PROJECT REPORT

ON

SENTIMENT ANALYSIS ON DEMONETIZATION BASED ON TWITTER DATA

SUBMITTED BY :

ADITHYA H
ANAND P P
FARHAT IMAM
MAHITH P
PRIYANKA K

Under the guidance of Mr. PRASOON KUMAR KG
(SENIOR TECHNICAL OFFICER)

Submitted in partial fulfilment of the requirements for qualifying
ADVANCED DIPLOMA IN BIG DATA ANALYTICS

National Institute of Electronics and Information Technology
(An Autonomous Scientific Society of Department of
Electronics and Information Technology (Deity)
Ministry of Communications and Information Technology, Government of India)
NIT CAMPUS Calicut 673601

ACKNOWLEDGEMENT

It’s a great responsibility to express our sincere thanks to all the people, who have contributed to the successful completion of our project work through their support, encouragement and guidance. Our first and foremost thanks go to Dr. M P PILLAI, Executive Director (Scientist / Engineer ‘G’) NIELIT Calicut , for all the inspiration and encouragement and gave the opportunity to associated with NIELIT by studying in the Advanced Diploma in Big-Data Analytics. It is our privilege to place on record our indebtedness to Mr. PRASOON KUMAR K G, Senior Technical Officer, NIELIT Calicut, for his immense support at every stage of the project and course work. We owe our gratitude and thanks to, Mrs. VIMALA MATHEW, Scientist /Engineer ‘D’, NIELIT Calicut, for her constant guidance and help all through our work. We owe our gratitude and thanks to, Mrs. RESHMA C B, Senior Technical Officer, NIELIT Calicut, for her constant guidance and help all through our work. We owe our gratitude and thanks to, Mr. Hari K, Technical Officer, NIELIT Calicut, for his help in resolve software installation work. We would like to express our thanks to all teaching and non-teaching staffs of our institute and our friends for their help in successful completion of this project work. We would like to express our sincere thanks to God almighty, for His love and grace that He has bestowed upon us, finally we thank our parents for their moral, financial support and encouragement.

ABSTRACT

After the Indian government banned Rs.500 and Rs.1000 currency notes, the economy has taken a hit. While the jury is still out as to whether the move is a good one or not and it has created trouble for common people. Tools such as Big Data can be effective in collecting information about financial misappropriation. This project describes various steps required to perform sentiment (positive /negative / neutral) analysis on Demonetization by analysing the twitter dataset available. Sample datas are ingested from twitter. Python ,Hive ,Pig and Hadoop are used for analysis of these data.

CONTENTS

1. INTRODUCTION

SENTIMENT ANALYSIS, also called opinion mining, is the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc.

The most important indicators of sentiments are sentiment words, also called opinion words. These are words that are commonly used to express positive or negative sentiments. For example, good, wonderful and amazing are positive sentiment words, and bad, poor, and terrible are negative sentiment words. Apart from individual words, there are also phrases and idioms, e.g., cost someone an arm and a leg. Sentiment words and phrases are instrumental to sentiment analysis for obvious reasons. A list of such words and phrases is called a sentiment lexicon (or opinion lexicon). Over the years, researchers have designed numerous algorithms to compile such lexicons. Although sentiment words and phrases are important for sentiment analysis, only using them is far from sufficient. The problem is much more complex. In other words, we can say that sentiment lexicon is necessary but not sufficient for sentiment analysis. Below, we highlight several issues:

A positive or negative sentiment word may have opposite orientations in different application domains. For example, 'suck' usually indicates negative sentiment e.g. 'This camera sucks', but it can also imply positive sentiment. Many sentences without sentiment words can also imply opinions. Many of these sentences are actually objective sentences that are used to express some factual information. The sentence “This washer uses a lot of water” implies a negative sentiment about the washer since it uses a lot of resource (water). All these sentences have no sentiment words.

BIG DATA means really a big data, it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, technqiues and frameworks.

Big data is really critical to our life and its emerging as one of the most important technologies in modern world. Follow are just few benefits which are very much known to all of us:

1. Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.

2. Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.

3. Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business.

To harness the power of big data, you would require an infrastructure that can manage and process huge volumes of structured and unstructured data in realtime and can protect data privacy and security.

There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. This includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data.

MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high and low end machines.The major challenges associated with big data are as follows:

- Capturing data
- Storage
- Searching
- Sharing
- Transfer
- Analysis

2. PROGRAMMING LANGUAGES USED

2.1 HADOOP

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage. Hadoop MapReduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner. HDFS uses a master/slave architecture where master consists of a single NameNode that manages the file system metadata and one or more slave DataNodes that store the actual data. A file in an HDFS namespace is split into several blocks and those blocks are stored in a set of DataNodes. The NameNode determines the mapping of blocks to the DataNodes. The DataNodes takes care of read and write operation with the file system. They also take care of block creation, deletion and replication based on instruction given by NameNode.

MapReduce: It is a parallel programming model for processing large amounts of structured, semi-structured, and unstructured data on large clusters of commodity hardware.

2.2 APACHE PIG

APACHE PIG is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets. It is a procedural language platform used to develop a script for MapReduce operations.

The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark Pig Latin abstracts the programming from the Java MapReduce idiom into a notation which makes MapReduce programming high level, similar to that of SQL for RDBMSs. Pig Latin can be extended using User Defined Functions (UDFs) which the user can write in Java, Python ,Javascript ,Ruby or Groovy and then call directly from the language.

2.3 PYTHON

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.It supports functional and structured programming methods as well as OOP. It can be used as a scripting language or can be compiled to byte-code for building large applications. It provides very high-level dynamic data types and supports dynamic type checking.

There is a collection of command style functions in python ie: matplot.pyplot that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

2.4 HIVE

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. It is a platform used to develop SQL type scripts to do MapReduce operations.

It stores schema in a database and processed data into HDFS. It is designed for OLAP. It also provides SQL type language for querying called HiveQL or HQL. It is familiar, fast, scalable, and extensible.

Hive chooses respective database servers to store the schema or Metadata of tables, databases, columns in a table, their data types, and HDFS mapping. Hadoop distributed file system or HBASE are the data storage techniques to store data into file system.

2.5 R LANGUAGE

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.

R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac. This programming language was named R, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs Language S.

3. PROGRAMS

3.1.1 Using PIG

Step 1: Loading the data into pig using PigStorage

```
  tweets = LOAD '/usr/local/hadoop/pig_project/demonetization-tweets.csv' USING PigStorage(',');
```

Step 2: After loading you can see the tweets loaded successfully into pig by using the dump command

```
  dump tweets;
```

Step 3: Extract the id and the tweet_text

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

Step 4:Now we will divide the tweet_text into words to calculate the sentiment of the whole tweet.

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

Step 5: Loading AFFIN Dictionary using pig

```
    dictionary = load '/usr/local/hadoop/pig_project/AFINN.txt' using PigStorage('\t') AS(word:chararray,rating:int);
```

Step 6:Joining the tokens statement and the dictionary contents

```
   word_rating = join tokens by word left outer, dictionary by word;
  describe word_rating;
```

Step 7:Extracting the id,tweet text and word rating

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

Step 8:Grouping the rating of all the words in a tweet

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

Step 9:performing the Average operation on the rating of the words

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

Step 10: Storing the avg_rate in local file

```
  STORE avg_rate  INTO '/usr/local/hadoop/pig_project/avg_data/' USING PigStorage(',');
```

3.1.2 Extracting Negative Response

Step 1: Loading the data into pig using PigStorage

```
 tweets = LOAD '/usr/local/hadoop/pig_project/demonetization-tweets.csv' USING PigStorage(',');
```

Step 2: After loading you can see the tweets loaded successfully into pig by using the dump command

```
 dump tweets;
```

Step 3: Extract the id and the tweet_text

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

Step 4:Now we will divide the tweet_text into words to calculate the

sentiment of the whole tweet.

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

Step 5: Loading AFFIN Dictionary using pig

```
        dictionary = load '/usr/local/hadoop/pig_project/AFINN.txt' using PigStorage('\t') AS(word:chararray,rating:int);
```

Step 6:Joining the tokens statement and the dictionary contents

```
  word_rating = join tokens by word left outer, dictionary by word;
  describe word_rating;
```

Step 7:Extracting the id,tweet text and word rating

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

```
describe rating;
```

Step 8:Grouping the rating of all the words in a tweet

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

Step 9:performing the Average operation on the rating of the words

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

Step 10: Filter the negatives tweets

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

Step 11: Storing the negative_tweets in local file

```
STORE negative_tweets INTO '/usr/local/hadoop/pig_project/neg_data' USING PigStorage(',');
```

3.1.3 Extracting Positive Response

Step 1: Loading the data into pig using PigStorage

```
tweets = LOAD '/usr/local/hadoop/pig_project/demonetization-tweets.csv' USING PigStorage(',');
```

Step 2: After loading you can see the tweets loaded successfully into pig by using the dump command

```
dump tweets;
```

Step 3: Extract the id and the tweet_text

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

Step 4:Now we will divide the tweet_text into words to calculate the sentiment of the whole tweet.

