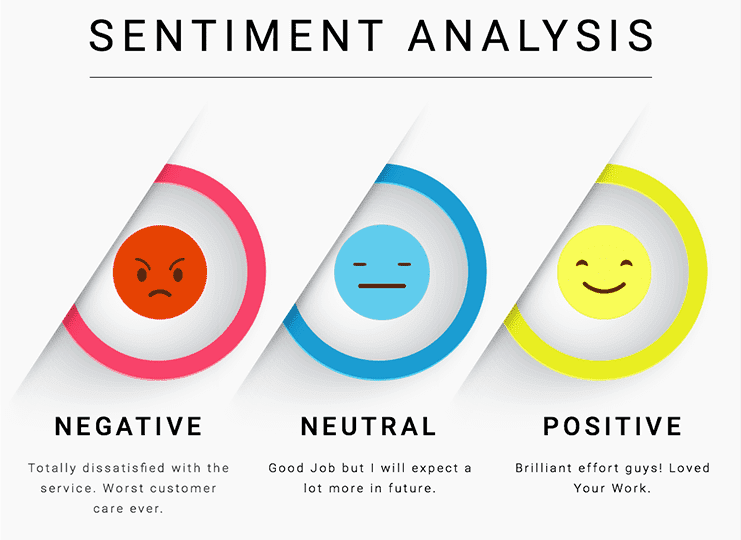
Devi Ahilya Vishwavidyalaya, Indore.

School Of Data Science and Forecasting



ASSIGNMENT ON

ANALYSIS OF SOCIAL MEDIA POSTS USING DATA MINING MODEL



GUIDED BY: SUBMITTED BY:

PROF. KAPIL SHARMA AYUSHI JAIN (DS7B-2006) GAUTAM (DS7B-2008) NISHITA JAIN (DS7B-2010)

PARITOSH LAHRI (DS7B-2011)

**M.Tech. (Big Data Analytics)**

# Index

|  |  |
| --- | --- |
| 1. | Abstract |
| 2. | Data mining |
| 3. | Real time applications |
| 4. | Sentiments analysis in everyday life |
| 5. | Literature review |
| 6. | Problem statement |
| 7. | Objective |
| 8. | Algorithm 1 |
| 9. | Python code 1 |
| 10. | Algorithm 2 |
| 11. | Python code 2 |
| 12. | Sentiments analysis in Excel |
| 13. | Future research work |
| 14. | Conclusion |
| 15. | References |

# Abstract

Knowledge plays a vital role in every human life. Data mining supports to Knowledge discovery, pattern discovery/correlation among data. And these are information applied in various applications like business, media, medical, agriculture etc. Data mining field has been very much successful from its early design to present date. Though it has many drawback while handling social media. Social media is very important source that provide huge amount of data that are in unstructured and heterogeneous form.

Social media is process of gathering and analyzing data from social Networks such as Facebook, Twitter, Instagram etc. Social media has got remarkable attention in last decade. And due to multi-dimensionality and heterogeneousity property of data, handling of this data become the major issues. At present day, a number of algorithms are present with their merits and demerits.

Social media mining requires human data analysts and automated software programs to shift through massive amounts of raw social media data in order to recognize patterns and trends relating to social media usage, online behaviors, sentiment behavior, sharing of content, and connections between individuals, online buying behavior, and more. So this project provides us basic idea about the sentiment analysis with some data mining techniques, on social network platform.

**Data Mining**

Data Mining is referred to as Information Harvesting/ Knowledge Mining/ Knowledge Discovery in Databases/Data-Dredging/Data-Pattern Processing/Data Archaeology/Database Mining, Knowledge Extraction etc. and many more. Data Mining is a process of analyzing data from different dimensions or angles and summarizing it into useful information that can be applied in different fields to take proper decision. It increases profits and cuts costs, or both. Technically, data mining is a computing process of discovering patterns or correlations in large relational databases involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems.

***Evolution:***

In 1763 Bayes’ theorem was published, including fundamental of data mining and conditional probability. It allows understanding of complex realities based on estimated probability. Regression analysis was later introduced in year 1805, was used to find the relationship among variable. Regression is one of major tool of data mining technique. In 1936, When computer was started where collection and processing of large amount of data was possible. The basic of object oriented concepts started in mid 1960s; Then Programmer started giving importance to data storage and processing.

Data plays an important role in all the fields and started to attract the attention of researchers. Due to fast growth of technology, huge volumes of data are available. As a result researchers put their interest in managing and analyzing the data.

In mid 1970s, researcher make possible to store and query tera bytes and penta bytes of data are being stored in much complex databases systems. The term “Knowledge discovery in databases” (KDD) was found in 1989. In 1990s, the “data mining” concept appeared in the database. Retail companies and the financial communities apply data mining to analyze data and recognize trends to identify customers’ behavior, predict stock prices and increase their customer base.

As size and complexity of data set increases rapidly, ordinary tools for data analysis is not enough. Hence data analysis has been enhanced with indirect, automated data processing, such as neural networks, cluster analysis, genetic algorithms(1950s),decision trees and decision rules(1960s), and support vector machines(1990s).Though there are more number of data mining algorithms under various techniques, still research is going on in this area.

