CHAPTER -1

PROJECT INTRODUCTION, PURPOSE AND SCOPE

INTRODUCTION

This project is the based on the data science as we know that data science is the combination of various fields like machine learning, Big data, data visualization, data mining, deep learning etc. This project is based on Big data and data mining. We talk about data structure so in this project we used semi structure data so it regard to big data problems also we used large amount of data. And another point in the project is the data mining as mentioned above it contain so many outliers in the dataset so we refine that outlier and select only that dataset field that is useful for us.

In this project we used dummy dataset that take place from Github repository.

For processing and storage we used hadoop ecosystem it totally complete on hadoop framework. We provide storage on hadoop distributed file system and for processing purpose we used Pig Latin languages. this languages implement those developer that is not good in writing a map reduce code. It just like SQL like language so it is very easy to learn.

We store raw data on hadoop distributed file system. After processing or finalise the result we store that data again in 3 distributed file.

The data set that used in this project is tweets based so we process tweets data. In this project we find out the sentiment of tweets that comes in form of tweets.

We check polarity in this project. Polarity is of three type positive polarity mean positive sentiment in form tweets, negative polarity, and neutral polarity or sometimes is called no polarity in term of sentiment means tweets is form of felling no positive view or negative sentiment in tweets. All mining of data is done with help of Pig Latin language that was developed in 2006 when hadoop comes in market.

For check the polarity of tweets we give rating to the word that part of tweets. Every word has unique rating. Some words have a positive rating or more that 0 numerical value and some have negative rating means less than 0. And some words have 0 rating.

After taking useful data with the help of data mining. We select the tweets attribute and serial number and with the help of Pig Latin language we apply matching or joining operation and calculate the polarity of whole line that comes in tweets format. From here we get three type of numerical values first is positive (>0), negative (<0) or equal to zero (=0). What means numerical values we discuss above?

Finally output result stored on hadoop in three distributed file system one for positive sentiment, second for negative sentiment , another one is 0 sentiment means no sentiment tweets in form of felling type.

1.2 PURPOSE OF PROJECT

The main motive of this project is to find the sentiment and opinion of the tweets and also we can say that we check the quality of tweets. So this type of project is necessary because today is the era of socialization anyone want to connect through social media for taking a more advantage. So there are various problems in existing system so every organisation need such type of person or technology that tell the organisation of their brands what the people thinks about any organisation so that the take decision of their problems so that organisation and enterprise take more and more benefits.

The main purpose of this project is as follows:-

- Determining marketing strategy
- Improve campaign success
- Improve product messaging
- Improve customer services
- Generate leads

1.3 SCOPE OF PROJECT

Sentiment analysis is a uniquely powerful tool for businesses that are looking to measure attitudes, feelings and emotions regarding their brand.

The majority of sentiment analysis projects have been conducted almost exclusively by companies and brands through the use of social media data, survey responses and other hubs of user generated content

.

- The future of sentiment analysis is going to continue to dig deeper, far past the surface of the number of likes, comments ,shares and reach, and truly understand significance of social media interaction and what they tell us about the brand.
- Data pre-processing using more parameters to get best sentiment.
- Updating dictionary for new synonyms and antonyms already existing words.

- It helps to analyze the intent like views complain suggestions etc.
- This type of project is useful for natural processing language.
- This type of project is help in Consumer voice

- This project based on rating of tweets so this type of project help to tells the brand reputation checking the sentiment of peoples.
- This type of project is playing a wonderful role in online advertising and online commerce

So this project is best projects related to the data science it also so this type so project is help in politic also for their Voting advise application and Clarification of politicians positions.

This type or sentiment analysis is help in Public Actions like real-world events monitoring, legal matters, and intelligent transaction system.

CHAPTER – 2

SOFTWARE AND HARDWARE REQUIREMENT STUDY

INTRODUCTION :-

This chapter is regards study of hardware and software required during project complementation. As you know this project is regard the problems of Big Data so we need high configuration system hardware.

This project based new technology so here we used new software like Hadoop framework and new languages.

So here we mentioned that technologies, hardware and software-

2.1 HARDWARE REQUIREMENT

It deal what type of hardware is required to complete the project and what is the system configuration is required is mandatory to install is available because Hadoop framework require red very high configuration Linux system.

2.1.1 LINUX AS OPERATING SYSTEM :

As we know that Hadoop does not support other operating system except Linux operating system. So it is mandatory to used Linux operating system for Hadoop.

