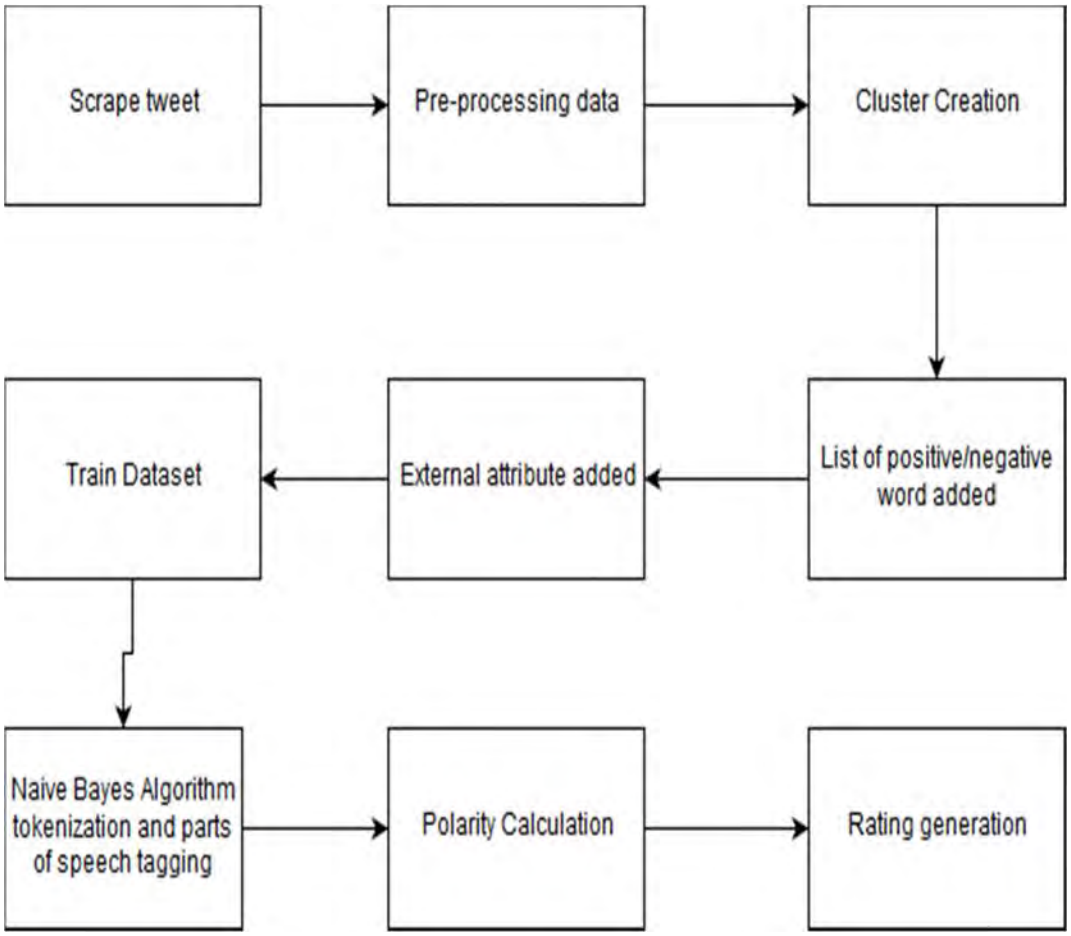


Figure 3: Proposed Model's workflow

3.2 Data Collection

For gathering data, we used a script to scrape Tweets from Twitter using the Python package which requests to retrieve the contents and Beautiful soup4 to parse the retrieved content.

Our scraper also uses –

- lxml – A library for processing XML and HTML in Python
- coala-utils 0.5.0 – A collection of utilities.
- fake_useragentrequests - simple useragent faker with real world database.

We did not use the Twitter search API/OAuth key as by using it we can only access tweets from past 7 days. Also twitter search API only allows us to send 180 requests every 15 minutes. For this data from past is needed as that would give more precise results but using Twitter search API would only let us access data from past 7 days. Our script allows us to scrape tweets from any date to any date. Also we needed a large amount of data so limited request would have been troublesome to collect data. Our script has no limitation like that. In within few seconds plenty of tweets are scraped for use. So, our script allows us to overcome these limitations which Twitter search API cannot.

Per tweet this script scrapes the following information –

1. Username & Full name
2. Tweet id
3. Tweet text
4. Timestamp
5. Number of likes
6. Number of replies
7. Number of retweets

Figure 4 is the twitter advanced search. Here by giving the right parameters, tweets are assembled. Then some part of link of the search is given on the script which finally leads to data collection.

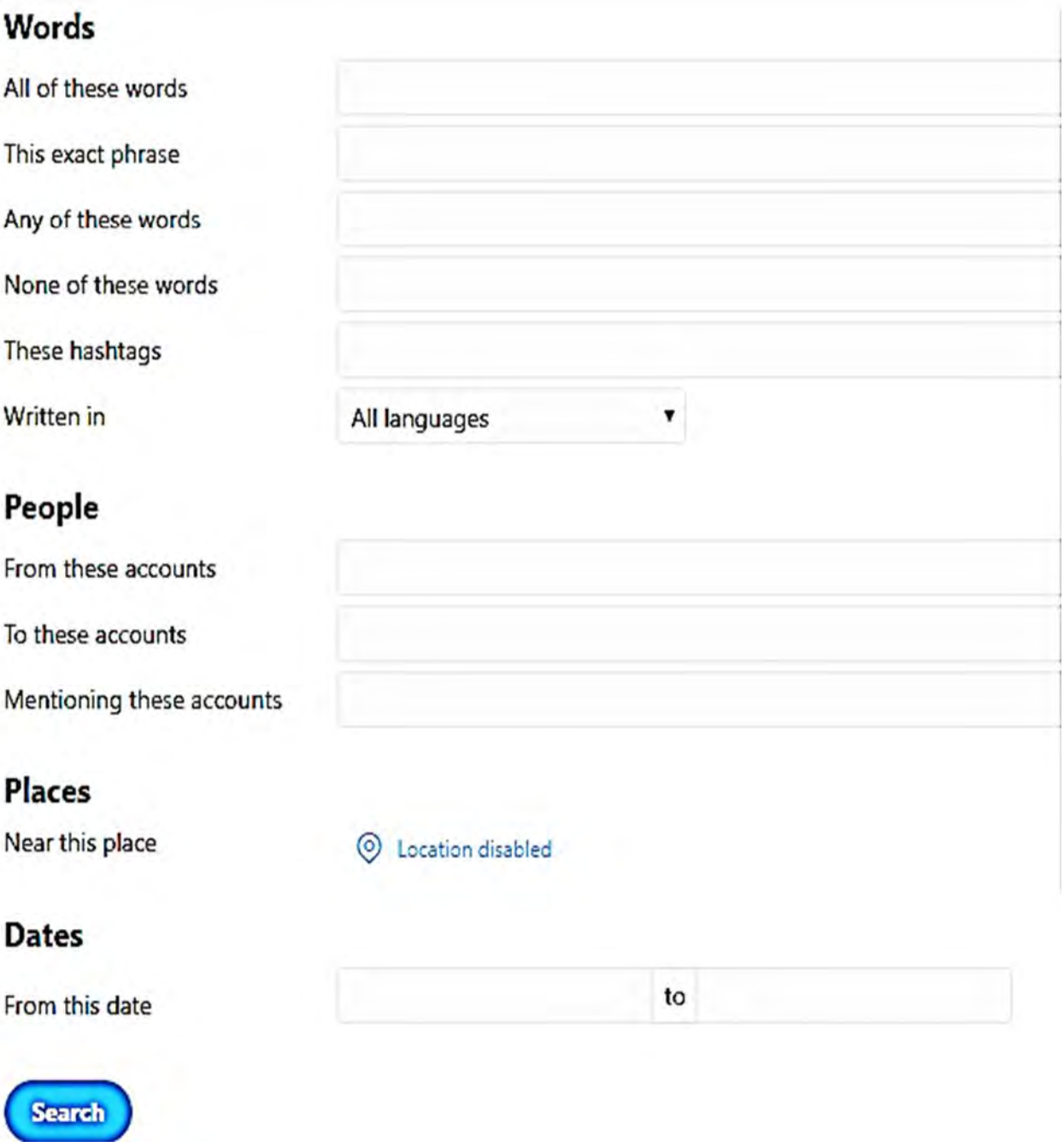


Figure 4: Twitter advanced search

3.2 Preprocessing and Filtering Data

After the collection of data which is in JSON format, collected data's need to be preprocessed in a supervised form. It means there cannot be any punctuation or additional symbol. To preprocess the data we used a java program. This program removes all the unnecessary symbols, links, punctuations and other unimportant symbols from the data and saves it as a CSV file.