## Processes:

The processes involved in the knowledge discovery are: (i) Data cleaning (to remove noise or irrelevant data), (ii) Data integration (where multiple data sources may be integrated / combined), (iii) Data selection (where data relevant to user’s request / analysis task are retrieved from the database), (iv) Data transformation (where data are transformed or consolidated into suitable forms for mining, by performing summary or aggregation operations, for instance), (v) Pattern discovery (Patterns are discovered by applying intelligent methods), (vi) Pattern evaluation (identify the truly interesting patterns representing knowledge based on some interestingness measures), (vii) Knowledge presentation (visualize the patterns in different forms).

*Architecture:*

The architecture of a data mining system has the following major components as depicted in figure Fig: 1

Pattern Evaluation

Knowledge base

Data Mining Engine

Graphics User Interface

Database or Data warehouse Server

Data Base

***Figure 1 Architecture of Data Mining System***

Data warehouse

## Database And Data warehouse:

The Database / Data warehouse server fetch the relevant data, Based on the users request through the data mining process called selection and transformation.

## Knowledge Base:

This is the domain knowledge that includes concept hierarchies, user beliefs, thresholds, and metadata(data about data) that is used to guide the search, or evaluate the interestingness of resulting patterns.

## Pattern Evaluation Module:

This component is integrated with the mining module and interestingness pattern stored in the knowledge base so as to focus the search towards interesting patterns. Pattern evaluation module plays an important role in efficient data mining to confine the search to only the interesting patterns.

## Graphical User Interface:

User can communicate with data mining system through this module and interact with system by providing information through query or task to help focus the search. In addition this component help user to browse database, Visualize the patterns in different forms and evaluate mined patterns.

Based on the type of databases used , kind of information to be extracted and type of technique and tools used, Data mining system are categorized are listed below.

*Type of data sources mined*

In any organization, huge amount of data are available in different data repositories at different Geo-location and in various formats (audio, video, text etc) and therefore data are classified according to their types.

*Data model*

At present time different data mining model are available such as Relational data model, hierarchical model etc. The data mining system classifies the data, from the model.

*Kind of knowledge discovered*

This classification based on data mining functionalities, such as association, clustering, classification, characterization, discrimination, etc. Some systems provide hybrid data mining functionalities together.

# Real Time Application

As the demand for data analyses is increasing day by day, It has touched ever sector of the industries at every level. These sectors include *sales / marketing, banking, agriculture, insurance, Transportation, Telecommunication, fraud detection, finance, education sector, medical and so on*.

Some important sectors are noted.

➔ Data Mining in Education Sector:-

Educational Data Mining (EDM) is an emerging interdisciplinary research field. Educational Data Mining refers to techniques and tools applied on information generated from educational settings related to students’ learning activities and investigate scientific questions within educational research. It is very much useful to understand students’ learning behaviors and the settings which they learn in. Use student’s data to analyze their learning behavior to predict the results.

➔ Data mining in Medicine

An Enormous Electronic Health Records (EHRs) are available in medical field. Accuracy is considered an extremely important factor while handling these EHRs as it is related with patient’s health. DM can generate information that can be useful to all stakeholders in health care, including patients by identifying effective treatments and best practices.

➔ Data Mining in Agriculture

Agriculture is base for human food, and sustain life. Recent study has addressed certain problem in agriculture. And by applying Data mining in agriculture which can solve the problem of seeds, or water and by tackling these problem we can yield of the crops.

➔ Data mining in Marketing

Due to rapid fluctuation in the value of currency and dynamic customers’ behavior, it is very difficult to take investment decision in business. Also stock market is also being generating huge volume of data. This nature has attracted researchers to apply mining on these data and find patterns to predict the probability of purchasing the product and the future trend of business.

➔ Data mining in Social Media

Social media mining analyzes and extracts patterns or correlations or trends from raw social media data e.g., social media usage, online behaviors, sharing of content, connections between individuals, online buying behavior, etc. These patterns give valid information to companies, governments and non -profit organizations, to design their strategies or introduce new programs / products / services.

### Sentiment Analysis in Everyday Life

As various technologies are coming into life, so is the analysis of sentiments of a human being are playing an important role in order of efficient functioning of the businesses. Social media monitoring tools use sentiment analysis to extract useful insights from the data which is being generated exponentially every day for the feedback of a business.

However, sometimes sentiment analysis has been used in more unclear circumstances. Facebook, for example, came into light when it was observed that they were using sentiment analysis to check if they can [influence user emotions](https://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds) by changing their algorithms to show negative or positive posts more often into their users’ news feeds. By using this process of “emotional contagion,” they found that they could easily influence their users’ emotions by showing more relative posts which they have been following lately.

The issue is that Facebook never informed its users that they were under some experiment and may have caused emotional distress to them in some cases. Clearly we can see how this use of sentiment analysis can be unethical.

But our main interest lies in how sentiment analysis is used with social media monitoring tools, Social media sentiment analysis determines whether a user is using your product, service, or brand in a positive, negative, or neutral way.

Why is social media sentiment analysis important?

1. Improved customer satisfaction.
2. Understand your audience.
3. Prevent social media crisis.
4. Measure the results of a Public Relation campaign.
5. Monitor your competitors.
6. Improve your service according to your customers’ needs.



**LITERATURE REVIEW**

In [2], the authors have studied the techniques that are currently used to analyze Social Media(SM).In this paper, analysis of social media data has proved to be effective, because of the capacity possessed by data mining in handling unstructured data. Author suggested more data mining techniques in future are yet to be explored.