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

Step 5: Loading AFFIN Dictionary using pig

```
    dictionary = load '/usr/local/hadoop/pig_project/AFINN.txt' using PigStorage('\t')
AS(word:chararray,rating:int);
```

Step 6:Joining the tokens statement and the dictionary contents

```
  word_rating = join tokens by word left outer, dictionary by word;
  describe word_rating;
```

Step 7:Extracting the id,tweet text and word rating

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

Step 8:Grouping the rating of all the words in a tweet

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

Step 9:performing the Average operation on the rating of the words

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

Step 10: Filter the positive tweets

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

Step 11: Storing the positive_tweets in local file

```
   STORE positive_tweets INTO '/usr/local/hadoop/pig_project/positive_data/positive.csv' USING
pigStorage(',');
```

3.2.1 Using Hive -Positive

```
CREATE TABLE IF NOT EXISTS positive_rating1 ( id String, text String, rating FLOAT)  ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' ;
```

```
load data local inpath '/home/sw3/Documents/jan11/positive1.csv'into table positive_rating1 ;
```

```
select  rating , COUNT(rating)  from positive_rating1 GROUP BY positive_rating1.rating;
```

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/sw3/Documents/jan11/positive1'ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' select
rating , COUNT(rating)  from positive_rating1 GROUP BY positive_rating1.rating ;
```

3.2.2 Using Hive -Negative

```
CREATE TABLE IF NOT EXISTS negative_rating1 ( id String, text String, rating FLOAT)  ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' ;
```

```
load data local inpath '/home/sw3/Documents/jan11/negative.csv'into table negative_rating1 ;
```

```
select  rating , COUNT(rating)  from negative_rating1 GROUP BY negative_rating1.rating;
```

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/sw3/Documents/jan11/negative1'ROW
FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' select
rating , COUNT(rating)  from negative_rating1 GROUP BY negative_rating1.rating ;
```

3.3.1 Python (Positive Response)

Step 1: importing packages for graph plotting and for using array

```
 import numpy as np
import pandas as pd
from pandas import *
```

Step 2: reading csv file

```
p=pd.read_csv('/home/sw3/Documents/jan11/positive1/positive_trating.csv',header = None)
```

Step 3: initialise the arrays of rating and occurance for positive response

```
p=p.rename(columns={0:'rating'})
p=p.rename(columns={1:'total'})
print p
```

Step 4: dataframe to csv

```
p.to_csv('/home/sw3/Documents/jan11/hivepgms/negative1panda.csv')
```

3.3.2 Python (Negative Response)

Step 1: importing packages for graph plotting and for using array

```
import numpy as np
import pandas as pd
from pandas import *
```

Step 2: reading csv file

```
p=pd.read_csv('/home/sw3/Documents/jan11/negative1/negative_trating.csv',header = None)
```

Step 3: initialise the arrays of rating and occurance for positive response

```
p=p.rename(columns={0:'rating'})
p=p.rename(columns={1:'total'})
print p
```

Step 4: dataframe to csv

```
p.to_csv('/home/sw3/Documents/jan11/hivepgms/positive1panda.csv')
```