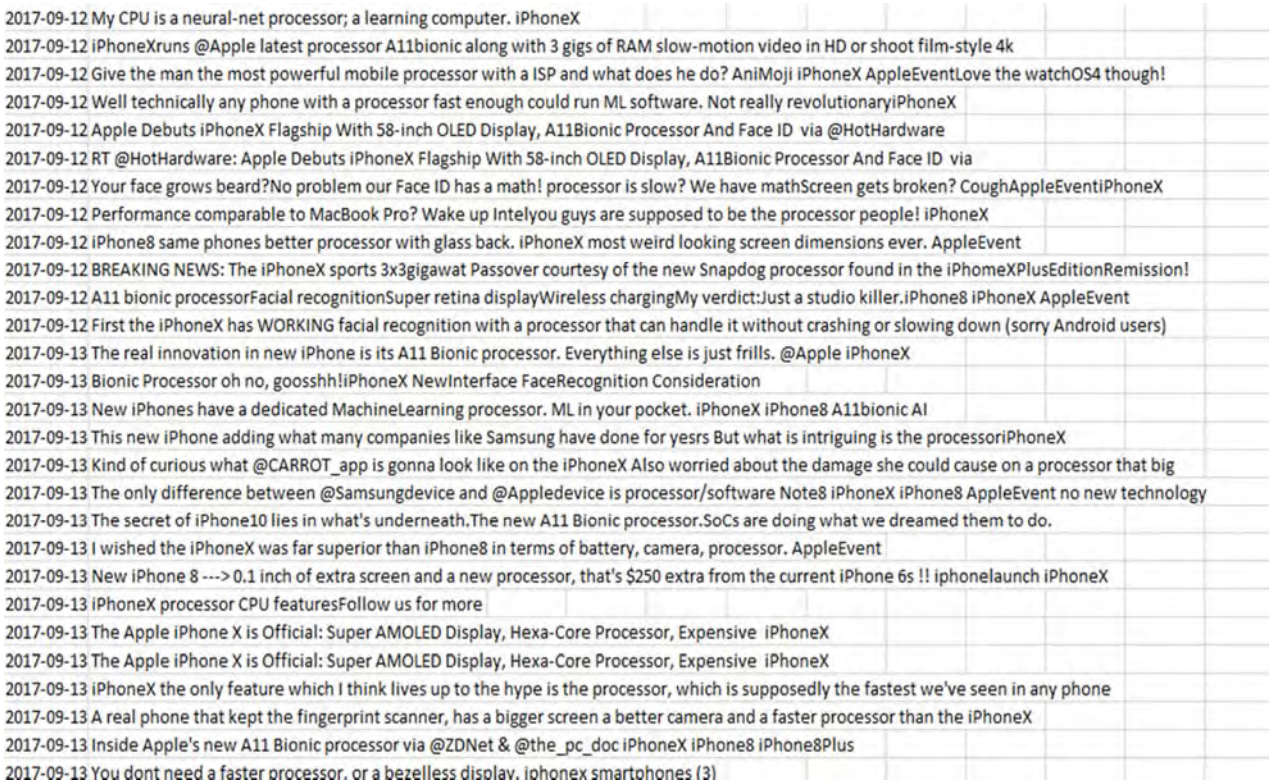
To remove garbage data caused by script's parameter username and full name, number of likes, number of retweets, number of replies are removed by just eliminating these from the script. However id and timestamp data could not be removed using the same technique. So to remove those, the java program was used again to remove the id and timestamp from data to filter it and just keep the text or tweet of a user.

The program then saves this processed and filtered supervised data in a CSV file. In Figure 5 the JSON format data is shown before pre-processing.


[{"id": "906835073437192192", "text": "Get ready for #iPhone8/#iPhoneX. Add ARMv8.1 opcode support to IDA with this little processor extender plugin https://gith
-to-ios-11-leak\u00a0\u2026 pic.twitter.com/kK5upMQ9qS", "timestamp": "2017-09-10T17:18:17"}, {"id": "906929836827992064", "text": "A11 Fusion in iPhone X appear
tamp": "2017-09-10T17:24:43"}, {"id": "906931522199896065", "text": "'#A11Fusion' in #iPhoneX appears to be a six core processor, according to #iOS11 leak http:/
ion-in-iphone-x-appears-to-be-a-six-core-processor-according-to-ios-11-leak\u00a0\u2026 #iPhoneX", "timestamp": "2017-09-10T18:15:39"}, {"id": "907049430288945150", "text": "#iPhoneX may pack a 6-core processor https://www.engadget.com/2017/09/10/iphone-x-6-core-processor-leak/\u00a0\u2026 #technology #feedly", "timestamp": "2017-09-11T07:07:36"}, {"id": "907290495331729408", "text": "#iPhoneX may pack a 6-core processor http://j.mp/2vQ7V7O\u00a0 #Apple #Mobile pic.twitter.com/X9UoWxGtpp", "timestamp": "2017-09-11T14:05:07"}, {"id": "907250097309831172", "text": "\"iPhone X A11 chip with six processor cores \u03b5\u03af\u03bd\u03b1\u03b9 \u03c4\u03bf\u03c5 iPhone X \u03bc\u03b5 6 processor cores \u03b5\u03af\u03bd\u03b1\u03b9 \u03c4\u03bf\u03c5 @jonfingaspic.twitter.com/X9UoWxGtpp", "timestamp": "2017-09-11T14:05:07"}, {"id": "907290495331729408", "text": "#iPhoneX may pack a 6-core processor http://j.mp/2vQ7V7O\u00a0 #Apple #Mobile pic.twitter.com/X9UoWxGtpp", "timestamp": "2017-09-11T17:07:36"}, {"id": "907676774737301505", "text": "So, #iPhone10 has an insane camera, reliable FaceID, fast processor, Airpower, no buttons, and 2.5hr longer battery announced, comes in two colors: space gray and a polished silver. Does feature the A11 Bionic processor and...", "timestamp": "2017-09-12T18:58:38"}, {"id": "907673295637676033", "text": "My #iPhoneX CPU is a neural-net processor; a learning computer. But luckily Skynet presets the switch to read-only when w
43:33"}, {"id": "907676774737301505", "text": "So, #iPhone10 has an insane camera, reliable FaceID, fast processor, Airpower, no buttons, and 2.5hr longer batter
ly announced, comes in two colors: space gray and a polished silver. Does feature the A11 Bionic processor and...", "timestamp": "2017-09-12T18:58:38"}, {"id": "
aster processor. #WaitAndSave", "timestamp": "2017-09-12T19:29:33"}, {"id": "907691234222895104", "text": "\"My CPU is a neural-net processor; a learning compute
buts-iphone-x\u00a0\u2026 via @HotHardware", "timestamp": "2017-09-12T20:45:11"}, {"id": "907706863294320647", "text": "RT @HotHardware: #Apple Debuts #iPhoneX F
7721160367816705", "text": "BREAKING NEWS: The #iPhoneX sports 3x3gigawat\u2122 Passover courtesy of the new Snapdog processor found in the #iPhomeXPlusEditionRe
com/p/BY9k61gj9kD/\u00a0", "timestamp": "2017-09-13T01:06:03"}, {"id": "907785108396142593", "text": "New iPhones have a dedicated #MachineLearning processor. ML
Phone10 lies in what's underneath.\nThe new A11 Bionic processor.\nSoCs are doing what we dreamed them to do.", "timestamp": "2017-09-13T03:19:21"}, {"id": "907866"}, {"id": "907902919084670977", "text": "The Apple iPhone X is Official: Super AMOLED Display, Hexa-Core Processor, Expensive http://filipinotechaddict.net/201
#iphonex #smartphones (3)", "timestamp": "2017-09-13T16:33:06"}, {"id": "908012537572081665", "text": "iPhone X A11 Bionic processor defeats MacBook Pro in #Geek
"timestamp": "2017-09-14T06:53:28"}, {"id": "908237118216044545", "text": "Relieved that we get a spaceshuttle processor for the 3-4 apps we are using daily and
1", "text": "The iPhone X\u2019s processor is more powerful than the newest MacBook Pro http://bgr.com/2017/09/14/iphone-x-vs-iphone-8-a11-bionic-benchmarks-macb
int sensor/home button when #iPhone8Plus is $200 cheaper and same #atom11 processor", "timestamp": "2017-09-14T22:05:14"}, {"id": "908500593626566657", "text":
ne X is sourced from Samsung... \n#iPhoneX #iPhone8", "timestamp": "2017-09-15T18:57:12"}, {"id": "908790068759994369", "text": "The #iPhoneX's #processor is mor
{"id": "908908689641361408", "text": "#iPhoneX may pack a 6-core processor #mobile #uxdesign #iOSApp #Android", "timestamp": "2017-09-16T04:21:32"}, {"id": "909
pl http://bit.ly/2xskKJl\u00a0pic.twitter.com/GPUJjGO8u4", "timestamp": "2017-09-18T14:05:34"}, {"id": "909809652556337152", "text": "GeekBench 4 points of iPhon
, "text": "Can #GooglePixel2 processor beat #iphonex #A11 #bionic chip in speed test ?? #android #Oreo #iOS11 #Apple", "timestamp": "2017-09-21T06:54:03"}, {"id"
the new #iPhoneX will be: Will the audio processor be improved and how much better is the water resist rating", "timestamp": "2017-10-18T01:46:43"}, {"id": "920520037", "text": "#Apple vs Samsung, un nuovo processo per ricalcolare il risarcimento http://ift.tt/2xiD3xG\u00a0 #Settembre #iPhoneX #iOS11 #Apple #Tech #RT", "
iOS11 #Apple #Tech #RT", "timestamp": "2017-10-26T14:37:30"}, {"id": "923588252765970433", "text": "#Apple vs Samsung, un nuovo processo per ricalcolare il risar
ice.", "timestamp": "2017-10-27T08:58:27"}, {"id": "923879549468794880", "text": "#QStick have Atom x7-z8750 processor with #8GB of #RAM with LPDDR3-1600 as RAM
ceApple's own A11 Bionic chip is the fastest smartphone processor but @CNBC cautions that \"the iPhone X gets hot\" pic.twitter.com/wF3dp4mLAw", "timestamp": "20
"}, {"id": "927709874976444416", "text": "@reneritchie not sure if it\u2019s the processor or the new screen tech, but man is this #iPhoneX buttery, almost predi