2.1.2 RAM REQUIRED:

For Hadoop minimum 8 GB RAM is required.

2.1.3 HARDDISK :

Minimum 1TB hard disk is enough because we are not used Hadoop as distributed manner.

2.2 SOFTWARE REQUIRMENT

We used various tools and languages as per project requirement. So we used here Hadoop framework, Pig, and Bitwise software for data ingestion.

2.2.1 HADOOP

Hadoop is the solution of Big Data problems so we used Hadoop framework. With the help of hadoop we can process, and store the large amount of data.

Hadoop ecosystem contains so many tools to handle this type of problems that discuss above.

For storage Hadoop provide storage for large amount of data in distributed manner through which we handle such type of problems very easily. It provide fast processing system than the other commodity system. It is developed by Dug Cutting and Mike Cafarella in 2005 it the project of yahoo that is Nutch search engine. We used HDFS for storage in ecosystem for raw data storage and also final result or data.

2.2.2 PIG

Pig is the tool or platform that used Pig Latin languages for process the large amount of data. This is developed in 2006 when hadoop comes in market because most of the programmer faces so many problems to written Map Reduce code than Pig Latin language is developed by Yahoo because it is the part of Yahoo project.

2.2.3 BITWISE

This software is used for data ingestion process with help of this software we transfer data from one system to another by taking remotely the both system.

CHAPTER – 3

PURPOSED METHODOLOGY

INTRODUCTION

This chapter is regards with the purposed methodology used during the project starting till the complementation. This project divided into the various phase like data ingestion, data wrangling, refining of data and last phase of this project is data analysis or result analysis. So understand one by one.

3.1 Data Ingestion

This is the first stage of this project in this phase we take data from Github repository and import that data into hadoop distributed file system.

Data ingestion deals with data import and export process from one database to another database. We stored raw data in to distributed manner through which we can process that data easily and fast as we can see figure no. 2.0

Figure 2.0 show data stored in hadoop distributed file system in that figure one is .text file that is the AFFIN word rating dictionary and another is the excel file that contain demonetization of tweets.

3.2 Data wrangling

This the second phase of project. This stage comes after the data ingestion in this stage we take data in standard format through which we can handle or process easily when data comes un this format every rows become observation and every column become variable so we can handle data easily.

In this stage we load data from hadoop distributed file system to Pig engine because we use Pig Latin languages to process the data and pig also a platform that stored data temporary for processing purpose and after processing we can stored data into hadoop distributed file system.

3.3 Data refining

This is the third stage of project and this is very important phase of project or we can say that in data science project this is the very complex and important stage because in this stage we deal with the outlier of the data.

Outlier may something that affect our data. In this tweets section there are various type of outlier like in comment most tweets contain braces, some type of special symbol. They type object gives error during the process of data

So in this stage we select tweets part of data and clean it for further usage and this is done by the help of Pig Latin languages.

3.4 Data Mining

This is main phase of the project in the phase we deal with the data mining stage as we know that in Data mining we select that data that is useful for us. Means that is the form of information.

In the phase we select two variables one is text means the column that contain demonetization tweets and process it further as per requirement.

3.5 Analysis Methodology

The is the final stage of data set or this project in the phase we process data as per our requirement like first we find out the average rating of the demonitization tweets by the data that obtained from mining process.

In the divided data or information into three category one for average rating, second is positive rating, and last one is negative rating. Positive rating means sentiment of tweets is positive and negative means negative sentiment and if average rating is equal to the zero means there is no sentiment and we can say that tweets is the polar form means this tweets is the felling type.

Finally we stored process information into three directory in the hadoop.

CHAPTER – 4

CODING AND SCREEN SHOTS

INTRODUCTION

This chapter include all coding part of project because project have several phase like data ingestion, data wrangling, data mining and last one is data analysis of data that comes after the process the raw data or tweets data.

Here we mentioned all code that used for refining of data, mining, and analysis the result with their sentiment. Basically all code written in PIG Latin languages and also code of Hadoop that how to write a data on Hadoop in block format.

All snapshots of project is mentioned in this chapter that defined how HDFS stored data in block format. What is final output, and how namenode stored metadata.

