Meshank Hiteshbhai Adhia

School of Computer Science and Engineering

Vellore Institute of Technology

Vellore, Tamil Nadu, India

Meshank.adhia2020@vitstudent.ac.in

# ICANPredict

Prekshita Patil

School of Computer Science and

Engineering

Vellore Institute of Technology

Vellore, Tamil Nadu, India

Prekshita.vasudeo2020@vitstudent.ac.in

Rishabh Sharma

School of Computer Science and Engineering

Vellore Institute of Technology

Vellore, Tamil Nadu, India

Rishabh.sharma2020@vitstudent.ac.in

**Abstract**

Prof. Prabu sir

School of Computer Science and Engineering

Vellore Institute of Technology

Vellore, Tamil Nadu, India

sprabhu@vit.ac.in

In the present era everyone is addicted to internet and hence opinion on the internet are considered as a valid statement as it is user opinion that he faced which can never be wrong. There are several of social media websites that provides such type of facility such as Facebook, LinkedIn, Google reviews, Twitter and so on. We will be considering the most official and trusted social networking site or application that is Twitter also it is gaining high popularity in terms of any trend It might be political, events and much more furthermore we get abundant data which is in raw format. Hence, according to today’s world Twitter plays an crucial role in giving a rise to a company. Hence developing a program such like that will help company to know customers or users are thinking about their company. So, reviews will be analyzed and will be represented in a statical format which can be either in the form of bar chart or pie chart.

## 1. Introduction

Basically, every person is on internet nowadays and they all are expressing their opinion with the help of the social media or applications that are easily available on the play-store, app-store and websites. Some of the famous social media websites and applications are LinkedIn, Facebook, Google, Amazon, Flipkart, Snapdeal, Alibaba, Myntra and Twitter. Furthermore, I these companies provide an opportunity to take and advertise their product online and connect with their buyers or users. Hence based on it, people write their own review on one of the applications which provides an open platform to raise and state their views on a particular thing. Consumer are considered equivalent to god and what they say represents the company/product identity. Hence, as the people like to read the reviews that has been stated by the other user and then take a move towards it. So, the reviews can revolve anything and can boom the news to a greater extent. Hence, a company needs to know whether the good is performing well or not in the present market or not so it will be helpful to them to carry on improvement if needed.

Hence, by going each and every review it becomes tedious task so we will be using Twitter for gathering the reviews which will help the company to conclude whether a

product is on the positive side or not. This will be in the form of pictorial representation which will be in bar chart or pie chart. We will be using a prestored data and it will be filtered means we will be removing the extraneous things which are unnecessary like removal of numerical digits, links and user name to keep it anonymous which will train the model and then we will be testing the reviews that were posted by the users on the twitter. And lastly the conclusion will be drawn on the basis of it.

## 2. Related Work

Many papers were there which were analysing twitter tweets for different purposes such as [1] tried to detect the abuse words on the twitter and they were collecting pre stored data to analyse. Furthermore, there were papers who were considering SVM and logistic regression where the accuracy was about 0.5 to 0.6 [2]. Other approaches that were followed were by using the inbuilt library such as NLTK there were predicting the sentences [3]. On the basis of [4] in which they were comparing different algorithm so by using them we found the accuracy of different algorithm.

## 3. Framework

We will be using python as a development/base programming tool. Moreover, we are going to use Pandas, Scikit learn (sklearn), Matplotlib and seaborn. Initially, we are going to read the comma separated value file with the help of Pandas library and then we will be using Scikit Learn to split the data into training and testing and to import the model. Eventually, we will be using Matplotlib and Seaborn to visualize the data. The task that we have done with the frameworks are

**Classification**: This is a classification application which will be having different categories where categories will be static and will be specified preliminary [1] with the help of our dataset. For an instance “This is bad day” is negative and “what a great day” will be positive.

**Training:** Some Columns or features from the data will be given to learn as our model will be supervised learning where we are training the model with adequate amount of data. This will be the words that are separated and will be 0 or 1 in Boolean values.