Figure 5: Data before processing in JSON format

In Figure 6, we are demonstrating the filtered data which is saved in CSV format. Here the first column represents posting date of the corresponding tweet that is in the second column. In this format, we omitted the punctuations, unnecessary symbols and all types of links and extra whitespaces from data.



2017-09-12 My CPU is a neural-net processor; a learning computer. iPhoneX
2017-09-12 iPhoneXruns @Apple latest processor A11bionic along with 3 gigs of RAM slow-motion video in HD or shoot film-style 4k
2017-09-12 Give the man the most powerful mobile processor with a ISP and what does he do? AniMoji iPhoneX AppleEventLove the watchOS4 though!
2017-09-12 Well technically any phone with a processor fast enough could run ML software. Not really revolutionaryiPhoneX
2017-09-12 Apple Debuts iPhoneX Flagship With 58-inch OLED Display, A11Bionic Processor And Face ID via @HotHardware
2017-09-12 RT @HotHardware: Apple Debuts iPhoneX Flagship With 58-inch OLED Display, A11Bionic Processor And Face ID via
2017-09-12 Your face grows beard?No problem our Face ID has a math! processor is slow? We have mathScreen gets broken? CoughAppleEventiPhoneX
2017-09-12 Performance comparable to MacBook Pro? Wake up Intelyou guys are supposed to be the processor people! iPhoneX
2017-09-12 iPhone8 same phones better processor with glass back. iPhoneX most weird looking screen dimensions ever. AppleEvent
2017-09-12 BREAKING NEWS: The iPhoneX sports 3x3gigawat Passover courtesy of the new Snapdog processor found in the iPhomeXPlusEditionRemission!
2017-09-12 A11 bionic processorFacial recognitionSuper retina displayWireless chargingMy verdict:Just a studio killer.iPhone8 iPhoneX AppleEvent
2017-09-12 First the iPhoneX has WORKING facial recognition with a processor that can handle it without crashing or slowing down (sorry Android users)
2017-09-13 The real innovation in new iPhone is its A11 Bionic processor. Everything else is just frills. @Apple iPhoneX
2017-09-13 Bionic Processor oh no, goosshh!iPhoneX NewInterface FaceRecognition Consideration
2017-09-13 New iPhones have a dedicated MachineLearning processor. ML in your pocket. iPhoneX iPhone8 A11bionic AI
2017-09-13 This new iPhone adding what many companies like Samsung have done for yesrs But what is intriguing is the processoriPhoneX
2017-09-13 Kind of curious what @CARROT_app is gonna look like on the iPhoneX Also worried about the damage she could cause on a processor that big
2017-09-13 The only difference between @Samsungdevice and @Appledevice is processor/software Note8 iPhoneX iPhone8 AppleEvent no new technology
2017-09-13 The secret of iPhone10 lies in what's underneath.The new A11 Bionic processor.SoCs are doing what we dreamed them to do.
2017-09-13 I wished the iPhoneX was far superior than iPhone8 in terms of battery, camera, processor. AppleEvent
2017-09-13 New iPhone 8 ---> 0.1 inch of extra screen and a new processor, that's $250 extra from the current iPhone 6s !! iphonelaunch iPhoneX
2017-09-13 iPhoneX processor CPU featuresFollow us for more
2017-09-13 The Apple iPhone X is Official: Super AMOLED Display, Hexa-Core Processor, Expensive iPhoneX
2017-09-13 The Apple iPhone X is Official: Super AMOLED Display, Hexa-Core Processor, Expensive iPhoneX
2017-09-13 iPhoneX the only feature which I think lives up to the hype is the processor, which is supposedly the fastest we've seen in any phone
2017-09-13 A real phone that kept the fingerprint scanner, has a bigger screen a better camera and a faster processor than the iPhoneX
2017-09-13 Inside Apple's new A11 Bionic processor via @ZDNet & @the_pc_doc iPhoneX iPhone8 iPhone8Plus
2017-09-13 You dont need a faster processor, or a bezelless display, iphonex smartphones (3)

Figure 6: Data after processing

3.3 Data Clustering

The supervised data saved in a CSV file is then clustered according to specific attributes. Let's say our system is rating a phone. There are many attributes of a phone which the user will want to know about like processor, camera, battery life etc. Now using these attribute name as keyword we would cluster the data and separate the dataset according to their attribute. We created the cluster to make the sentiment analysis relatively easier as we can easily show the graphical representation.

3.4 Word list and attribute addition

To make the classifier more efficient we added a list of positive and negative words [12, 13] with the provided wordlist by python NLTK. We also added some external attributes/features for our experimental product (phones). Like processor class includes words like Qualcomm, Snapdragon, ARM cortex, MediaTek, Samsung Exynos, Apple A series, etc. We did the same for UI and camera attribute.

3.5 Training Dataset

At first we used datasets made by us from collecting twitter then classifying them by hand. After that we used a dataset of amazon phone review [15].

3.6 Tokenization and POS tagging

After reading the tweet the classifier first tokenizes the words based on comma (,), full stop (.) space and many other punctuation. After that the stop words are removed. By using parts of speech tagging we can indicate subjectivity of a tweet better [1][16]. Parts of speech tagging was done by NLTK pos tagger. Which was pre-trained by python NLTK.

3.7 Sentiment Analysis on Data

Sentiment Analysis is the process of detecting the contextual polarity of text. In other words, it determines whether a piece of writing is positive, negative or neutral. After clustering the data, we did sentiment analysis on the datasets. For sentiment analysis we used the Naïve Bayes classifier algorithm in TextBlob [3].

This algorithm is used for predicting the probability of words being in any particular class. This is used due to its ease during both training and classifying steps. Preprocessed data is given as input to train the classifier and that model is applied on test to generate positive or negative or neutral sentiment.