In [3], the authors have studied the results of the survey papers and have provided some useful ideas for controlling the attributes which mostly affect social media. The Data mining techniques provide a better data control facility. The data mining techniques support for discovering the similarities among the patterns which exist among the unstructured data set. From the outcomes and the results produced, there searchers make a new dimension for the researcher to control the uncontrollable data existing in the social Medias and social networks.

In the author used systematical data mining architecture to extract useful knowledge from social media. In this research they have used Facebook as a primary data source and collected different attribute such as age gender, wall counts, number of music, there are number of activities and interests are present. After analyzing the mined data they suggested that social data mining is an challenging to mine knowledge which can be used in human behavior prediction, decision making, pattern recognition, social mapping and many more.

In, according to technocratic, about 75,000 new blogs and 1.2 million new posts giving opinion on products and services are generated every day.

In Data mining algorithms have been used to analyze opinion/sentiments expressed on different Social Media sites. Research revealed that Support Vector Machine, Naive Bayes and Maximum Entropy were majorly considered among other data mining techniques available when analyzing opinion/sentiment in Social Media.

### Problem Statement

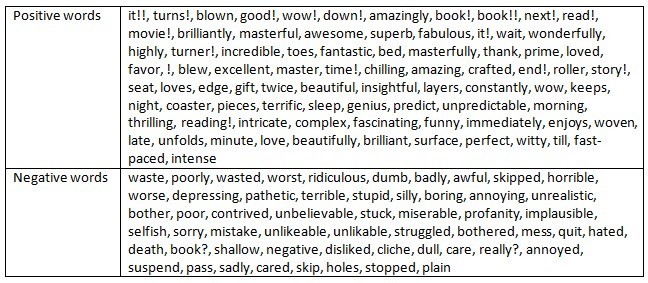
Social Media carries tones of post every single minutes, And due to which number of comments increases, we as a team want to analyze every comment that had be written on the post and try to find the *biasness of the sentence* and Persons attitude toward the post or the content.

### Objective

As our Problem is simple, we took a set of keywords through which we will analyze the data, And keyword is divided into three categories:-

* 1. Positive words
  2. Negative words
  3. Neutral words

After finding type of words used in any sentence, we can calculate skewness of the sentence that where it lies, in positive side or on the negative or neutral.



**Algorithm**

In our approach, we had used two different feasible solution applied on the different datasets. Data mining is collecting relevant data from unstructured data. His purpose of data mining effort is normally either to create a descriptive model or a predictive model.

Algorithm 1:-

In this approach our aim is to check that how many comments in the given list of comments are positive, negative or neutral.

1. For that, We have made a program using class as analysis, and defining its methods
2. Under the class we have assigned class variables including positive, negative, neutral keywords as pchar, nchar, ichar
3. We define a constructor init method to initialize values through arguments passed in given list.
4. For this we will run comment wise for loop in given list li. Using split() we have further split the string in the list into substrings ,placed in res variable.

5 We will run second loop for each element (i) in list to identify that the substring exists in the given list of keywords and identify the category in which it falls out of the three .

6. If the substring is found to be present in pchar, then p stack will get appended under the method reaction or else is tack or nstack will get appended

1. In \_count method we have counted the number of keywords that are appended in pstack, nstack, istack through len(pstack) ,len(istack) and len(nstack)
2. In cal\_percent method we have calculated the percent of positive, negative, neutral comments among total number of comments.
3. And finally in the class \_plot we have used matplotlib to plot bar chart for count of category wise comments and pie chart to represent percentage of category wise sentiments.

**Python code 1:-**

**#**importing matplotlib and numpy

import matplotlib.pyplot as plt import numpy

#%matplotlib inline

#class definition analysis

class analysis():

pchar=["Hurray","Yeah","Wow","amazing","delighted","joy"]

ichar=["calm","cool","relax","chill","soothing"]

nchar=["OhNo","sad","gloomy","angry","lost","bad"] p\_stack = []

i\_stack = [] n\_stack = []

def init (self,li):

self.li=li

def reaction(self):

for i in self.li:

res=i.split() for j in res:

print(j)

if j in self.pchar:

self.p\_stack.append(j) elif j in self.ichar:

self.i\_stack.append(j) elif j in self.nchar:

self.n\_stack.append(j)

#precentage calculation or aggrate def cal\_percent(self):

self.pnum=len(self.p\_stack) self.inum=len(self.i\_stack) self.nnum=len(self.n\_stack) self.per\_p=len(self.p\_stack)\*100/len(self.li) self.per\_i=len(self.i\_stack)\*100/len(self.li) self.per\_n=len(self.n\_stack)\*100/len(self.li) print("% of positivity is:",self.per\_p) print("% of neutrality is:",self.per\_i) print("% of negativity is:",self.per\_n)

def \_count(self):

print("total number of positive characters:",self.pnum)

print("total number of neutral characters:",self.inum) print("total number of negative characters:",self.nnum)

def \_plot(self):

self.x=numpy.array(['positive count','neutral count','negative count'])

self.y=numpy.array([self.pnum,self.inum,self.nnum])

#ploting the graph plt.bar(self.x,self.y,label='reaction\_count',color='red',width=0.2)

plt.legend() plt.show()