3.1 CODING

```
/*Loading the csv file using pig*/

load_tweets = LOAD '/bharat_project_data/demonetization-tweets.csv' USING
PigStorage(',');


/* Extracting the needed data.... */

/* $0 represents first field and $1 represents the second field.... */

/* "id" is the alias name of $0 and "text" is the alias name of $1 */

extract_details = FOREACH load_tweets GENERATE $0 as id,$1 as text;


 /* The following characters are considered to be word separators: space, double
quote("), coma(,) parenthesis(()), star(*).The following characters are
considered to be word separators: space, double quote("), coma(,)
parenthesis(()), star(*). */

/* TOKENIZE will split the records based on above seperators and give the bag
of words.... */

/* FLATTEN will remove the parenthesis like () and {} */
```

/* Below code of FLATTEN(TOKENIZE(text)) will remove "{}" from the bag of words and tokens i.e. words from each record will get associate with each record as 3rd field in a iterative manner until the tokens will be finished */

```
tokens = foreach extract_details generate id,text, FLATTEN(TOKENIZE(text))
As word;
```

/* For checking the affect of step 'tokens', use the below commands */

```
tokens_limit = LIMIT tokens 8;

dump tokens_limit;
```

/* Loading the dictionary which have rating for words */

```
dictionary = LOAD '/input_files/AFINN.txt' using PigStorage('\t')
As(word:chararray,rating:int);
```

/* 'replicated' join is the special type of join in which second relation is small enough to fit into the main memory which will help in efficient join.... */

```
word_rating = join tokens by word left outer, dictionary by word using
'replicated';
```

/* This will give sample data i.e. first 3 rows of word_rating */

```
word_rating_limit = limit word_rating 3;

dump word_rating_limit;
```

/* We are iteratively selecting the data from the relation "word_rating" and selecting fields from respective relations */

/* To select the field which is the part of a certain relation we need to use double colon "::" as below in case of some operation has

already been applied between two relations like in previous steps */

```
rating = foreach word_rating generate tokens::id as id,tokens::text as text,
dictionary::rating as rate;
```

/* This will give sample data i.e. first 3 rows of rating */

```
rating_limit = limit rating 3;
```

/* This below command will group relations "rating" on the basis of id and text combinedly */

```
word_group = group rating by (id,text);
```

/* First 2 rows of word_group */

```
word_group_limit = limit word_group 2;
```

/* This line is crucial in deciding the rating for any tweet on twitter data as it will sum up all the word's rating in a tweet */

/* Below line will generate group which is from relation word_group and as all the data will get group on the basis of (id,text) we can perform average of those tokens also as per their ratings */

```
avg_rate = foreach word_group generate group, AVG(rating.rate) as
tweet_rating;
```

/* First 100 rows of avg_rate */

```
avg_rate_limit = limit avg_rate 100;

dump avg_rate;
```

/* Filter the positive tweets*/

```
positive_tweets = filter avg_rate by tweet_rating >= 0;
```

/* Filter the negative tweets*/

```
negative_tweets = filter avg_rate by tweet_rating < 0;
```

/* Storing the positive and negative tweets output in output_files folder*/

```
store positive_tweets into '/project_files/positive_tweets_output1';

store negative_tweets into '/project_files/negative_tweets_output1';
```

NOTE : The final output data stored in HDFS in Hadoop framework . as we know that Hadoop stored data in block format so final output data is in two blocks but after filtering data it stored in 3 different file or folder format as given below.

1st - for positive tweets

2nd – for neutral tweets

3rd – for negative tweets

So we make three Excel file or sentiment of tweets one for Negative tweets, neutral or feeling and other Excel file contained positive sentiment of tweets data.

3.2 SCREEN SHOTS

This contain all the screen shots of project . Fig1.0 is contain raw data that used for project