In Figure 7, we can see the Spyder IDE console giving outputs for the tweets that were used as input (filtered CSV file). The two values under each tweet represents the polarity and subjectivity of the sentence respectively.

```
REVEALED: SA release date and price for new Apple Watchiphonex iphone8
0.13636363636363635 0.45454545454545453
RT @better_buy_ins: iPhoneX insurance BEST PRICE GUARANTEED! RT and Follow to win 1 of 5 M&S 20 gift cards Competition giveaway ends 30 Nov.
0.9 0.35
3-day SALE Get the Spigen Rugged Armor Slim iPhone X case at this limited time price!!  ebay ecommerce blackfriday iphonex iphonexcases mobilecases mobile apple
la dallas letgo offerup amazon spigen phonecases onlineshopping
-0.11160714285714285 0.14285714285714285
Made the mistake of finally taking the case off my iPhoneX Now I never want to put it back on but honestly the glass back and repair price scare me enough to put
case back on when leaving the house.
0.09999999999999999 0.39999999999999997
WhatThePrice is back with a bang! Each RT drops the price of iPhoneX by Rs. 50! Isn't that amazing? Hit the RT button NOW!YayvoWhiteFriday17 AbLootMachegi
0.3000000000000004 0.45
RT:WhatThePrice is back with a bang! Each RT drops the price of iPhoneX by Rs. 50! Isn't that amazing? Hit the RT button NOW! BlackFriday, Black_Friday
0.37500000000000006 0.45
iPhoneX is worth to buy because everything was good. It does a premium phone, I like the grip when I hold it. But sorry Apple, the price is too much for me.Perha
the iPhone8?
0.175 0.47500000000000003
The only thing I need this BlackFriday is a discounted price on IPhoneX . So no thanks I much rather sit back relax & watch Mnangagwa swearing in.
0.025 0.35
Retweeted Yayvo by TCS (@Yayvo_TCS):WhatThePrice is back with a bang! Each RT drops the price of iPhoneX by
0.0 0.0
iPhone X  !!!  9000.-  0% 10iPhone iPhoneX Apple SHOCK_PRICE Promotion  Online  shopping changsphopping promotions Sale
0.0 0.0
```

Figure 7: sentence polarity

3.8 Graphical Representation

For graphical representation, we used the function named MatPlotLib. This is a library of python that is mostly used for creating graphical representation for example: bar charts, pie charts, histogram etc.

The library requires three parameters. One input is the polarity, the second one is the brand and the last one is attribute. These are given as input and the outputs are shown in graphs. In Figure 8 we have shown a sample graphical representation of sentiment analysis result.

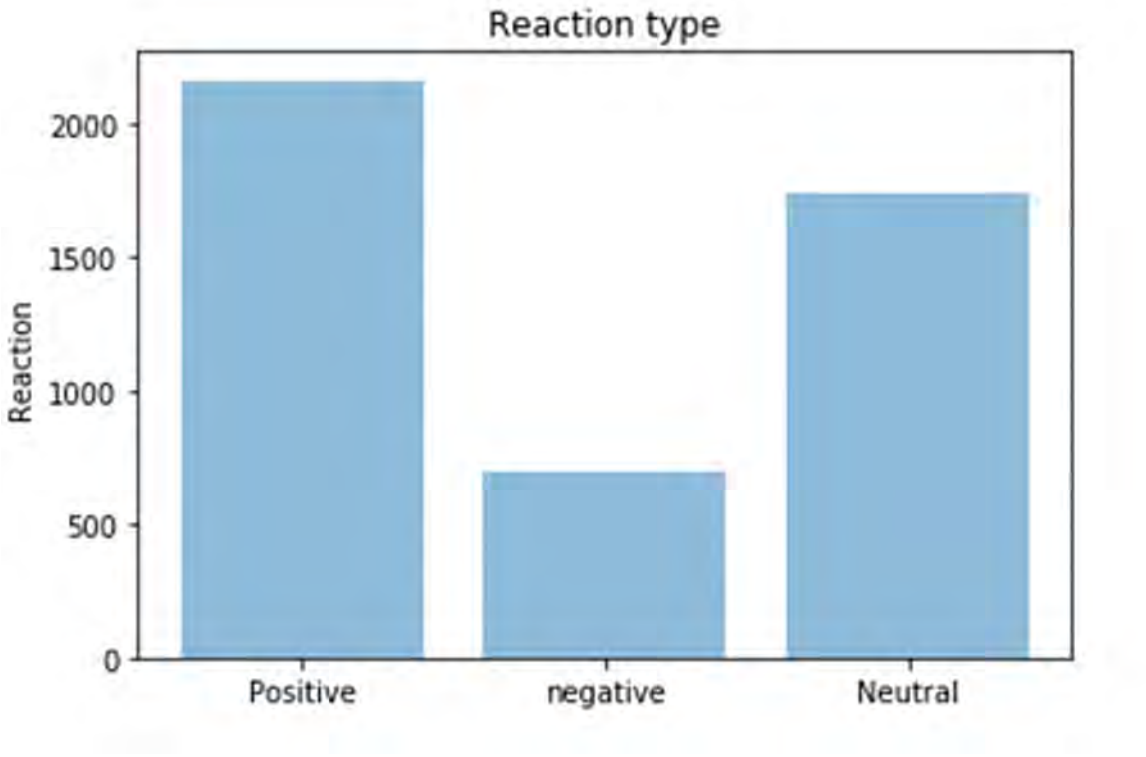


Figure 8: Initial polarity graph

3.9 Rating generation

Rating is calculated by averaging the polarity (positive, negative and neutral) of all attributes. As these polarities are fraction values, we convert it to decimal, round it up using ceiling/floor and show the rating. As neutral polarities are considered to have a polarity of 0 and it does not affect the average polarity count, these values are not taken in order to make our system more efficient. We save the number of positive, negative tweets and we average them to find out the rating on a scale of 1 to 10. This is how the polarity is calculated of a single product.

CHAPTER 04

Experimental Setup & Result Analysis

The experiment of this algorithm in this thesis is performed in the AMD FX^{tm}-8300 Eight-Core Processor, ~3.3GHz, 16384MB RAM environment using Spyder IDE. The codes were done in python v3.6.2. We used TextBlob library for building the Naïve Bayes classifier. We used pythons NLTK (natural language toolkit) for natural language processing basics. Our training dataset for naïve Bayes classifier were manual datasets we collected from twitter and later we used a dataset of amazon unlocked mobile [15].

In the classifier we have calculated the polarity. The polarity range is (-1.0 to 1.0) and if the polarity is less than 0 then the sentence is negative. If the polarity of the sentence is 0.0 then the sentence is neutral. Thus if the polarity greater than 0.0 then the sentence is positive.

- Positive (>0.0)
- Neutral (0.0)
- Negative (<0.0)

Figure 9 shows the graph for the review of processors of different phone models. The graph shows the positive and negative polarity of the sentiment.

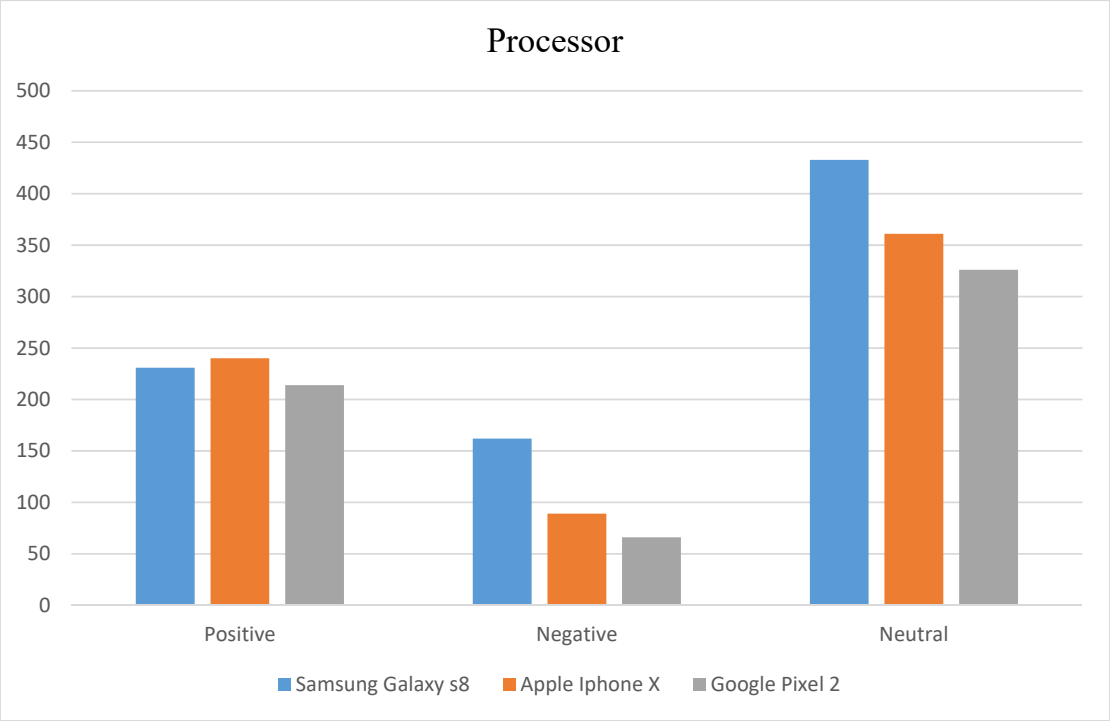


Figure 9: Processor sentiment graph

In Figure 9, the bars represent the number of positive, negative and neutral sentiment of Samsung Galaxy S8, IPhone X and Google Pixel 2's processor. Blue bar illustrates Samsung Galaxy S8, Orange bar illustrates Apple IPhone X and Gray bar illustrates Google pixel2.