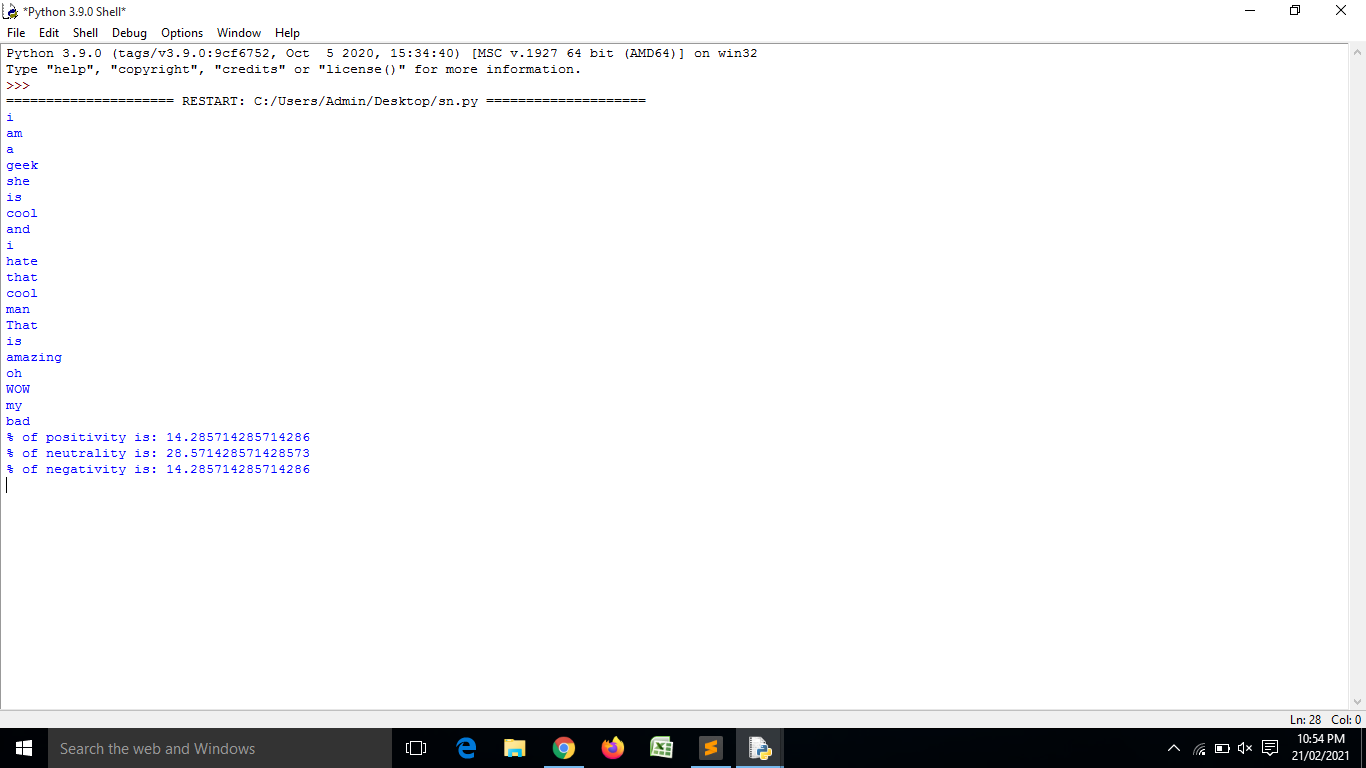
portion=[self.pnum,self.inum,self.nnum] reactions=['positive','neutral','negative'] plt.subplots() plt.pie(portion,labels=reactions) plt.legend(loc="upper right")

