SAVITRIBAI PHULE PUNE UNIVERSITY

A FINAL PROJECT REPORT

ON

TWITTER SENTIMENT ANALYSIS USING HADOOP

Submitted By

Anita Kumari	B120314209
Anjali Kante	B120314247
Shriya Samak	B120314279
Ajinkya Ingle	B120314204

Under The Guidance of

Mrs. J.M. Kanase

Department of Computer Engineering
PES Modern College Of Engineering
Shivajinagar, Pune 411043

[2015]-[2016]

Progressive Education Societys

PES Modern College Of Engineering

Shivajinagar, Pune 411043

CERTIFICATE

This is to certify that the project report entitled

Twitter Sentiment Analysis Using Hadoop

submitted by

1. **Anita Kumari - B120314209**
2. **Anjali Kante - B120314247**
3. **Shriya Samak - B120314279**
4. **Ajinkya Ingle - B120314204**

This bonafied work is carried out by the students and is submitted towards the fulfillment of the requirement of Savitribai Phule Pune University, Pune for the award of the degree of Bachelor of Engineering (Computer Engineering).

Date:

Place:

Internal Guide	Head Of Department	External Examiner
Mrs. J.M. Kanase	(Computer Engineering)	
	Dr. Prof. Mrs. S. A. Itkar	

Project Approval Sheet

Project Title

ON

TWITTER SENTIMENT ANALYSIS USING HADOOP

Is successfully completed by

Anita Kumari	B120314209
Anjali Kante	B120314247
Shriya Samak	B120314279
Ajinkya Ingle	B120314204

Under The Guidance of

Mrs. J.M. Kanase

Department of Computer Engineering

PES Modern College Of Engineering

Abstract

In todays highly developed world, every minute, people around the globe express themselves via various platforms on the Web. And in each minute, a huge amount of unstructured data is generated. This data is in the form of text which is gathered from forums and social media websites. Such data is termed as big data. User opinions are related to a wide range of topics like politics, latest gadgets and products. These opinions can be mined using various technologies and are of utmost importance to make predictions or for one-to-one consumer marketing since they directly convey the viewpoint of the masses. Here we have developed an application to analyze the sentiments of Twitter users through their tweets in order to extract what they think. Hence we are using dictionary based approach of sentiment analysis and map reduce technique of hadoop which will process the huge amount of data on a hadoop cluster faster.

In map-reduce we have implemented two functions which are mapper and reducer functions , in mapper function we compare each word in a tweet with positive and negative dictionaries and accordingly it assigns polarities to individual words. In reducer function we calculate the sum of positive and negative words and finally the result is displayed in the form of pie charts country wise

Acknowledgement

It gives us great pleasure in presenting the final project report on Twitter sentiment analysis using hadoop. The success of project work depends largely on the encouragement and guidelines of many others.

The guidance and support received from all the Professors who contributed and who are contributing to this project work, was vital for the success of the proposed work.

We would like to take this opportunity to thank our internal guide Mrs. J.M.Kanase for giving us all the help and guidance we needed. We are really grateful for her support. Her valuable suggestions and indispensable support were very helpful.

Anita Kumari 41063

Anjali Kante 41018

Shriya Samak 41036

Ajinkya Ingle 41012

(B.E. Computer Engg.)

Table Of Contents

List of Tables

List of Figures

1.

Synopsis

1.1 Project Title

Twitter Sentiment Analysis Using Hadoop

1.2 Project Option

Internal Project

1.3 Internal Guide

Guide : Mrs J.M. Kanse

1.4 Technical Keywords (As per ACM keywords)

- Opinion Mining
- Sentiment Analysis
- MapReduce
- cluster
- Unstructued
- Hadoop

1.5 Problem Statement

Using Dictionary based approach of sentiment analysis and map-reduce technique of hadoop assign contextual polarity to each word in a tweet in order to extract sentiments of people and classify them into positive, negative and neutral sentiments

1.6 Abstract

In todays highly developed world, every minute, people around the globe express themselves via various platforms on the Web. And in each minute, a huge amount of unstructured data is generated. This data is in the form of text which is gathered from forums and social media websites. Such data is termed as big data. User opinions are related to a wide range of topics like politics, latest gadgets and products. These opinions can be mined using various technologies and are of utmost importance to make predictions or for one-to-one consumer marketing since they directly convey the viewpoint of the masses. Here we have developed an application to analyze the sentiments of Twitter users through their tweets in order to extract what they think. Hence we are using dictionary based approach of sentiment analysis and map reduce technique of hadoop which will process the huge amount of data on a hadoop cluster faster.

In map-reduce we have implemented two functions which are mapper and reducer functions , in mapper function we compare each word in a tweet with positive and negative dictionaries and accordingly it assigns polarities to individual words. In reducer function we calculate the sum of positive and negative words and finally the result is displayed in the form of pie charts country wise

1.7 Goals and Objectives

The main objective of this project is to classify the sentiments of people into positive and negative sentiments. Consumers opinions provide valuable information about the companies as they help to understand how their products and services are perceived Extracting sentiments of people out of their opinions play an imporatnt role to increase their market value and help them to improve their products with respect to customer demands After calcualting the no of positive and negative sentiments of people we represented data in the form of graphs and pie charts using R programming

1.8 Plan of Project Execution

Dates	**Task**	**Resposible Person**
7th July 2015	Domain selection	Team
2nd week of July 2015	Topic Finalization	Team
3rd week of July 2015	Feasibility Study and Market Potential Analysis	Team
28th July 2015	Abstract Submission Team	Team
20 august 2015	System Architecture Design	Team
22 august 2015	Discussion about Platform issues	Team
23rd august 2015	Preparing Synopsis	Team
Last week of September 2015	Project Report Submission for Stage I	Team
1st week of December 2015	Discussion about platform Selection	Team
2nd week of January 2016	Design and Coding	Team
3rd of February 2016	Testing	Team
April 2016	Final Report Preparation	Team

Table 1.1: Project Plan Table

2.