Table 1: Processor attribute based sentiment count table.

	Samsung Galaxy S8	Apple iPhone X	Google Pixel 2
Total	826	240	606
Positive	231	59	214
Negative	162	361	66
Neutral	433	690	326

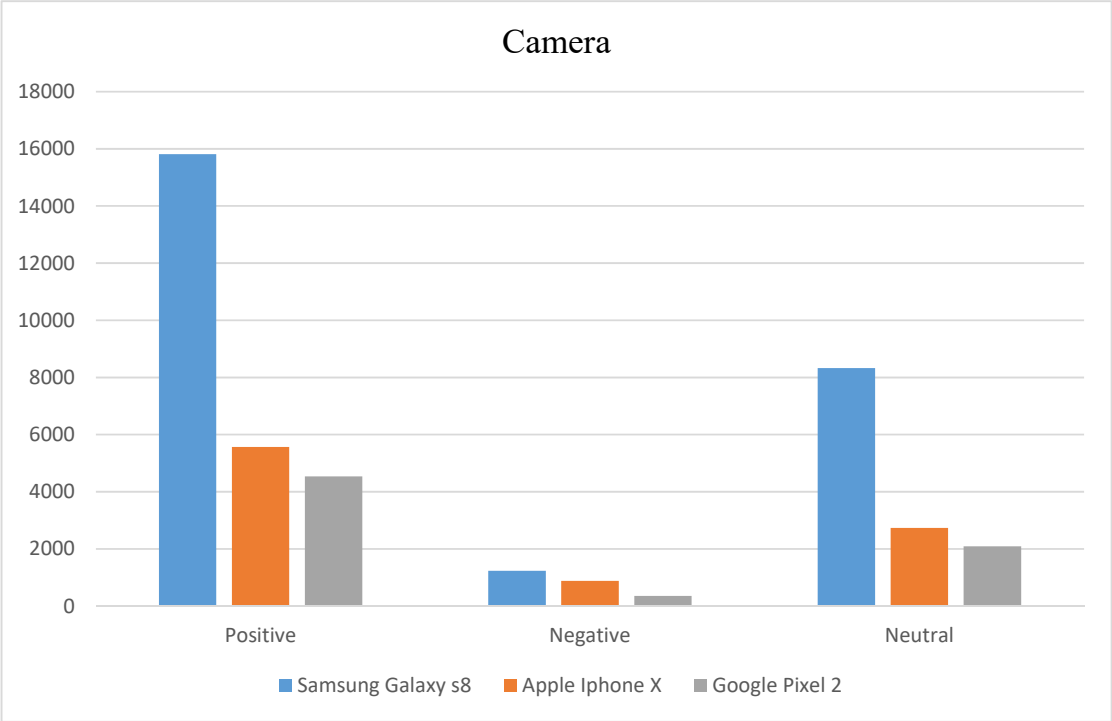


Figure 10: Camera sentiment graph

In Figure 10, the bars represent the number of positive, negative and neutral sentiment of Samsung Galaxy S8, IPhone X and Google Pixel 2's camera. Blue bar illustrates Samsung Galaxy S8, Orange bar illustrates Apple IPhone X and Gray bar illustrates Google pixel2.

Table 2: Camera attribute based sentiment count table.

	Samsung Galaxy S8	Apple iPhone X	Google Pixel 2
Total	24747	9289	6989
Positive	15182	5669	4537
Negative	1236	881	355
Neutral	8329	2739	2097

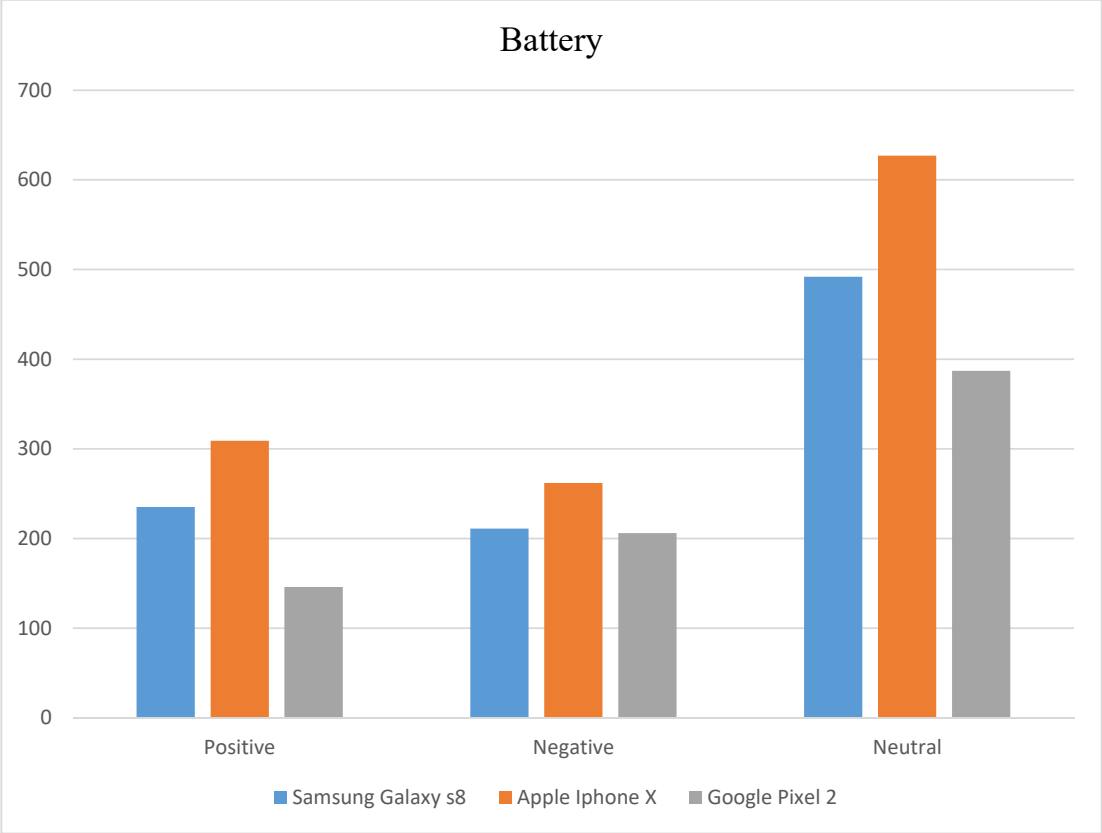


Figure 11: Battery sentiment graph

In Figure 11, the bars represent the number of positive, negative and neutral sentiment of Samsung Galaxy S8, IPhone X and Google Pixel 2's battery. Blue bar illustrates Samsung Galaxy S8, Orange bar illustrates Apple IPhone X and Gray bar illustrates Google pixel2.

Table 3: Battery attribute based sentiment count table.