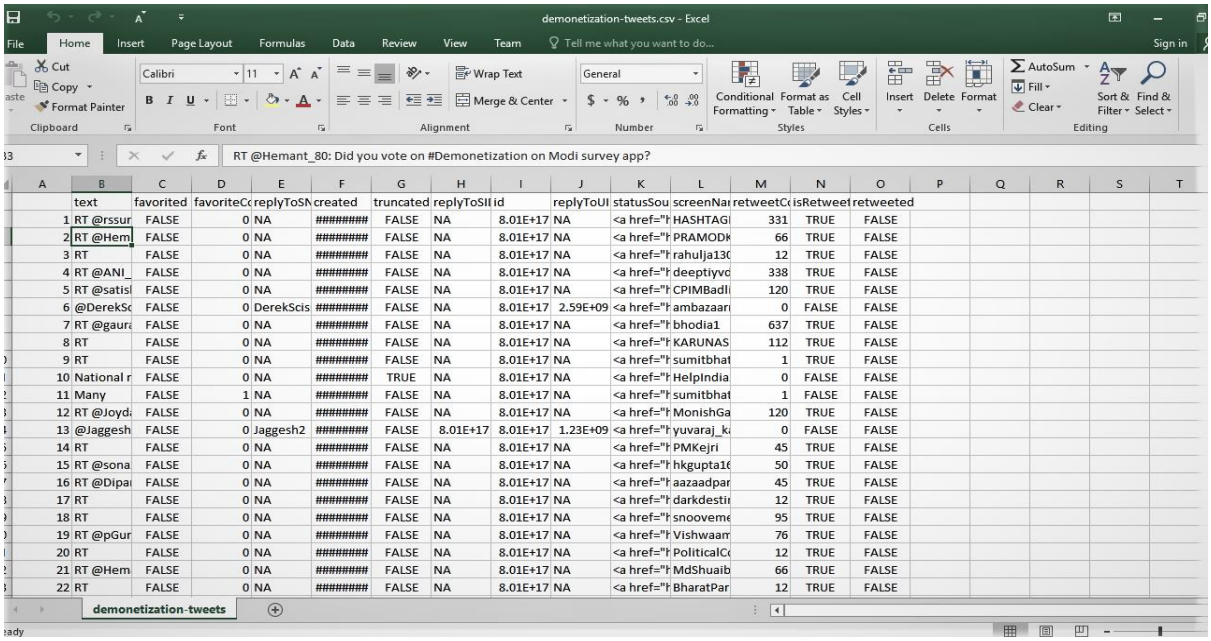

Fig.1.0 (Contain raw data)

This file contain various type attribute in data set. we take only 2 attribute dataset. Here first column is serial number and another filed is of text type that contain tweets of dataset

This figure1.2 contains the Dictionary that used for find out the sentiment form data. This figure contains words with their rating. Every word have own rating some word have negative rating, some have positive rating and some have 0 rating.

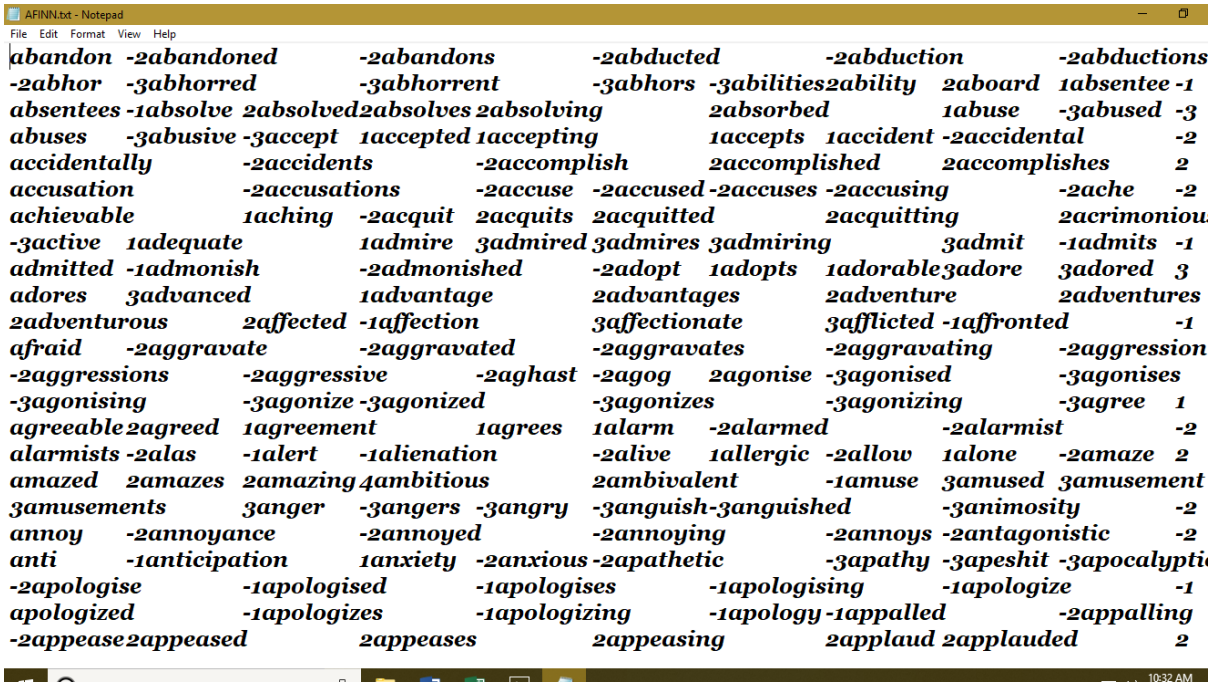

Fig 2(Words dictionary)