Introduction

2.1 Sentiment Analysis

Sentiment analysis also known as opinion mining.The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product.

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.

Hence, there is a need to develop a product which can analyze opinions of people.This product will be useful in increasing market value of industries.As well as satisfy needs of customers.

2.2 Motivation of The Project

Today we are living in the world which is surrounded by 99 percent of data. There are different microblogging sites where users express their views about different products these views are nothing but opinions of people and it will go waste if it is not used in proper way so there is a need to use opinions of people in improving productivity, usefulness, functionality of particular product or application or technique or any entertainment resource.

Hence, there is a need to develop a product which can analyze opinions of people. This product will be useful in increasing market value of industries. As well as satisfy needs of customers.

2.3 Literature Survey

Title	**Author**	**Description**
Sentiment analysis of movie Review	Shravan Vishwanathan	In this paper they have analyzed sentiments of people to predict the result that are based on users opinion about movies.
Sentiment Analysis of Twitter Data	Apoorv Agarwal ,BoyiXie ,Ilia Vovsha ,Owen Rambow ,Rebecca Passonneau	In this paper they have used two methods:- (1) Introduction of POSspecific prior polarity features. (2)Use of a tree kernel to Obviate the need for tedious.
Real time sentiment analysis of twitter data using Hadoop	Suni B Mane and	This paper describes real time sentiment analysis of Twitter data.
Twitter sentiment analysis using Apache Storm	Ishana Raina,Sourabh gujjar	Here they have proposed a system to analyze the tweets of Twitter users through their tweets in order to extract what they think. We classify their sentiments into three different polarities positive, negative and neutral.
Evaluation Datasets for Twitter Sentiment Analysis	Hassan Saif, Miriam Fernandez, Yulan He and Harith Alani	In this paper they have presented an overview of eight publicly available and manually annotated evaluation datasets for Twitter sentiment analysis
Decision Making Using Sentiment Analysis from Twitter	M.Vasuki, J.Arthi and K.Kayalvizhi	This paper focused to predict the polarity of words and then classify them into positive and negative feelings with the aim of identifying attitude and opinions that are expressed in any form of language
Hive A Petabyte Scale Data Warehouse Using Hadoop	Ashish Thusoo,Joydeep Sen Sarma ,Namit Jain,Zheng Shao,Prasad Chakka ,Ning Zhang	This paper describes sentiment analysis of data using Hive data warehouse.

3.

Problem Defintion and Scope

3.1 Problem Statement

Using Dictionary based approach of sentiment analysis and map-reduce technique of hadoop assign contextual polarity to each word in a tweet in order to extract sentiments of people and classify them into positive, negative and neutral sentiments

3.1.1 Goals and Objectives

- In recent times the use of Social Networking sites have increased tremendously
- Mainly on twitter people share their opinions about any event,product, movies or any treding topics
- Their opinions express their likes and dislikes which can be very beneficial from business point of view if used in a proper manner
- Hence we need a system that analyses the opinions of people and givee review in the form of counts of people who like or dislike a particular product
- This system analyses the tweets of people by breaking them them into individual words and analyse each and every word in the tweet and finally classify them into positive and negative sentiments
- This system can be used in business analytics and in different organizations which desire to satisfy consumers needs

3.1.2 Statement of Scope

- This project is proposed to analyse the sentiments of people by classifying into positive, negative and neutral based on polarity using dictionary based approach of sentiment analysis.
- At first, twitter datasets are downloaded using flume-agent. Necessary fields are taken for further processing using MapReduce
- In the next step, the structured input data is given to mapper function In which, each tweet is split into individual tokens.

- It also assings polarity to each token based on positive,negative and neutral dictionary.
- Individual tokens which are having certain polarities are collected by reducer function and reducer function also calculates the sum of positive,negative and neutral sentiments.
- The proposed system can find most popular information about the people, organizations and can be used in the fields of analytics
- Our proposed system is not analyzing real time tweets and emoticons

3.2 Software Context

The software can be used by cyber crime departments to predict the possibility of physical attacks in any location.

3.3 Major Constraints

- The data we are processing is not live. It is mainly the stored data in text file so our project will not be able to analyze real time tweets.
- Our project will not be able to detect the foreign scripts because the dictionaries that we are using have only English words with their sentiment scores

3.4 Methodologies of Problem Solving and Efficiency Issues

3.5 Outcome

The outcome of the project is to detect and classify the sentiment of word from large collections of tweets on different subjects and finally displaying the total of positive and negative sentiments of people about a particualr topic