plt.show() #manually passed the string as below

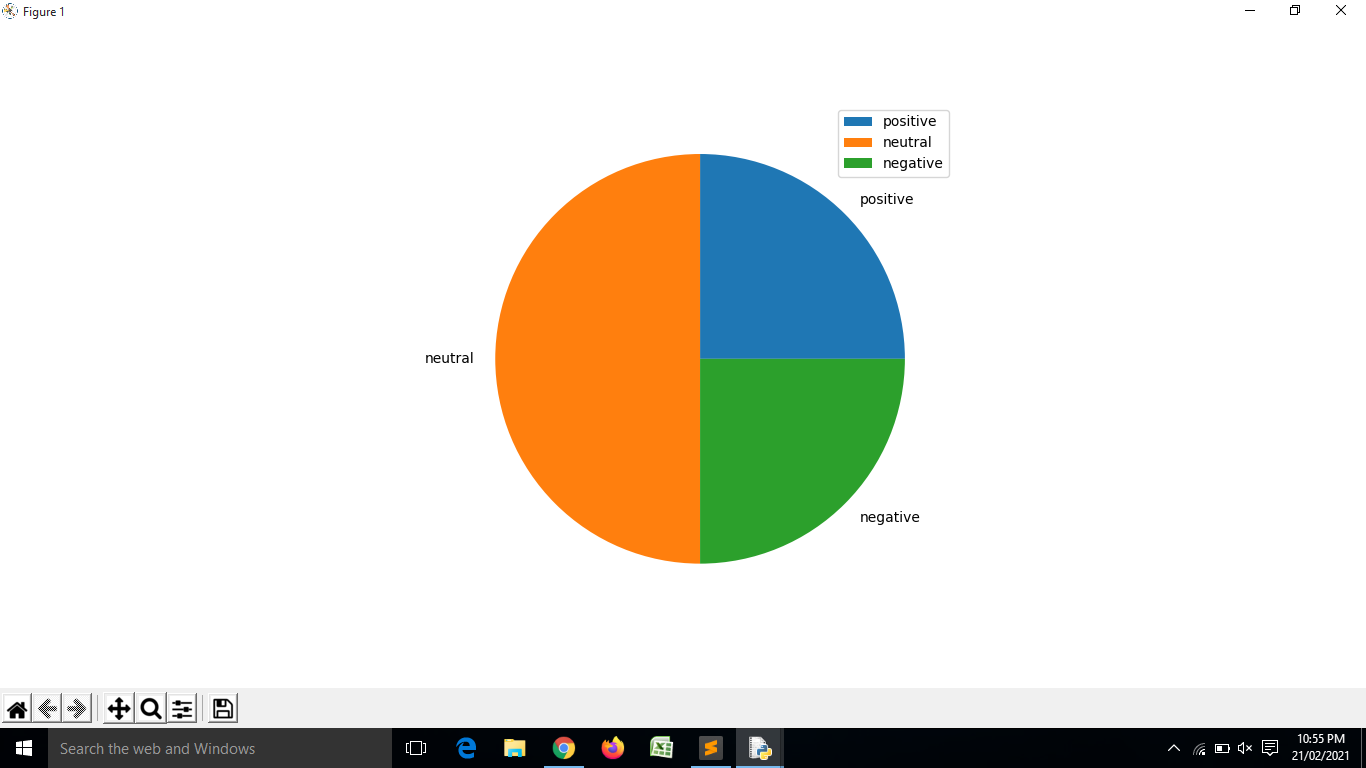
l=analysis(["i am a geek","she is cool","and i hate that","cool man","That is amazing","oh WOW","my bad"])

l.reaction() l.cal\_percent() l.\_plot()

After executing this program and produced an output:-







**Algorithm for 2nd program-**

Firstly we will create a variable named as ‘chat’.

Then we will take an input from user as a string which will be his comment on any particular post.

We will assign that input string to our previously defined variable ‘chat’.

We will convert the whole string into lower case and then split every word in the string separately.

Then we will create three lists named as Positive[], Negative[], and Neutral[].

We store some used positive keyword in our Positive[] list.

Then we will store some frequently used negative keywords in our Negative[] .

Also do the same with the Neutral[] list and store some frequently neutral used keywords in it.

Then we will create a for loop in which with the help of a variable we will traverse that particular variable in our positive list.

Same we will do with the other two lists also, we will iterate a for loop in other two lists also.

Then we will apply a condition in all the loops separately to find the keyword, and check what keywords are present in that string which will we take as input from the user his comment.

Then after we will check that what types of keywords present in that string which was given as an input from the user whether they are positive, negative or neutral.

We will do this to all the three different for loop we created to check whether the given comment is positive, negative or neutral.

Here we have also used Matplotlib for the visualization of the data.

With the help of matplotlib we have represented the percentages of the different sentiments in the pie chart as well as bar chart.

And at last we will get the result of the sentiments in that particular comment.

Here we have also used a counter with every loop to check that what percent of sentiments of each type present in that comment.

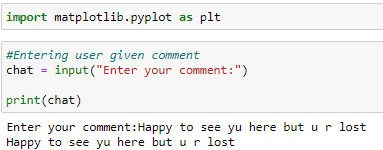
At last we will get our desired result and in it we will get to know that which type of sentiments the person is showing in his comment.

We will also get to know the percentage of which sentiment is present in that comment.

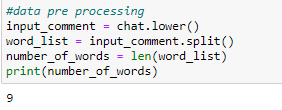
Like this our whole program will run and give us our desired result as an output with simple visualization.

Python code 2

--Initialization of script and scanning the user input--

******

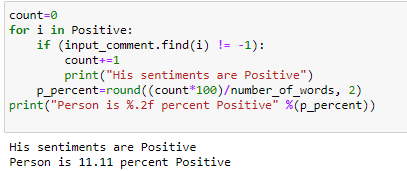
--Preprocessing of the scanned data--



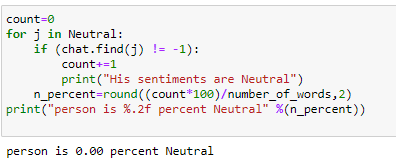
--List creation of keywords of different sentiments--



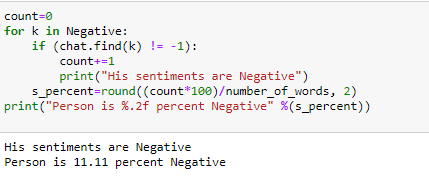
--Creating loop to check positivity in the data—



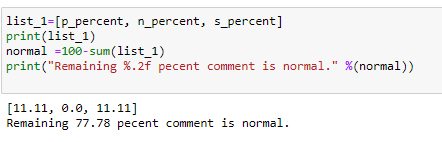
--Creating loop to check neutrality in the data—