The word have rating 0 is the type of felling so thier is no sentiment in the that word, the word have positive rating means good sentiment regards the tweets and negative rating means negative sentiment towards tweets.

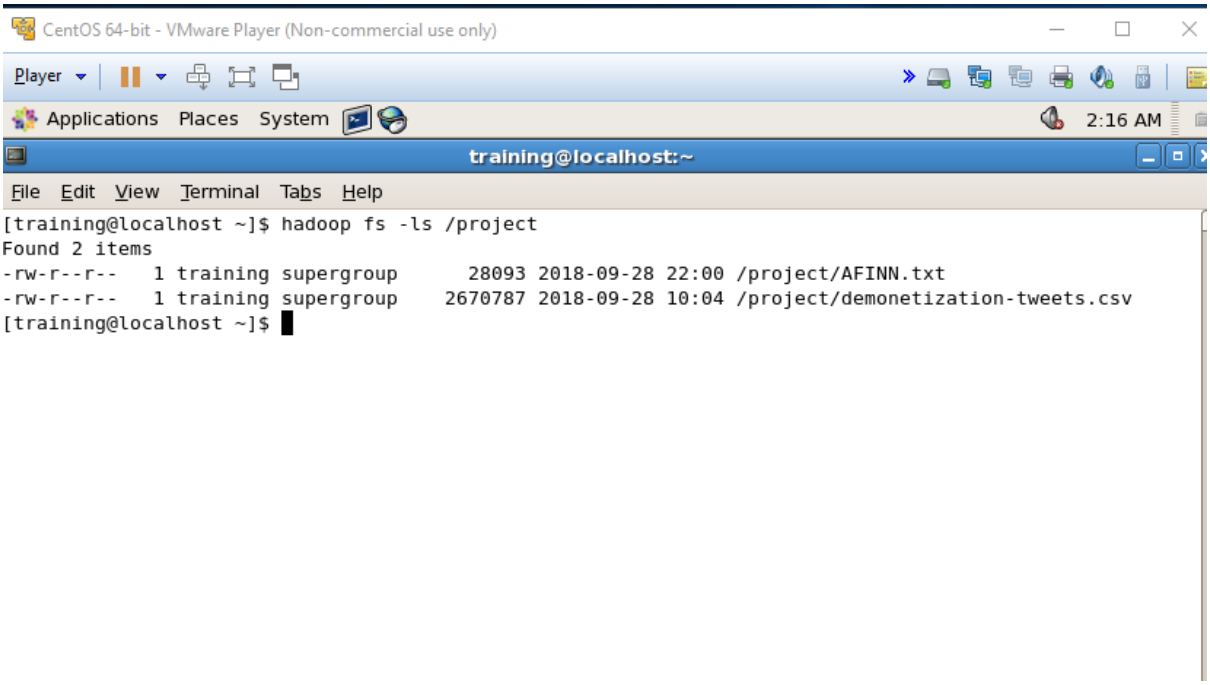

Fig 3.0 (raw data)

This figure contain raw data of the project that stored on the hadoop distributed file system. First file contain words dictionary and second file contain raw data or demonetization of tweets dataset.