3.6 Application

- Business Analytics: Consumers opinions provide valuable information about the companies as they help to understand how their products and services are perceived. So sentiment analysis is used in: Consumer voice, Brand reputation of the products, Online advertising: Blogger Centric Contextual Advertising and Dissatisfaction oriented online advertising, On-line commerce
- Politics: Sentiment analysis is used in voting advise applications and clarification of politicians positions
- Public Actions: Sentiment analysis gives an important contribution in monitoring real world events for example for monitoring critical information about earthquake locations and magnitude, riot locations, this monitoring helps policy makers to minimise damage in areas which are expected to be affected next by such events.
- Policy or government-regulation proposals: Another important application of sentiment analysis is the monitoring of the opinions that people submit about pending policy or government-regulation proposals
- Intelligent transportation system: A new emerging domain of sentiment analysis is Intelligent transportation system(ITSs), for the completeness of ITS space, it is necessary to collect and analyze the public opinions exchange. Traffic sentiment analysis has been developed which allows analysing the traffic problem in a humanizer way

3.6.1 Product Features

- It analyzes opinions of people and classify the sentiments into positive,negative and neutral sentiments
- It can help companies to understand how their products and services are perceived based on opinions of people
- It can help companies to increase the market value of their products by analysing opinions of people about their products
- It can help movie makers to understand the moods of people about what geners of movie people like to watch the most nowdays

- It can help people to have an idea of which movie is worth watching based on movie review

3.7 Harware Resources Required

- CPU Processor:intel i5
- RAM-8 GB

3.8 Software Resources Required

- OS-ubuntu 14.04 LTS
- Virtual Box
- Language used-Java

4.

Project Plan

4.1 Project Estimates

4.1.1 Time Estimates

- It took us approximately 15 days to decide the topic.
- It took us approximately 20 days to do survey of algorithmic approaches.
- It took us approximately 5 days to define the project scope.
- It took us approximately 10 days to finalize the algorithmic approach.
- It took us approximately 10 days to do requirement analysis and feasibility study.
- It took us approximately 55 days to do literature survey and preliminary report preparation.
- It will take us approximately 85 days to do system design and coding.
- It will take us approximately 31 days to do testing.
- It will take us approximately 31 days to do final report preparation.

Cost Estimates

- All open source technologies are used in the project therefore investment in software resources is zero.
- But we do require a computer with minimum RAM of 8GB and processor of 2Ghz.

4.2 Project Resources

Project resources that will be required are as follows:

- A dedicated team of developers.
- A computer with specifications as mentioned in hardware specification.
- Softwares with specifications as mentioned in software specification.

4.3 Risk Management w.r.t to NP hard analysis

This section discusses Project risks and the approach to managing them.

4.3.1 Risk Identification

Risk identification can be defined as the efforts taken to specify threats to the project plan. Risk identification can be done by identifying the known and predictable risks.

The risks that we might face during the developmet of our project are:

1. Performance risk.
2. Cost risk.
3. Support risk.
4. Schedule risk.

4.3.2 Risk Analysis

The risks for the Project can be analyzed within the constraints of time and quality.

ID	Risk Description	Probability	Impact		
			Schedule	Quality	Overall
1	Wrong Functiionality	Medium	Low	High	High
2	Wrong User Interface	Low	Low	High	Medium
3	Power Failure	Medium	Low	High	High
4	Unnecessary Features	Low	Low	Low	Low
5	Hardware Failure	Medium	Low	High	High

Table 4.1: Risk Table

Probability	Value	Description
High	Probability of occurrence is	$> 75\%$
Medium	Probability of occurrence is	$26 - 75\%$
Low	Probability of occurrence is	$< 25\%$

Table 4.2: Risk Probability Definitions

Impact	Value	Description
Very high	$> 10\%$	Schedule impact or Unacceptable quality
High	$5 - 10\%$	Schedule impact or Some parts of the project have low quality
Medium	$< 5\%$	Schedule impact or Barely noticeable degradation in quality Low Impact on schedule or Quality can be incorporated

Table 4.3: Risk Impact Definitions

4.3.3 Overview of Risk Mitigation, Monitoring, Management

Following are the details for each risk.

Risk ID	1
Risk Description	Wrong functionality
Category	Technology
Source	Software requirement Specification document.
Probability	Low
Impact	Very High
Response	Modify Functionality
Strategy	Requirement Analysis
Risk Status	Identified

Risk ID	2
Risk Description	Wrong User Interface
Category	Development Environment
Source	Software Design Specification documentation review.
Probability	Low
Impact	Very High
Response	Modify Interface
Strategy	Requirement Analysis
Risk Status	Identified

Risk ID	3
Risk Description	Power Failure
Category	Technology
Source	This was identified during early development and testing.
Probability	Low
Impact	Very High
Response	Accept
Risk Status	Identified

Risk ID	4
Risk Description	Unnecessary Features
Category	Development Interface
Source	Software Requirement Specification document
Probability	Low
Impact	Low
Response	Remove Features
Strategy	Requirement Analysis and Testing
Risk Status	Identified

Risk ID	5
Risk Description	Hardware failure
Category	Technology
Source	This was identified during early development and testing.
Probability	Low
Impact	Very High
Response	Change components
Strategy	Replace components
Risk Status	Identified

4.4 Team Organization

The manner in which staff is organized and the mechanisms for reporting are noted.

4.4.1 Team structure

The project team consists of four people:

Ajinkya Ingale

Role : Study of the available approaches and code implementation.

Shriya Samak

Role : Algorithmic Analysis, Database Management and data processing,

Anita Kumari

Role : Documentation, Queries application and testing .

Anjali Kante

Role : Partial documentation and GUI design and Testing.

4.4.2 Management reporting and communication

A specific role is given to each member and the role is fulfilled in parts.The parts of work are assigned to every member,per week.Weekly assessment of each part is done by every group member and future task is decided.