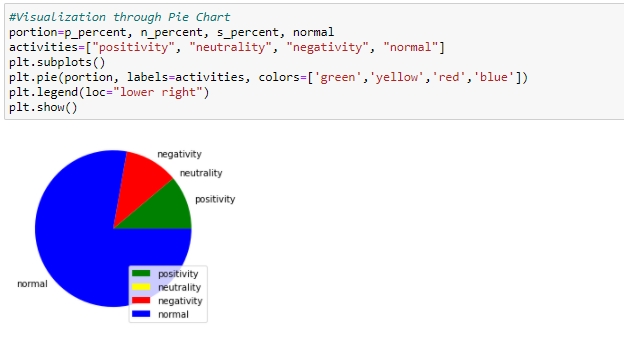
--Creating loop to check the negativity in the data --



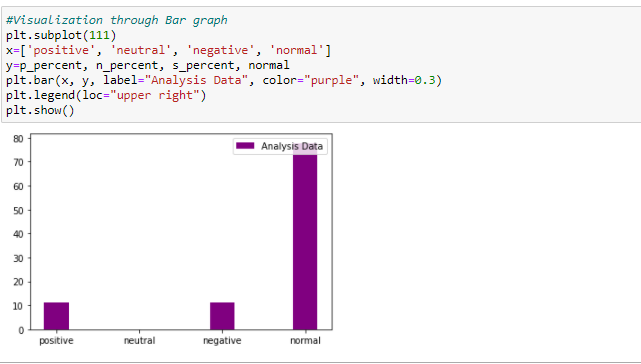
--Checking remaining data --



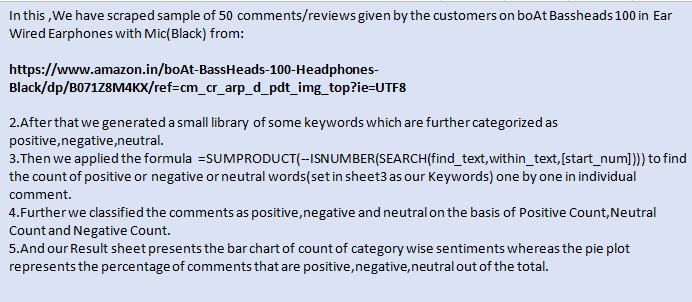
--Visualization of our Result through Pie Chart—



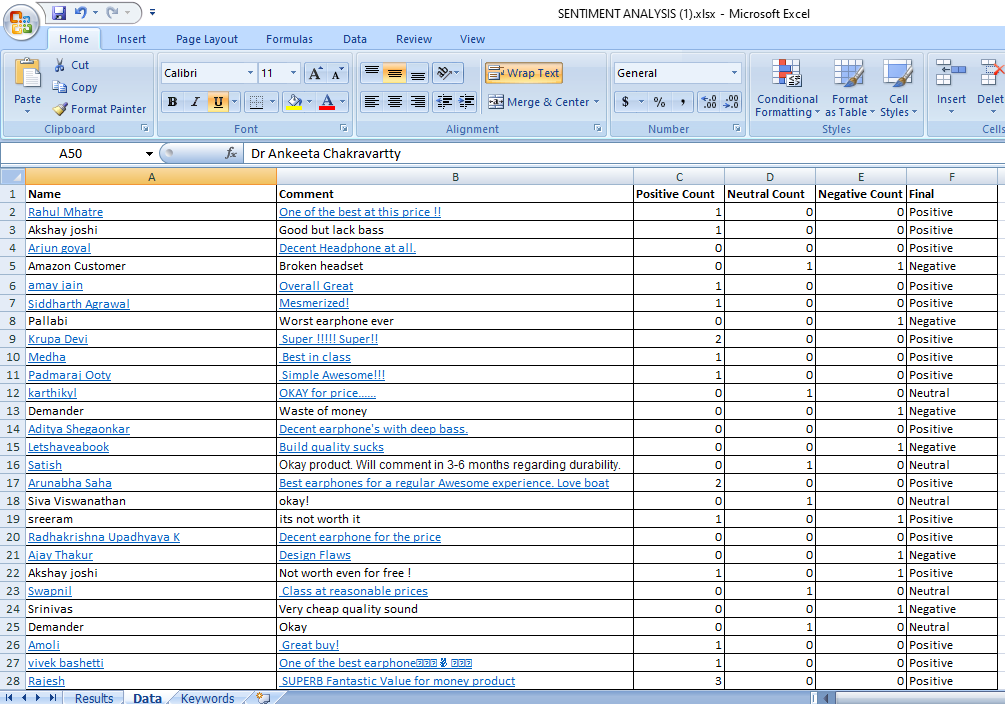
--Visualization of our Result through Bar Graph --