	Samsung Galaxy S8	Apple iPhone X	Google Pixel 2
Total	938	1198	739
Positive	235	309	146
Negative	211	262	206
Neutral	492	627	387

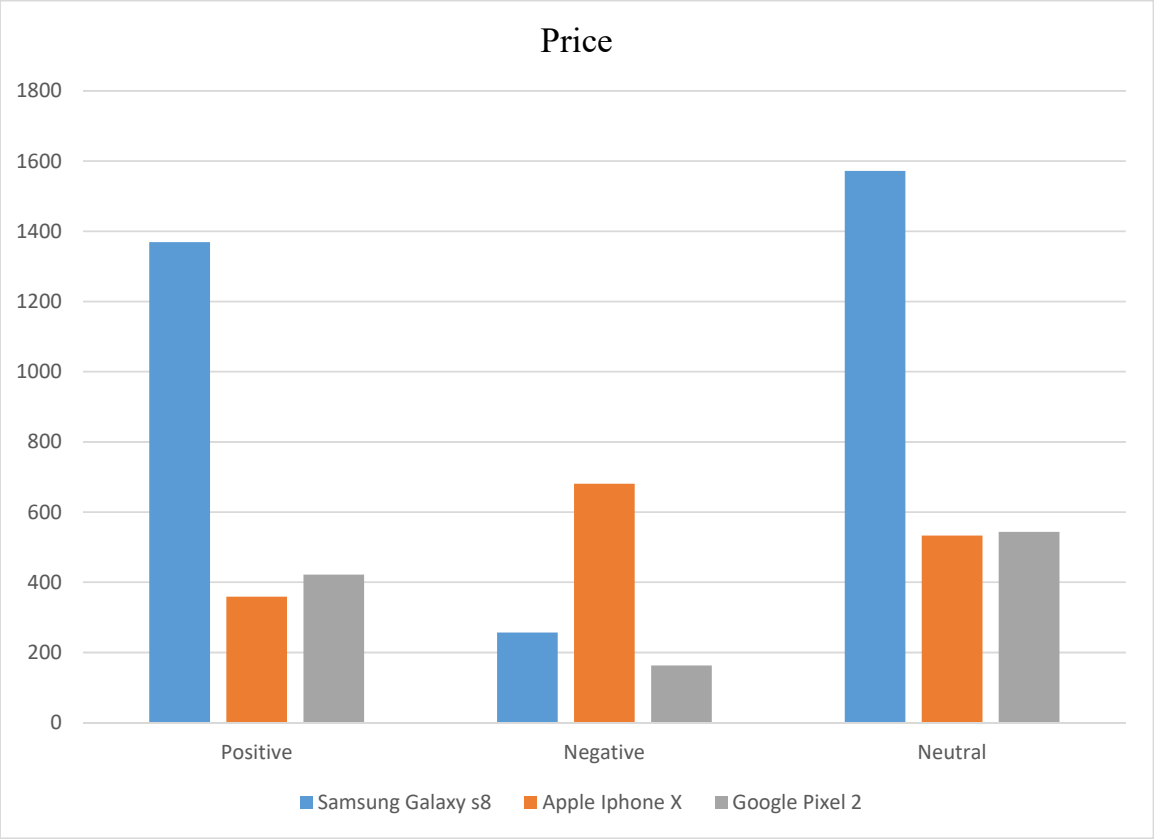


Figure 12: Price sentiment graph

In Figure 12, the bars represent the number of positive, negative and neutral reviews of Samsung Galaxy S8, IPhone X and Google Pixel 2's price. Blue bar illustrates Samsung Galaxy S8, Orange bar illustrates Apple IPhone X and Gray bar illustrates Google pixel2.

Table 4: Price attribute based sentiment count table.

	Samsung Galaxy S8	Apple iPhone X	Google Pixel 2
Total	3198	1573	1129
Positive	1369	359	422
Negative	257	681	163
Neutral	1572	533	544

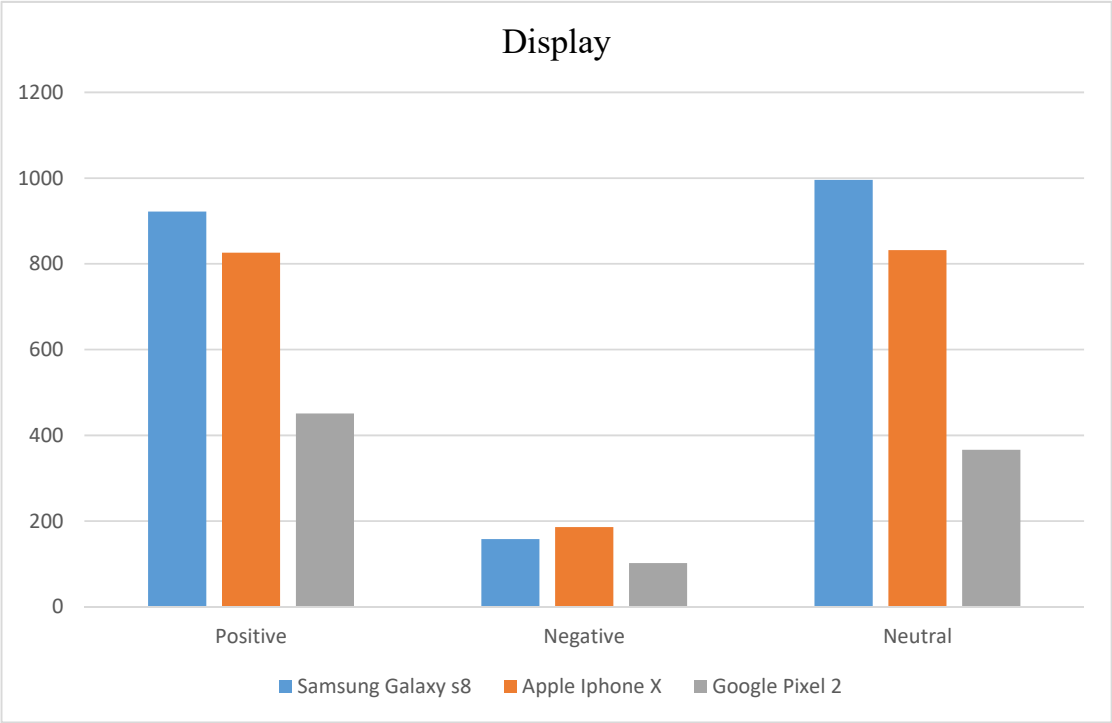


Figure 13: Display sentiment graph

In Figure 13, the bars represent the number of positive, negative and neutral reviews of Samsung Galaxy S8, IPhone X and Google Pixel 2's display. Blue bar illustrates Samsung Galaxy S8, Orange bar illustrates Apple IPhone X and Gray bar illustrates Google pixel2.

Table 5: Display attribute based sentiment count table.

	Samsung Galaxy S8	Apple iPhone X	Google Pixel 2
Total	2076	1844	919
Positive	922	826	451
Negative	158	186	102
Neutral	996	832	366

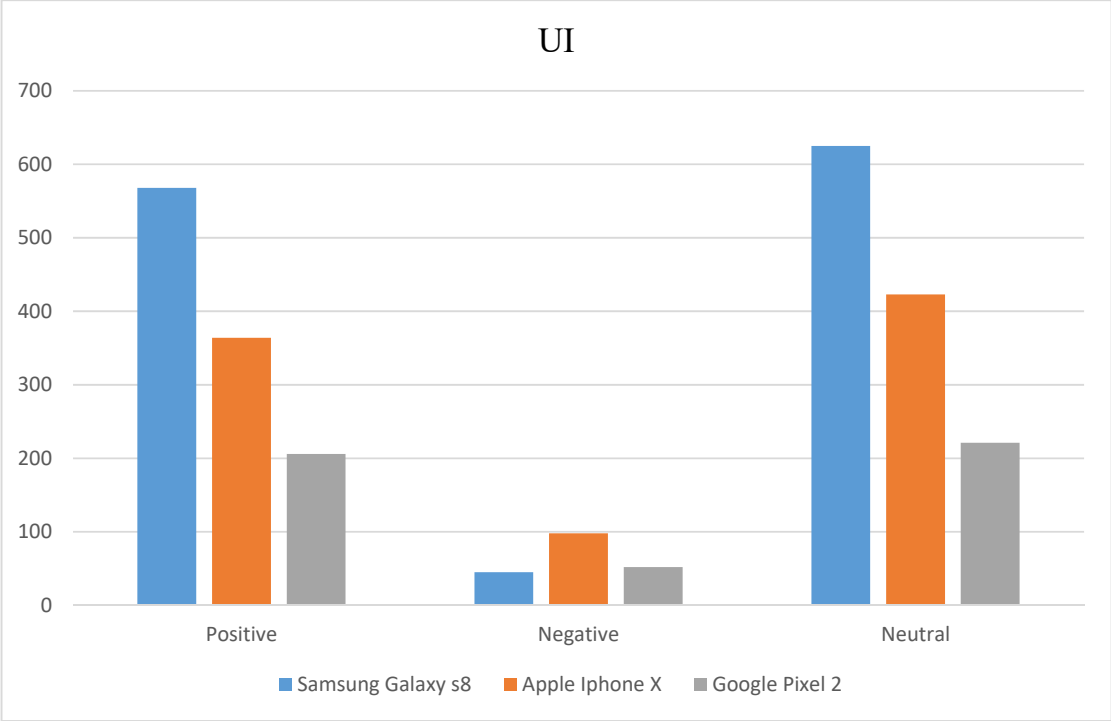


Figure 14: UI sentiment graph

In Figure 14, the bars represent the number of positive, negative and neutral reviews of Samsung Galaxy S8, IPhone X and Google Pixel 2's UI. Blue bar illustrates Samsung Galaxy S8, Orange bar illustrates Apple IPhone X and Gray bar illustrates Google pixel2.