The two guides that we will be reporting to are:

1. Internal Guide
2. Internal Co-Guide

- Any progress is to be first reported to the internal co guide.
- Upon approval, reporting is done to the internal guide.
- Any changes or suggestions are made/adhered to.
- The presentation is given to the internal guide and any further changes are made.

5.

Software Requirements Specification

5.1 Introduction

5.1.1 Purpose

The purpose of the project is to understand the opinions or sentiments of people through twitter data by applying dictionary based sentiment analysis approach,where one can understand the difference between opinions of different people and their views related to a particular thing or product

5.1.2 Project Scope

- This project is proposed to analyse the sentiments of people by classifying into positive, negative and neutral based on polarity using dictionary based approach of sentiment analysis.
- At first, twitter datasets are downloaded using flume-agent. Necessary fields are taken for further processing using MapReduce
- In the next step, the structured input data is given to mapper function In which, each tweet is split into individual tokens.
- It also assings polarity to each token based on positive,negative and neutral dictionary.
- Individual tokens which are having certain polarities are collected by reducer function and reducer function also calculates the sum of positive,negative and neutral sentiments.
- The proposed system can find most popular information about the people, organizations and can be used in the fields of analytics
- Our proposed system is not analyzing real time tweets and emoticons

5.2 System Features

At first we collected the twitter data Using API streaming of tokens and apache flume.During collection of data we passed keyword of required topic into config file of flume, then flume gets started collecting tweets related to that keyword. Tweets of different users are moved into Hadoop File system

In the next step using HIVE we apply hive queries on unstructured data which converts it into structured data ,then we create different tables based on different IDs of tweets.Then we apply MapReduce function on structured data which includes following functions:

1. Tokenization:All the words in a tweet are broken down into tokens. For example, '@username I had an amazing time today!' is broken down into individual tokens such as '@username', 'I', had, 'an', 'amazing', 'time', 'today'.

2. Comparing with dictionary:We have created dictionaries of positive and negative words. Tokens produced in tokenization are compared with positive and negative dictionaries and based on that different polarities are assigned to words

3. Classification:Reducer function calculates sum of words of different polarities and classify them into positive sentiments if polarity is 1 and negative sentiments if polarity is -1 At the end we are displaying result in the form of pie charts using R programming

5.3 Nonfunctional Requirements

5.3.1 Performance Requirements

- As for this prototype version we will keep on detecting if the system crashed, hanged or an operating system error occurred.

- Also detecting the performance of the system in terms of the efficiency of integration of the different components

5.3.2 Safety Requirements

- Unauthorized access to the system and its data is not allowed
- Ensure the integrity of the system from accidental or malicious damage.
- The access permissions for system data may only be changed by the systems data administrator

5.3.3 Security Requirements

There are no specific security requirements, anyone can access and use the portal but only authorized persons who are allowed to use and access the database, web pages and the product engine

5.3.4 Software Quality Attributes

- **Availability**

  Software must keep functioning in spite of problems. Since the 'problems' can be of many types, different technologies should work in tandem to achieve high availability for the overall system.

- **Security**

  Software should remain protected from unauthorized access. This includes both change access and view access.

- **Maintainability**

  Software must be easy to improve, correct any bugs and be proactively fixed through preventive maintenance.

- **Testability**

  Software must get tested thoroughly before release.

- **Reliability**

  High Reliability is the measure of how a product behaves in varying circumstances.

- **Efficiency**

  Software should be efficient enough to do the required processing in least amount of time.

- **Compatibility**

  Compatibility is the ability of the software to work with other systems.

- **Modularity**

  Software must compose of separate, interchangeable components, each of which accomplishes one function and contains everything necessary to accomplish goal. Modularity increases cohesion and makes it easier to maintain the code.

- **Usability**

  Software should able to offer its interfaces in a user friendly and elegant way.