**Prediction:** The machine will see the features passed to the

model and will decide the output as per the trained model

## 4. Datasets

We have used training.1600000.processed.noemoticon.csv to train our model. Which contains "target", "ids", "date",

"flag", "user", "tweets"

* Target: - Tweet positive or negative
* Ids: - numerical ids of user
* Date: - on which it was posted
* Flag: - Query or not
* User: - username on twitter
* Tweets: - what they tweeted

The shape of the data will be 1600000 X 6 which indicates there will be 1600000 rows and 6 columns which are stated above. The data will be like: -

* Target :- 0
* Ids :- 1467810369
* Date :- Mon Apr 06 22:19:45 PDT 2009
* Flag :- NO\_QUERY
* User :- \_TheSpecialOne\_
* Tweets:-@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it.

This is a row which is represented column wise. But we will be selecting the target variable and the tweets. Target variable will be in 0 and 4 where 0 denotes negative and 4 will be positive.

**Data Pre-processing**

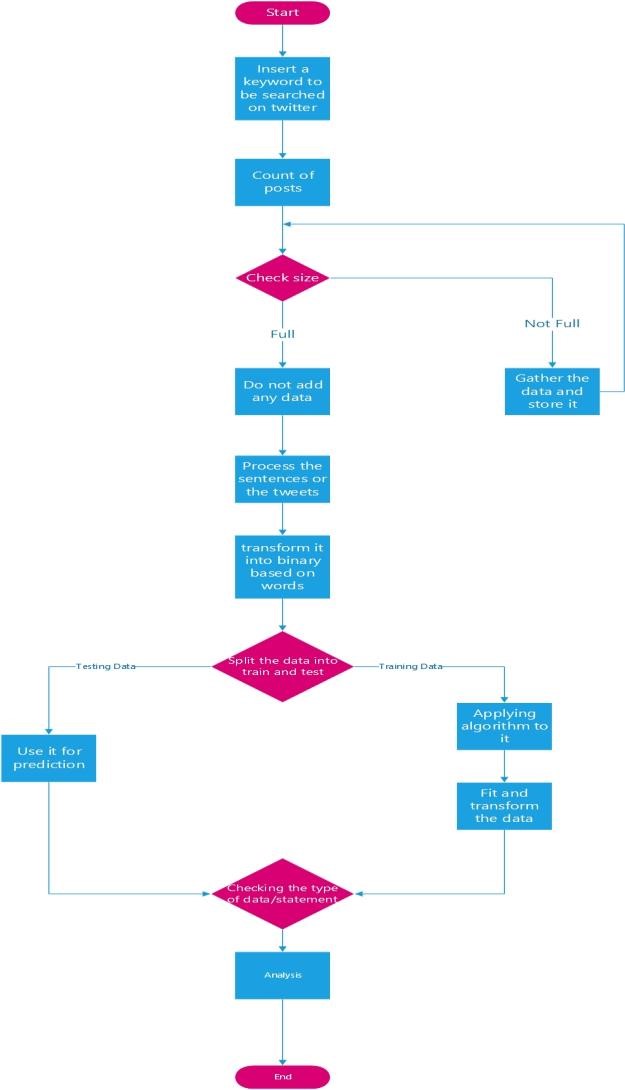
o Removal of Numeric values o Removal of links

o Replacing RT from tweets

o Lowering the tweets words

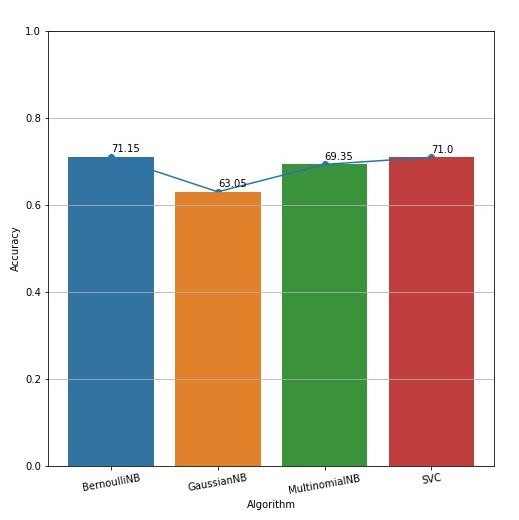
## 5. Architecture

The basic flow that we will follow is by inserting keyword into the python terminal by giving the credentials that are essentials and with the proper authorization later it will ask the number of posts that we want to analyse which will be added into data frame. The training data will be prestored data and the testing data will be the data accumulated from twitter. Hence, at last user will be given an option to choose a visualization option which will be either bar or pie.