Table 6: User Interface attribute based sentiment count table.

	Samsung Galaxy S8	Apple iPhone X	Google Pixel 2
Total	1238	885	479
Positive	568	364	206
Negative	45	96	52
Neutral	625	423	221

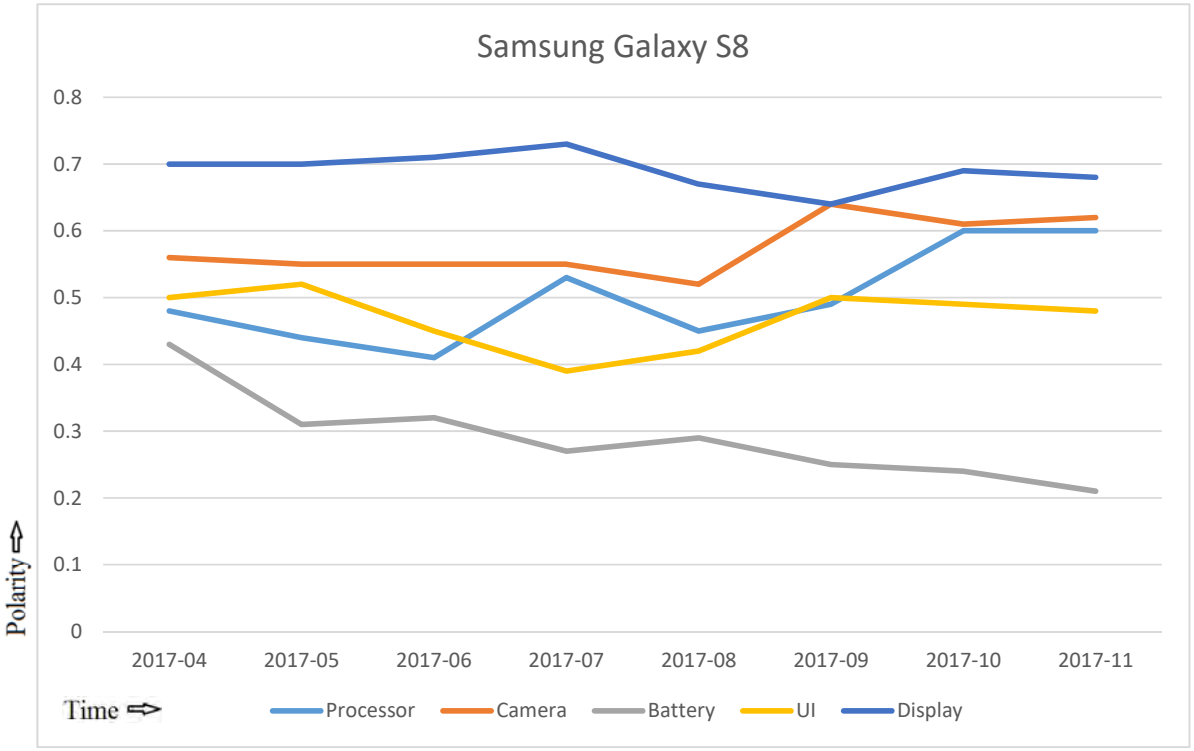


Figure 15: Time vs. Polarity graph of Samsung Galaxy S8

The above graph is a time vs. polarity graph of Samsung Galaxy S8. This phone was released at 21st April, 2017.In this graph, x-axis is the time and y-axis is the polarity. As this model was released before several months, we averaged the polarities of data from one month and input the results with respect to the month. If we did not plot the graph considering time in days, the graph will be scattered and very tough to understand. The dates are from 21st April to 30th November. This graph shows the users preference shifting with respect to time in a complex method. We did not include the attribute price on this graph as the cost of cellphones do not change much resulting in the sentiment of people being constant.

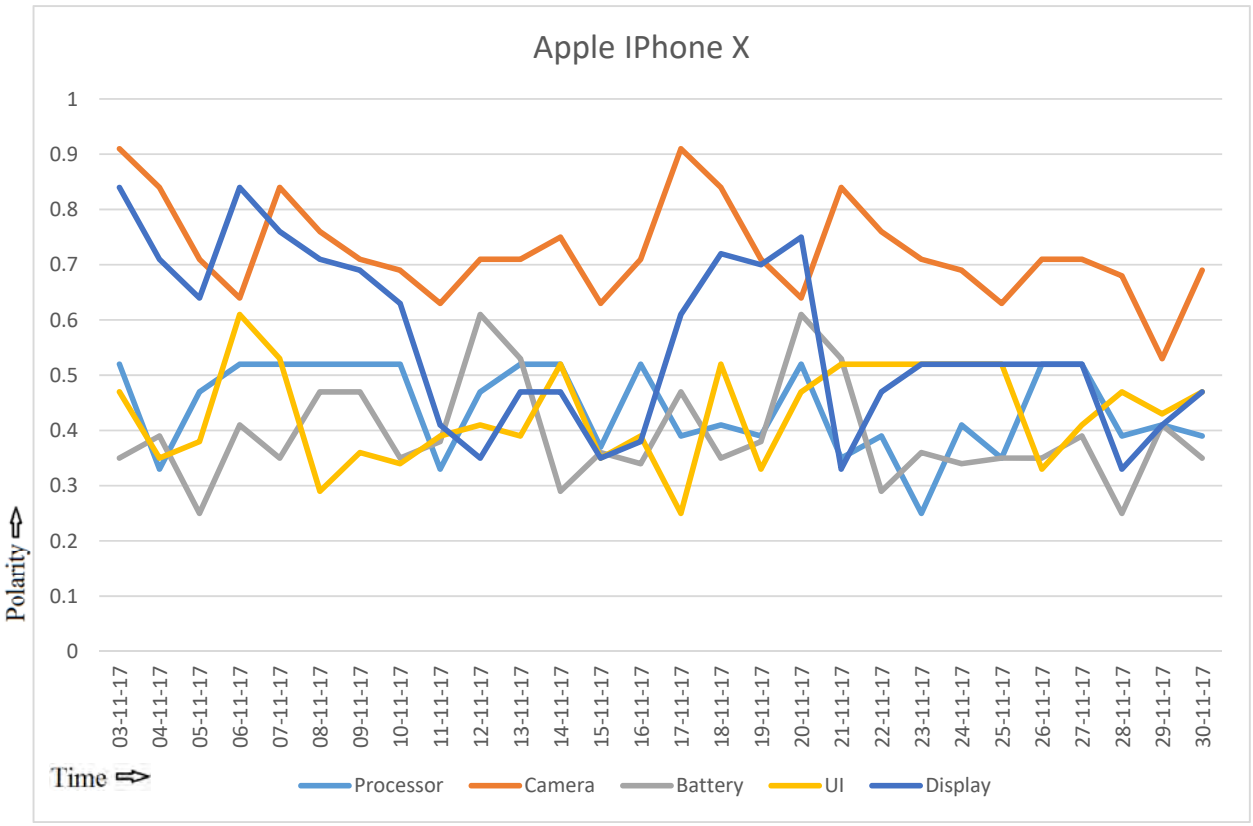


Figure 16: Time vs. Polarity graph of Apple IPhone X

The above graph is a time vs. polarity graph of Apple IPhone X. This phone was released at 3rd November, 2017. In this graph, x-axis is the time and y-axis is the polarity. As this model was released before few months, we averaged the polarities of data of one day and input the results with respect to the month. This phone was released recently, so we plotted time in days. The dates are from 3rd November to 30th November. This graph shows the users preference shifting with respect to time in a complex method. We did not include the attribute price on this graph as the cost of cellphones do not change much resulting in the sentiment of people being constant.