5.4 System Analysis Models

5.4.1 Data Flow Diagram / Use Case Diagram

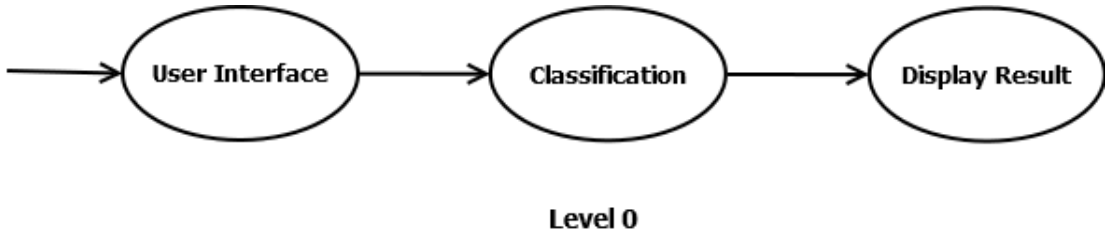


Figure 5.1: Data flow Diagram level 0

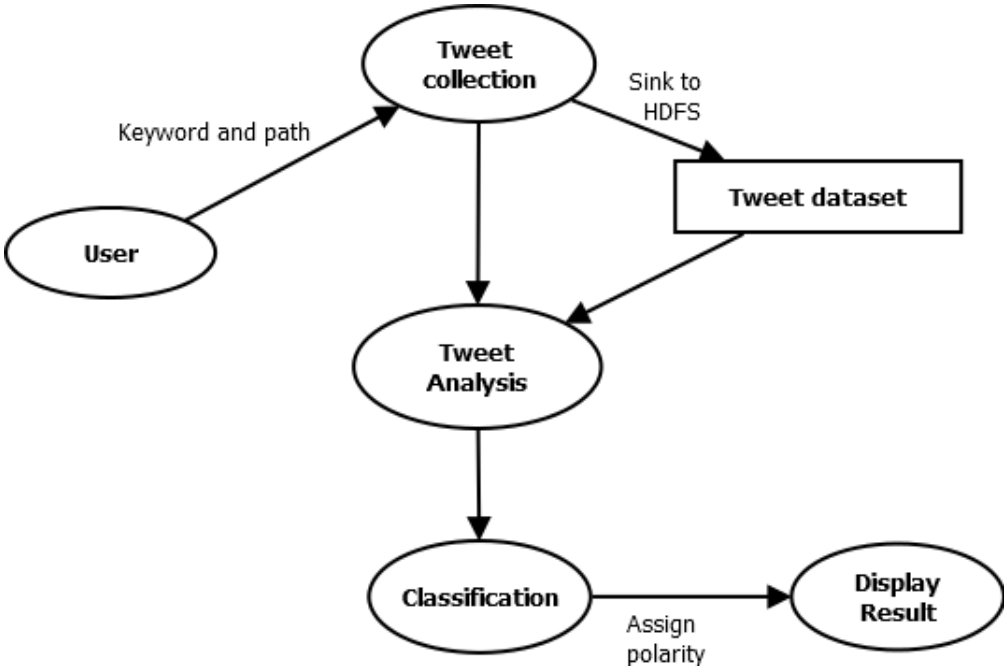


Figure 5.2: Data flow Diagram level 1

5.4.2 Usecase Diagram