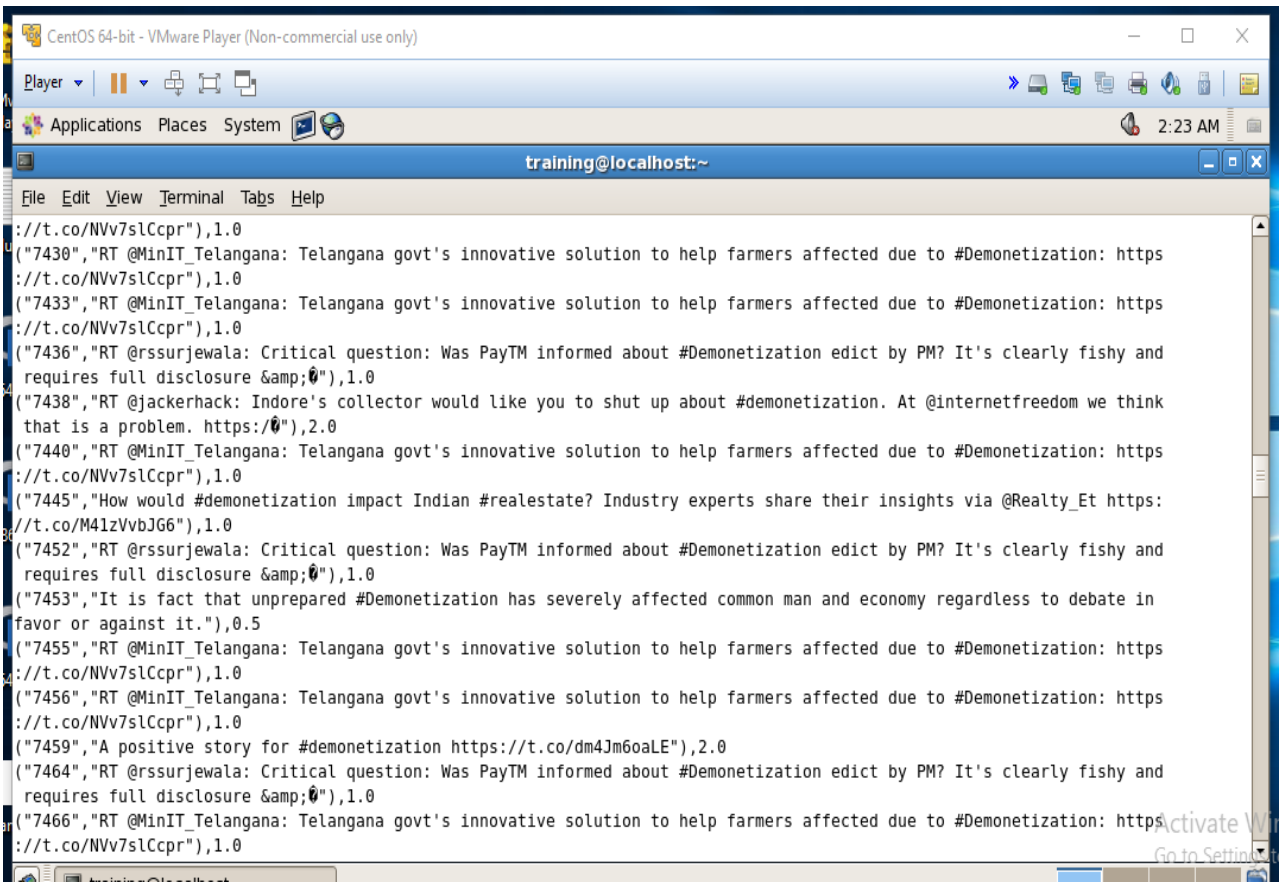

Fig 4.0(Positive rating)

This image or figure2.1 contain positive rating dataset that comes after processing the dataset.

This figure 2.2 contain the negative rating of tweets or we can say negative sentiment of tweets