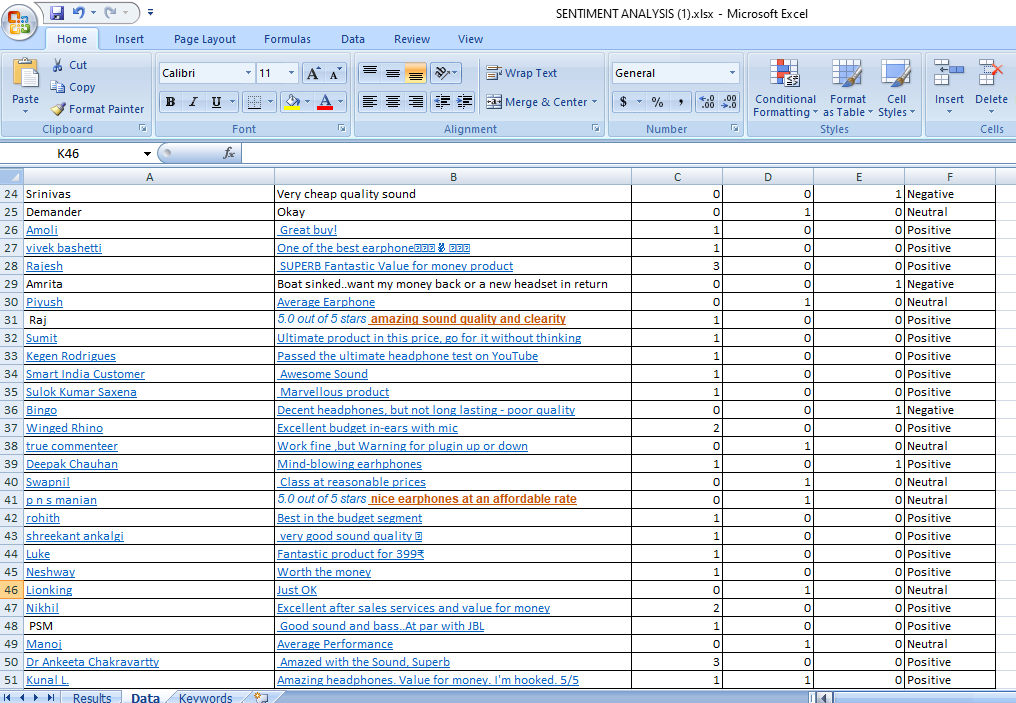


**Sentiments analysis in excel**

****

Data used --

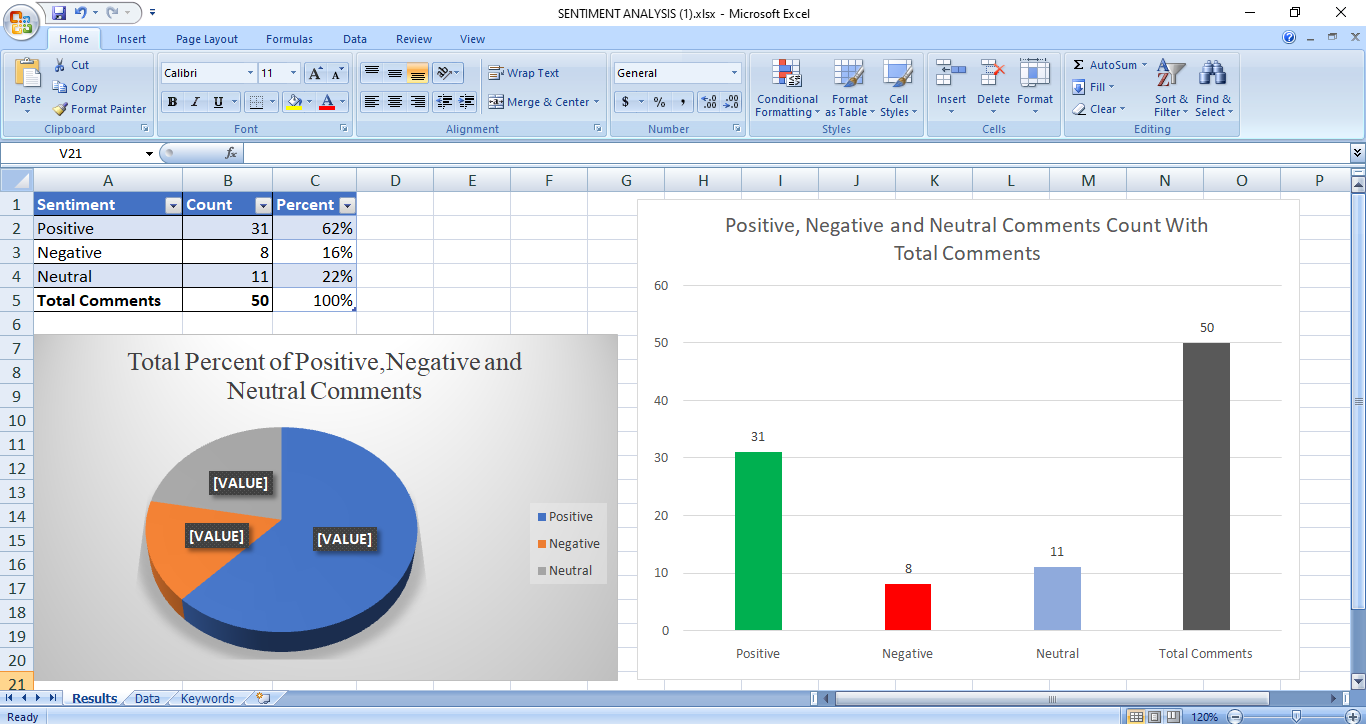
****

****

Keywords taken—



Result and visualization—

****

Link for above excel program-

[**..\Downloads\SENTIMENT ANALYSIS (1).xlsx**](../Downloads/SENTIMENT%20ANALYSIS%20(1).xlsx)

**FUTURE RESEARCH WORK**

In this era of infomation, getting a more simple and more accurate model for mining is tough job, because getting a effective and efficient algorithm involves a number of job aspect in the field of development. Minimizing Time complexity and Space complexity is anothor aspect, but in the field of unstructured dimensionality and multi-variant data make the job more complex.