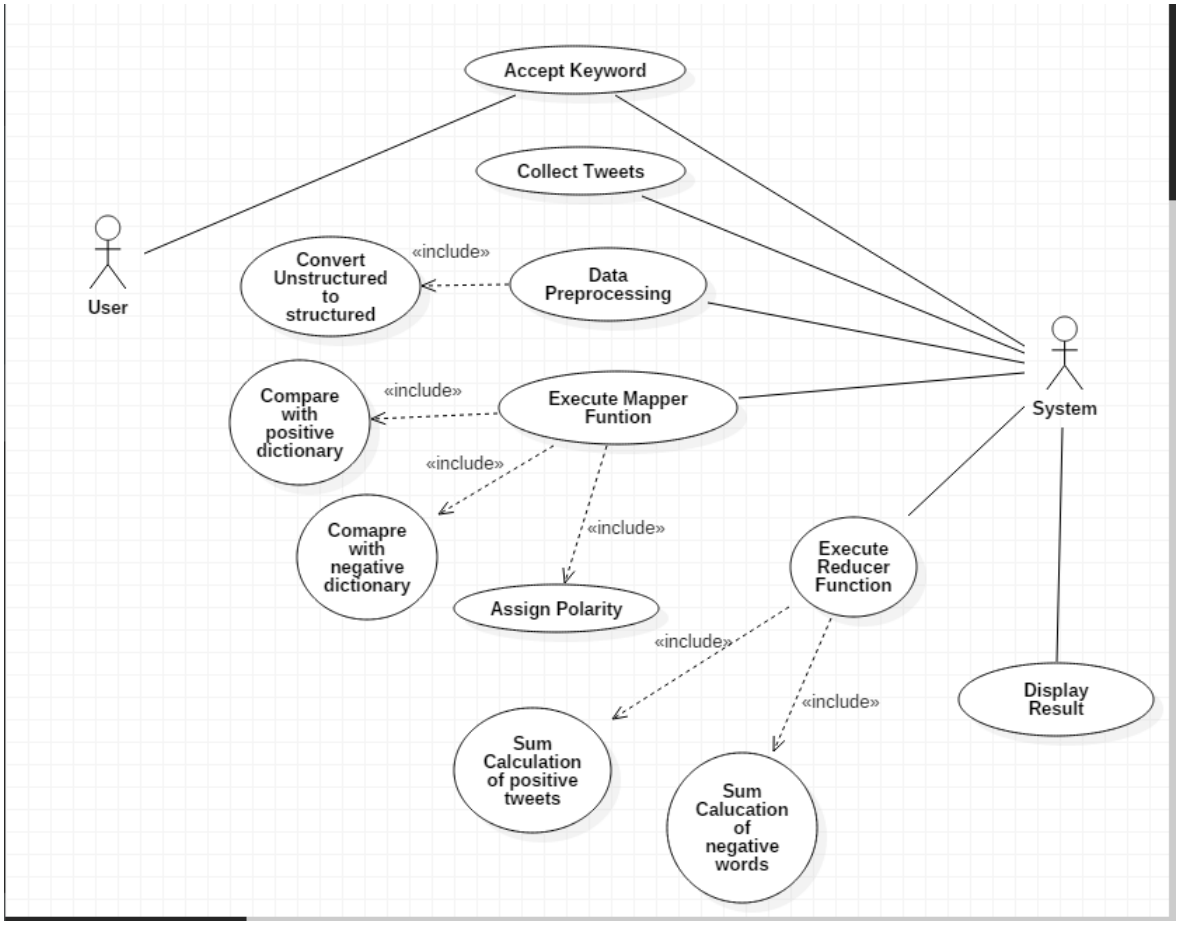


Figure 5.3: Use Case Diagram

5.4.3 Class Diagram

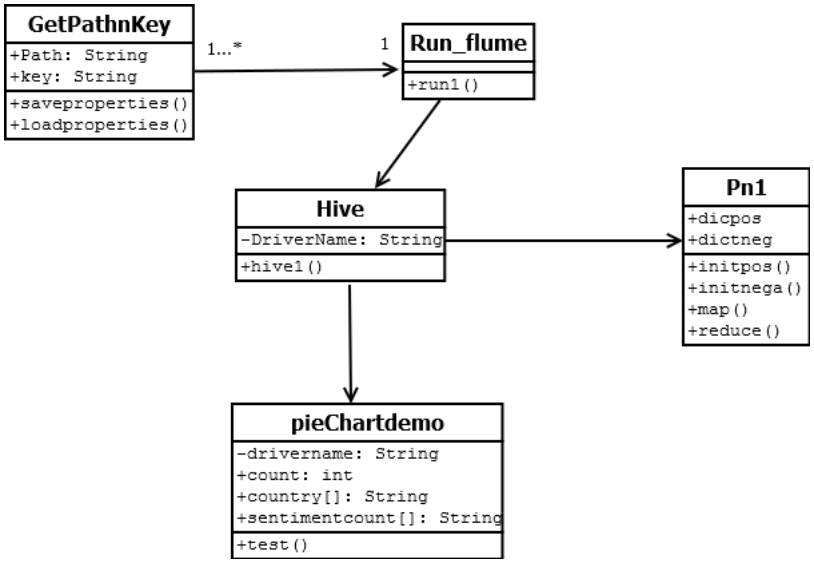


Figure 5.4: Class Diagram

5.4.4 State Transition Diagram

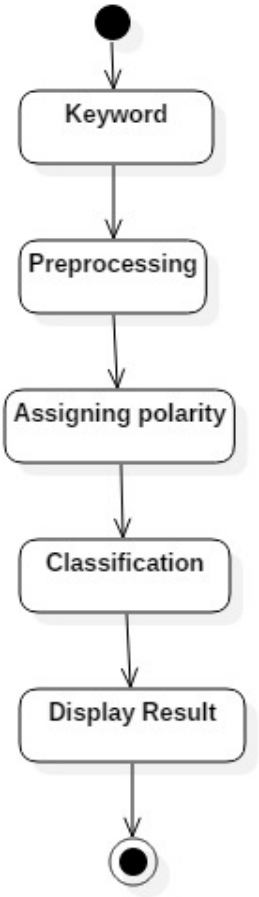


Figure 5.5: State Transition Diagram

6.