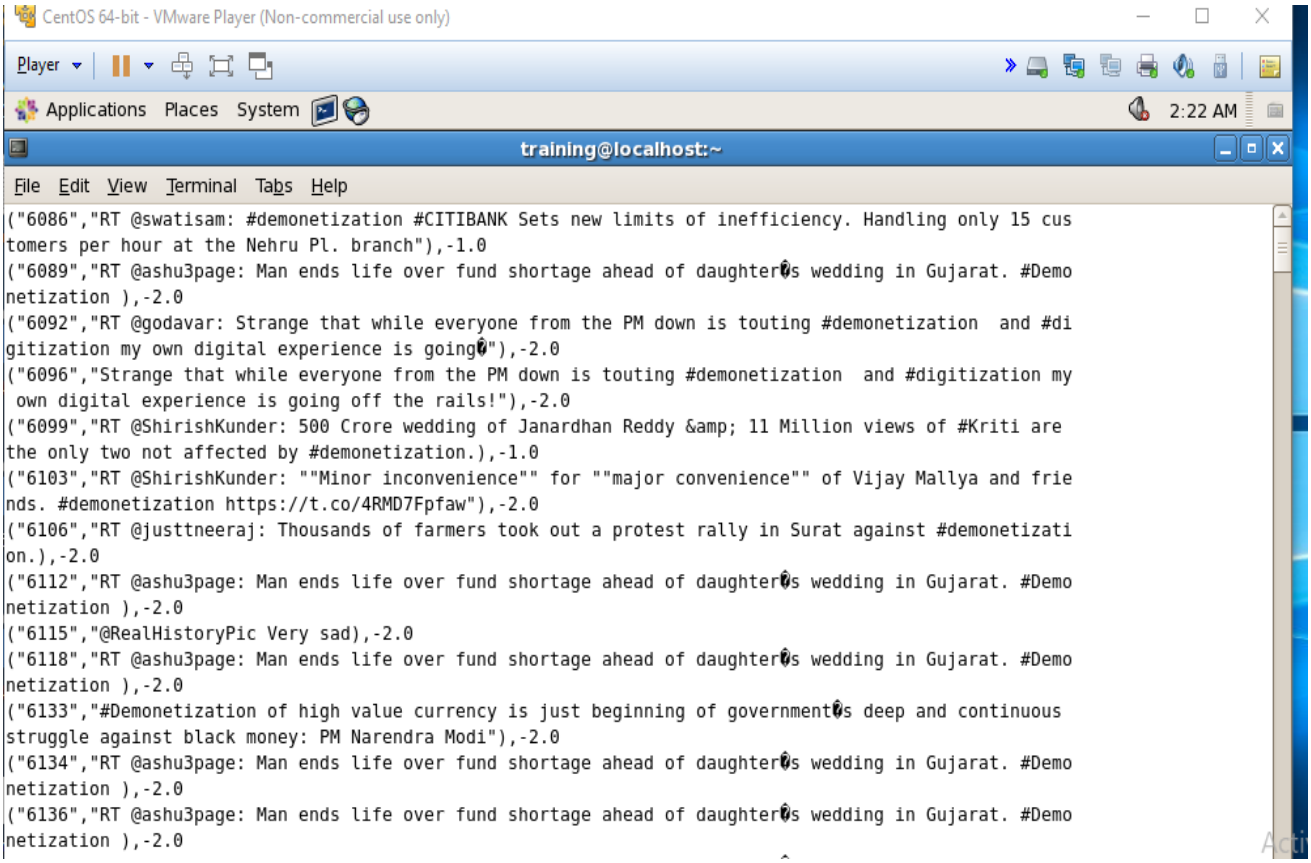

Fig 5.0(Negative Rating)

These all are the images of raw dataset and final result of dataset that obtained after processing.

CHAPTER – 5

CONCLUSION AND REFERENCES

5.1 CONCLUSION

This project is totally based on the Big data problems that Is the sub part of Data Science so this project is very good for any type of organization to help the understand the sentiment of their customers and users.

For doing this project there is another technologies that current comes in market is mandatory to learn because there is required much or deep knowledge of this area to complete such type of project. So for this project we need good knowledge of Hadoop framework.

For this project we are used Pig Latin Language for both processing the data.

We provide storage on Hadoop distributed file system.and analyze the result with Pig Latin. We separate the tweets on the based on the their polarity. Polarity in the sense their tweets quality of tweets whether tweet is negative, positive and and polarity is neutral(feeling type).

So this project is very help to understand the quality of tweets so it help to understand the sentiment of the people. It is the project of Big data problem so we are lots of think we learn.

In this project we take data from Github repository. Data is dummy because this project is learning purpose we stored that data on hadoop distributed file system to handle the data because data is semi structure so hadoop provide best solution for this type of problems.

Second phase of this project is the data cleaning or refining of data in the phase we clean the data as per our requirement and take off only that data are useful for us.

Last phase of this project is data analysis. In this stage we further store data(or clean data) on the hadoop distributed file system for further.

In this phase we checks the polarity of the sentiment of the comment. We define polarity term in chapter 2. Polarity are three type negative, positive and neutral(feelings) .and final stored result of hadoop distributed file system.

5.2 BIBLIOGRAPHY

Chuck Lam, Hadoop in Action,2010, www.manning.com/HadoopinAction.

Hadley Wicham , R for Data Science,2016, www.r4ds.had.co.nz

Allen B. Downey, Think Stats,2011, http://bit.ly/think_stats_2e.

Wikipedia.