3.4.1 Plotting in R - Negative

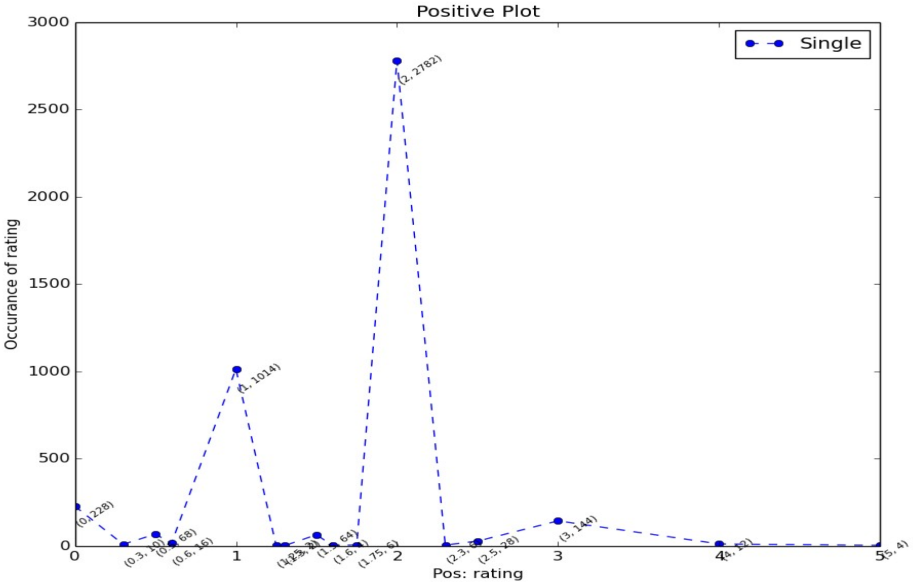
```
library(ggplot2)
library(reshape)

#read data
data = read.table(file.choose(), header=T,sep=",")

#melt data “rating vs. all”
data2=melt(data,id=c("rating"))
data2
max(data2$rating)
qplot(rating,value,data=data2,geom="line",linetype=variable)
```

3.4.2 Plotting in R - Positive

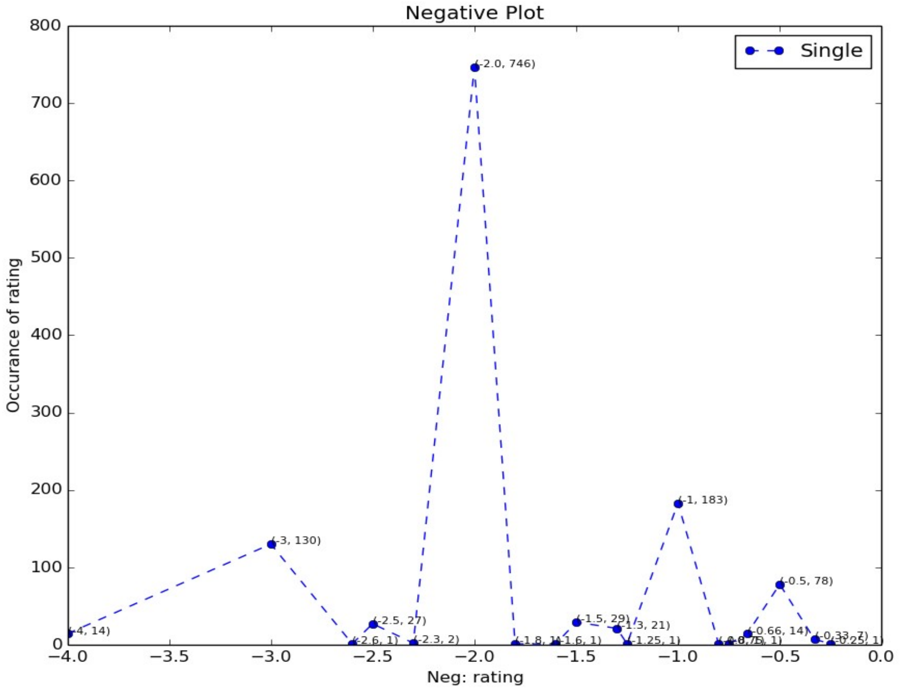
```
library(ggplot2)

library(reshape)


#read data

data = read.table(file.choose(), header=T,sep=",")


#melt data “rating vs. all”

data2=melt(data,id=c("rating"))

data2

max(data2$value)

qplot(rating,value,data=data2,geom="line",linetype=variable)
```

4. OUTPUT

4.1 Output of Positive Rating



4.2 Output of Negative Rating



5.CONCLUSION

Twitter, one of the largest social media site receives tweets in millions every day in the range of Zettabyte per year. This project provides a way of analyzing of big data such as twitter data using Python, Pig ,Hive, R and Hadoop which will process and analyze the sentiments in tweets on Demonetization topic . This also includes visualizing the results into a pictorial representations of Sentiments (Positive and Negative) tweets.

As in graph trend looks like the tweet sentiment becomes positive as time goes on in the first and later on the sentiment becomes more on Negative.

6. REFERENCES

1. International Journal Of Core Engineering & Management (IJCEM) Volume 1

-Tweet Analysis: Twitter Data processing Using Apache Hadoop

2. Sentiment Analysis and Opinion Mining - Bing Liu

3. https://gist.github.com

4.https://github.com/ujjwalkarn

5.https://acadgild.com/blog