Detailed Design

6.1 System Architecture

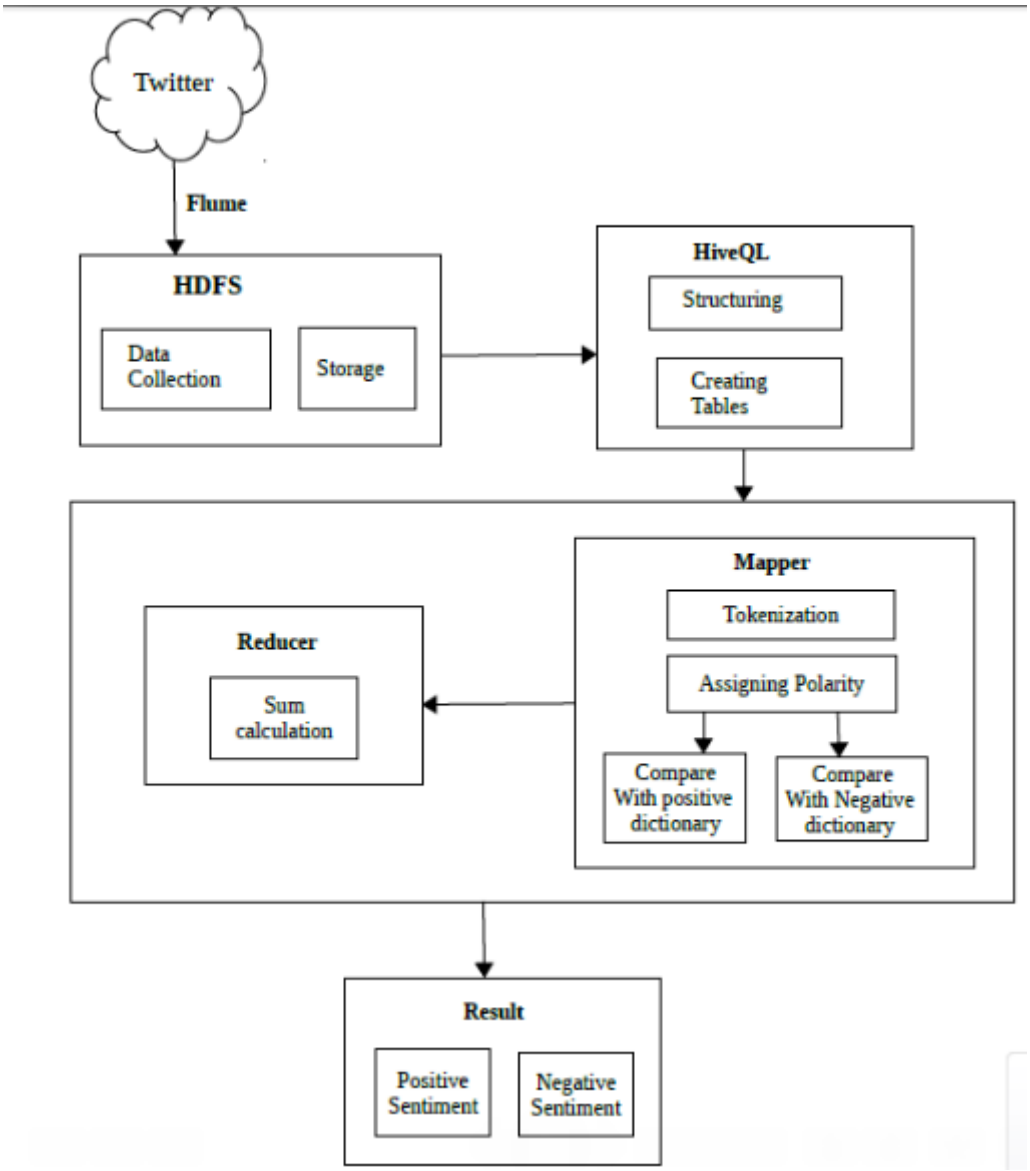


Figure 6.1: System Architecture

1. Twitter :

Twitter is an online social networking site which enables people to share their opinions about any trending topic in the form of short messages which are called tweets. Twitter datasets are freely available and it can be used to extract different sentiments of people about any topic or product and its beneficial to different companies for increasing their market value.

2. Flume : Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

3 Data collection : Data sets are collected from twitter using following steps:

- Creating Twitter application: At first we are creating twitter application to get access to Consumer key, Consumer secret key, Access token key and Access token secret key.These keys are essential to collect data from twitter application.
- Injection of data to HDFS: For downloading datasets we passed keyword of related topic into config file of flume and we also added different token keys into the config file then we run flume agent from command line using command as soon as flume gets configured correctly it starts downloading tweets based on keyword and it injects tweets into hdfs.

4 Hive : Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis.

- Structuring: Datasets we have downloaded from twitter is in unstructured format which cannot be analyzed so we have used hive queries to convert unstructured data into structured data
- Creating tables: From unstructured data we are selecting some fields that is twitterId ,message and locations for creating different tables for analysis and this is accomplished by using HiveQL

5 Mapper :

1. Tokenization: For analysing each tweet we need to break it into individual words which are called tokens and these tokens are compared with dictionary in order to generate sentiments.
2. Assigning Polarity: we have created positive dictionary which contains positive words and negative dictionary which is collection of negative words. that we have generated are compared with positive dictionary if it is a positive word and it is compared with negative dictionary if it is a negative word. Based on positive and negative word polarities are assigned to the tokens.

6.Reducer : Sum Calculation:- It calculates sum of positive and negative words and result is displayed.