In this paper we used Data mining framework to provide a basic glimpse How sentiment analysis is performed. We have basic theme of the analysis. We performed this analysis using two different platform, one from Ms-Excel and another Python. Our model to approach the problem where same in both the platforms. And the data used was filtered. And of one dimensioned, If we increase the volume of the data then this model might not work. But future study on this model would make it more efficient and effective that it can perfom much larger scaled data mining operation.

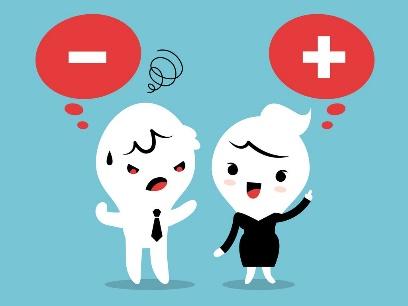
Different data mining of algorithm are present for different kind of algorithm. The broadness of our model is less meaning, the keywords that we used where less in number, indentifying more keywords means increasing the complexity. So there are many more aspect that we can keep in mind in order to increase the performance of the model in Natural language processing.

**CONCLUSION:**

Ultimately, sentiment analysis is not always accurate, but neither are we when trying to decode what someone means by their expressions or by any other communication means. Within social media monitoring, we need sentiment analysis as a starting point to understand general public sentiment in aggregate. From there, we can use the public’s general emotions to initiate campaigns based on their feedback.

Social media is perhaps the largest pool from which we can mine insights for public opinion and begin to gather informative data on the success or failure of our brand, products, or marketing campaigns in the eyes of the public.

Here, a simple yet powerful model using python has been built which precisely will analyse the mood of the person to a particular post he/she reads on the social media, the model will say about the person having positive, negative or neutral thoughts about that post and also, the results will be visualized through a bar and pie chart. Also, another model on Microsoft Excel has been developed where the comments of the users have been scraped and analysed their reviews on the particular product. The results are represented in the form of a dashboard which will show the output of the analysis. Initially, dataset of 50 comments is taken and analysed which shows accurate results, this can be extended to any number of comments in the datasets.



In sum, sentiment analysis helps in the interpretation of human behavior towards any situation. This typically involves taking a piece of text, whether it’s a sentence, a comment or an entire document and returning a “score” those measures how positive, negative or neutral the text is.

**REFERENCES**

1. Mohammad Noor Injadat, Fadi Salo, Ali Bou Nassif “Data Mining Techniques in Social Media: A Survey”, NEUCOM17295, Volume 214,PP:654-670, 2016
2. Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F, “A Survey of Data Mining Techniques for Social Media Analysis” in Journal of Data Mining & Digital Humanities, PP:1-27, 2014.
3. Dr.B.Umadevi1, P.Surya2, “A Review on Various Data Mining Techniques in Social Media”, in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5,

Issue 4, PP: 8082-8086, 2017.

1. Muhammad Mahbubur Rah-man, “Mining Social Data to Extract Intellectual Knowledge", International Journal of Intelligent Systems and Applications(IJISA), vol.4, no.10, pp .15-24,

2012. DOI: 10.5815/ijisa.2012.10.02

1. Kim, P.: “The Forrester Wave: Brand Monitoring, Q3 2006,” Forrester Wave (white paper), 2006.
2. Clarke, D., Lane, P., Hender, P.: Developing Robust Models for Favourability Analysis. UH Research archive, 2011.
3. Pang, B. and L. Lee, Vaithyanathan, S.: Sentiment Classification Using Machine Learning Techniques. In: Proceedings of Conference on Empirical methods in natural

Language Processing (EMNLP), Philadelphia, July 2002, 79 - 86. Association for Computational Linguistics, 2002.

1. Hemlata Sahu, “ A Brief Overview on Data Mining Survey” in International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3, PP: 114-121.
2. Dinesh Bhardwaj1, Sunil Mahajan2, “ANALYSIS OF DATA MINING TRENDS, APPLICATIONS, BENEFITS AND ISSUES”, in International Journal of Computer Science and Communication

Engineering, Volume 5 issue 1, PP:53-57, 2016.

1. Dr. Poonam Chaudhary , “Data Mining System, Functionalities and Applications:A Radical Review” in International Journal of Innovations in Engineering and Technology (IJIET), Volume 5,Issue

2 ,PP:449-455, 2015.

1. Neelamadhab Padhy1, Dr. Pragnyaban Mishra 2, and Rasmita Panigrahi3, “The Survey of Data Mining Applications And feature scope”, in International Journal of Computer Science, Engineering and

Information Technology (IJCSEIT), Vol.2, Issue.3, 2012.

1. mita1, Priti Sharma, “ Use of Data Mining in Various Field: A Survey Paper”, in IOSR Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 3, PP 18-21, 2014.
2. Mrs. Bharati M. Ramageri, ”DATA MINING TECHNIQUES AND APPLICATIONS”, in Indian Journal of Computer Science and Engineering, Vol. 1 Issue. 4, PP:301-305.
3. Annan Naidu Paidi “ Data Mining: Future Trends and Applications” in International Journal of Modern Engineering Research (IJMER), Vol.2, Issue.6, PP:4657-4663, 2012.
4. Umamaheswari. K, S. Niraimathi “A Study on Student Data Analysis Using Data Mining Techniques”, in International Journal of Advanced Research in Computer Science and Software

Engineering, Volume 3, Issue 8, PP:117-120, 2013