### 6. Results

We have trained the data on various models to see which algorithms suits the best so the conclusion that we have is shown in the bar graph



So, from the above-mentioned data we found that Bernoulli Nb was at the peak model among all having 71.15 as accuracy score.

By Using Bernoulli Nb we have tried by different size of data and hence the accuracy that we obtained is shown in the below table

|  |  |
| --- | --- |
| Data Size | Accuracy |
| 1000 | 0.615 |
| 2000 | 0.66 |
| 5000 | 0.711 |
| 10000 | 0.712 |
| 20000 | 0.706 |
| 40000 | 0.725375 |
| 60000 | 0.7364167 |
| 80000 | 0.7323125 |
| 100000 | 0.729 |

0.6

0.65

0.7

0.75

0

20000

40000

60000

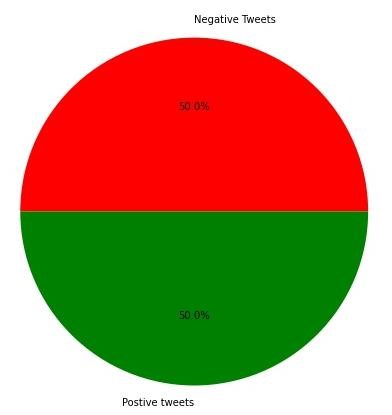
80000

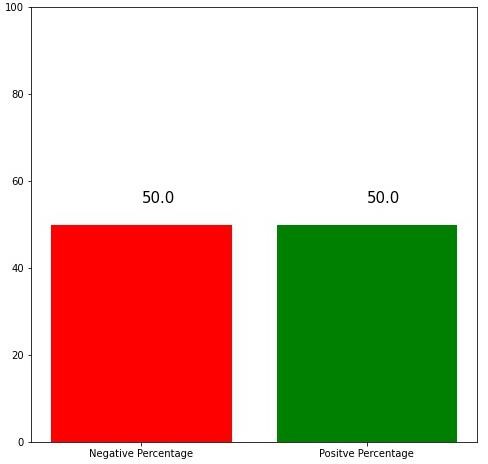
100000

120000

**Accuracy VS Data Size**

The charts show results of predicting live tweets on twitter which are represented in pie chart as well as in bar chart by giving a specific keyword.





### 7. Conclusion

Through this methodology we came on to conclusion that we can help the companies to know their reviews which may be positive or negative related to any product which can help them to modify their way of working and will draw to conclusion using charts.

### 8. Future Work

We will be using emojis and sentence for the predictions and will try to develop either a website or a software. Furthermore, we will try to use data from different social media sites.

## REFERENCES

* Comparative Studies of Detecting Abusive Language on Twitter Younghun Lee Seunghyun Yoon Kyomin Jung Dept. of Electrical and Computer Engineering Seoul National University, Seoul, Korea
* NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis
* Sentiment Analysis: An Overview Comprehensive Exam Paper by Yelena Mejova Computer Science Department, University of Iowa yelena-mejova@uiowa.edu November 16, 2009
* Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy Jin Huang Jingjing Lu Charles X. Ling Department of Computer Science The University of Western Ontario London, Ontario, Canada N6A 5B7 fjhuang, jlu, [clingg@csd.uwo.ca](mailto:clingg@csd.uwo.ca)
* Automatic Sentiment Analysis in On-line Text Eric Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens. Katholieke University Leuven, Tiensestraat 41B-3000 Leuven, Belgium.
* Twitter Sentiment AnalysisAliza Sarlan, Chayanit Nadam, Shuib Basri Computer Information Science Universiti Teknologi PETRONAS Perak, Malaysia [aliza\_sarlan@petronas.com.my,chayanit171@gmail.com,shuib\_basri@petronas.com.my](mailto:aliza_sarlan@petronas.com.my,chayanit171@gmail.com,shuib_basri@petronas.com.my)