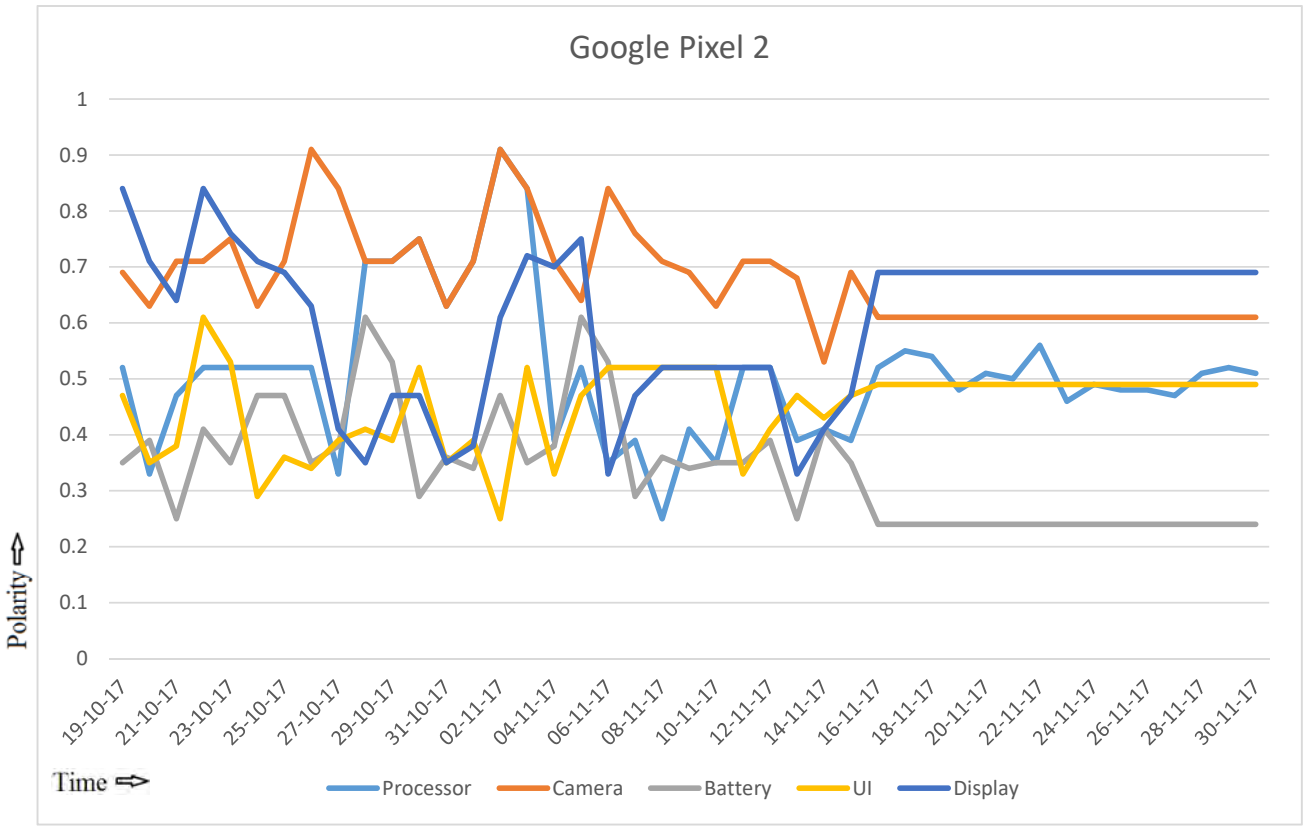


Figure 17: Time vs. Polarity graph of Google Pixel 2

The above graph is a time vs. polarity graph of Google Pixel 2. In this graph, x-axis is the time and y-axis is the polarity. This phone was released at 19th October, 2017. As this model was released before few months, we averaged the polarities of data of one day and input the results with respect to the month. Due to this phone being recently released, time was plotted in graph in days. The dates are from 19th October to 30th November. This graph shows the users preference shifting with respect to time in a complex method. And the frequency is high in this graphs it is daily average of the polarity We did not include the attribute price on this graph as the cost of cellphones do not change much resulting in the sentiment of people being constant.

CHAPTER 05

Conclusion & Future Work

5.1 Conclusion

In this paper, we have proposed a general product rating system based on public opinion. This can be widely used in the future to get proper reviews of any product to get the best review possible for a product. This system will allow user to briefly analyze bunch of products based on their rating and choose the best one depending on the budget and requirement of the user. This system is reliable as the rating is basically generated based on public opinion and this rating will reflect the actual state of the product in the market of buyers with limited information on the product.

5.2 Future Work

We have built this system to work on cell phone's rating. Our basic plan was to implement the system for any product. In future we wish to work to integrate the system so that it can generate rating for any product. That would not require much effort as we can always collect data from twitter on any product and we can train more datasets in our created classifier. We wish to make this system user friendly more.

We look forward to ease the process of decision making about buying any product for people. This system would be just one more step towards digitalization. We belief it will make people's lives easier as this system will allow users to see an accurate rating of any product and save them from the misery of finding a good review from various review sites.

Reference

[1] Vishal A, S.S. Sonawane, "Sentiment Analysis of Twitter Data:A Survey of Techniques" 2016 volume 139 No.11

[2] Prerna Chikersal, Soujanya Poria, and Erik Cambria "SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning"

[3] M. Vadivukarassi, N. Puviarasan and P. Aruna "Sentimental Analysis of Tweets Using Naive Bayes Algorithm". World Applied Sciences Journal 35 (1): 54-59, 2017 ISSN 1818-4952

[4] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan "Mining Social Media Data for Understanding Students' Experiences", IEEE transactions on learning technologies, vol. 7, no. 3, july-september 2014

[5] Luciano Barbosa Junlan Feng "Robust Sentiment Detection on Twitter from Biased and Noisy Data" Coling 2010: Poster Volume pages 36–44, Beijing, August 2010

[6] Vimalkumar B. VaghelaBhumika M. Jadav Analysis of Various Sentiment Classification Techniques

[7] Bingwei Liu ; Erik Blasch ; Yu Chen ; Dan Shen ; Genshe Chen "Scalable sentiment classification for Big Data analysis using Naïve Bayes".

[8] Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), pp.1-167.

[9] A. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," Proc. Human Language Technology Conf. and Conf. Empirical Methods in Natural Language Processing, pp. 339- 346, 2005.

[10] Chaovalit, P. and Zhou, L., 2005, January. Movie review mining: A comparison between supervised and unsupervised classification approaches. In System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on (pp. 112c-112c). IEEE.

[11] Gobinda G. Chowdhury (2006). Natural language processing.

[12] Bing Liu list of words, http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

[13] Bill McDonald list of words, http://www3.nd.edu/~mcdonald/Word_Lists.html,

[14] Montejo-Ráez, A., Martínez-Cámara, E.,Martín-Valdivia, M., Ureña-López, L.: Random walk weighting over sentiwordnet for sentiment polarity detection on Twitter.

WASSA '12 Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis

[15] Promtcloud, Amazone unlocked phone review, https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones

[16] R. Creţulescu ; A. David ; D. Morariu ; L. Vinţan "Part of speech tagging with Naïve Bayes methods" System Theory, Control and Computing (ICSTCC), 2014 18th International Conference 15 December 2014

[17] Abinash Tripathy Ankit Agrawal, Santanu Kumar Rath Classification of Sentimental Reviews Using Machine Learning Techniques

[18] Tripathy, A., Agrawal, A. and Rath, S.K., 2015. Classification of Sentimental Reviews Using Machine Learning Techniques.Procedia Computer Science,57, pp.821-829

[19] Chaovalit, P. and Zhou, L., 2005, January. Movie review mining: A comparison between supervised and unsupervised classification approaches. In System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on (pp. 112c-112c). IEEE.

[20] Pang, B., Lee, L. and Vaithyanathan, S., 2002, July. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10(pp. 79-86). Association for Computational Linguistics.

[21] Y. Jo and A.H. Oh, "Aspect and Sentiment Unification Model for Online Review Analysis," Proc. Fourth ACM Int'l Conf. Web Search and Data Mining, pp. 815-824, 2011.

[22] Hu Minqing, Bing Liu, "Mining opinion features in customer reviews", Proceedings of the national conference on artificial intelligence, vol. 4, no. 4, pp. 755-760, 2004.

[23] Mouthami, K., Devi, K.N. and Bhaskaran, V.M., 2013, February. Sentiment analysis and classification based on textual reviews. In Information Communication and Embedded Systems (ICICES), 2013 International Conference on (pp. 271-276). IEEE.

[24] Gautam, G. and Yadav, D., 2014, August. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In Contemporary Computing (IC3), 2014 Seventh International Conference on (pp. 437-442).IEEE.