6.2 UML Diagrams(Activity, Collaboration, Sequence)

6.2.1 Activity Diagram

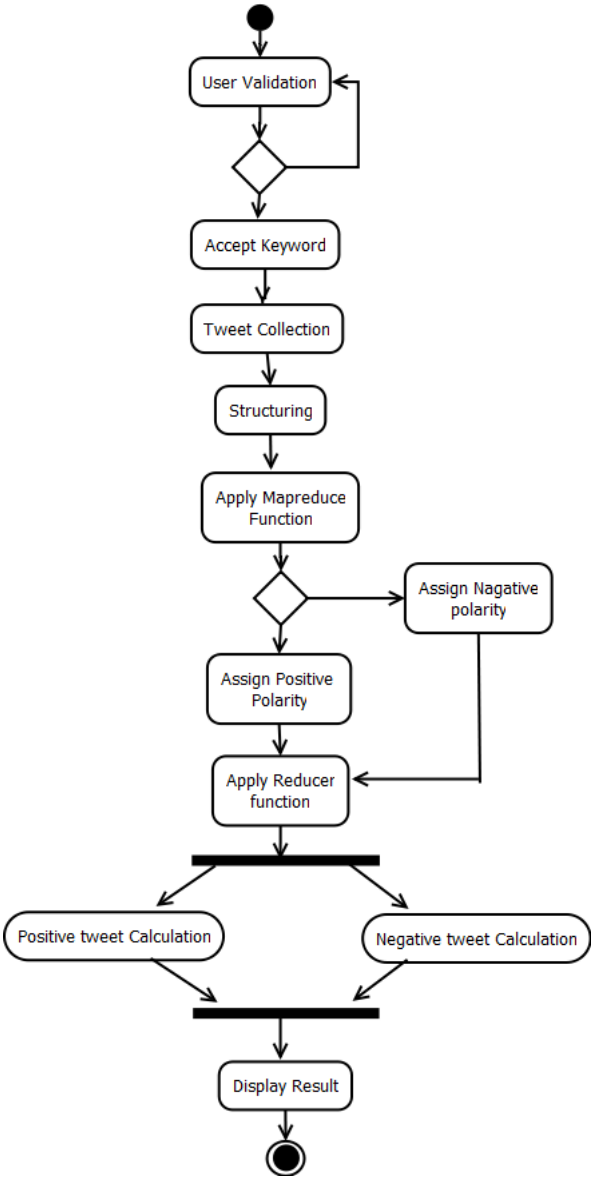


Figure 6.2: Activity Diagram

6.2.2 Collaboration Diagram

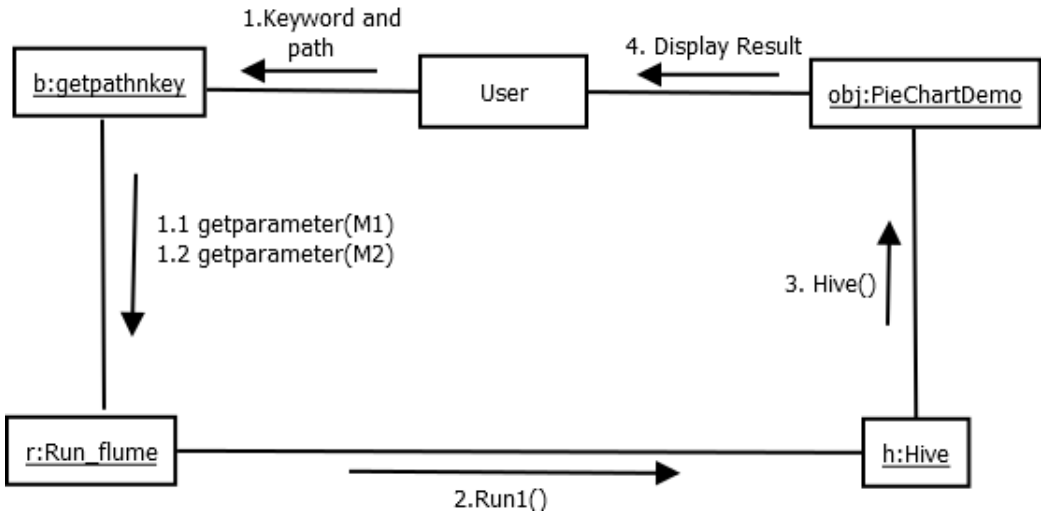


Figure 6.3: Collaboration Diagram

6.2.3 Sequence Diagram

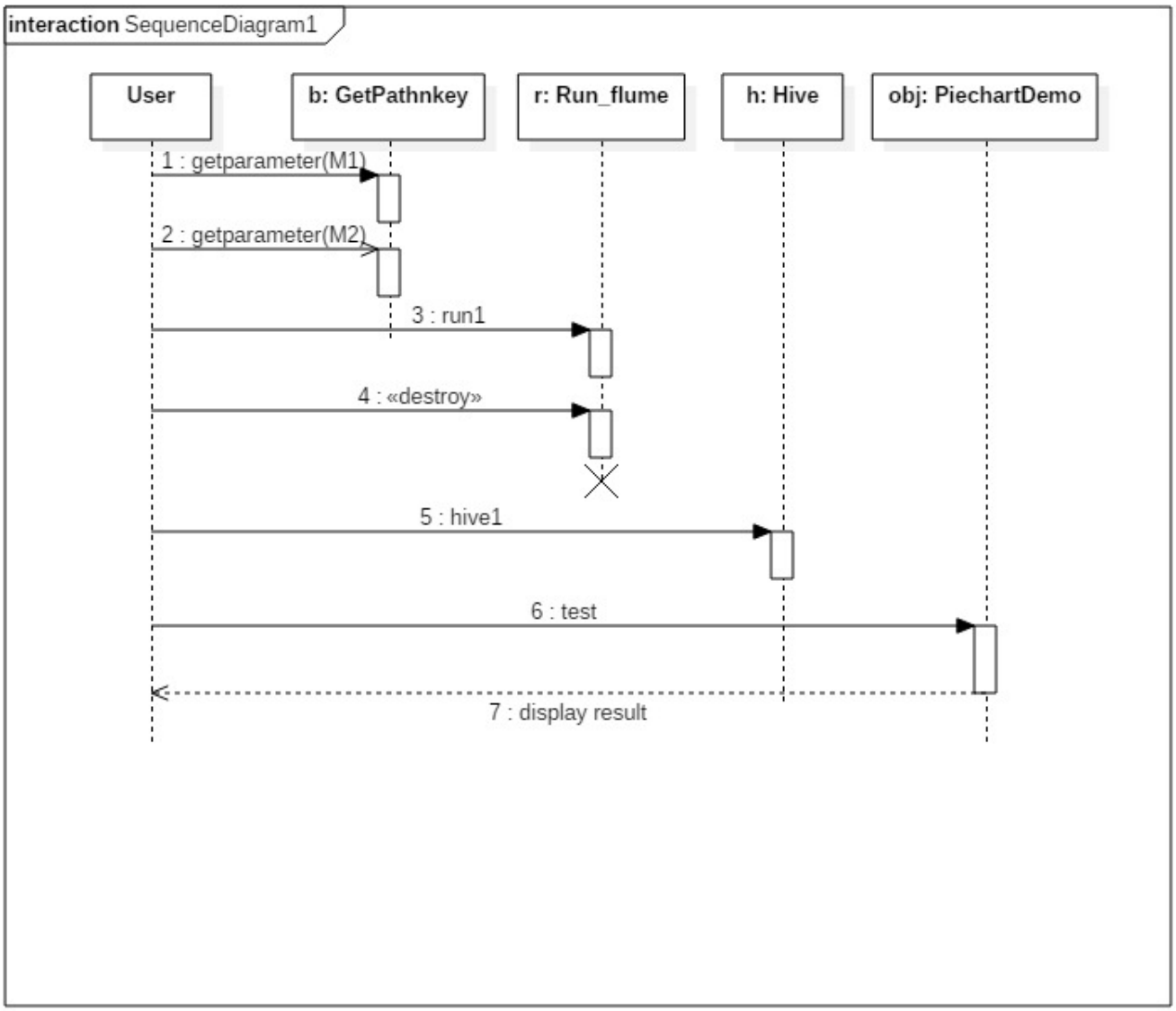


Figure 6.4: Sequence Diagram

6.3 Deployment and Component Diagrams

6.3.1 Component Diagram

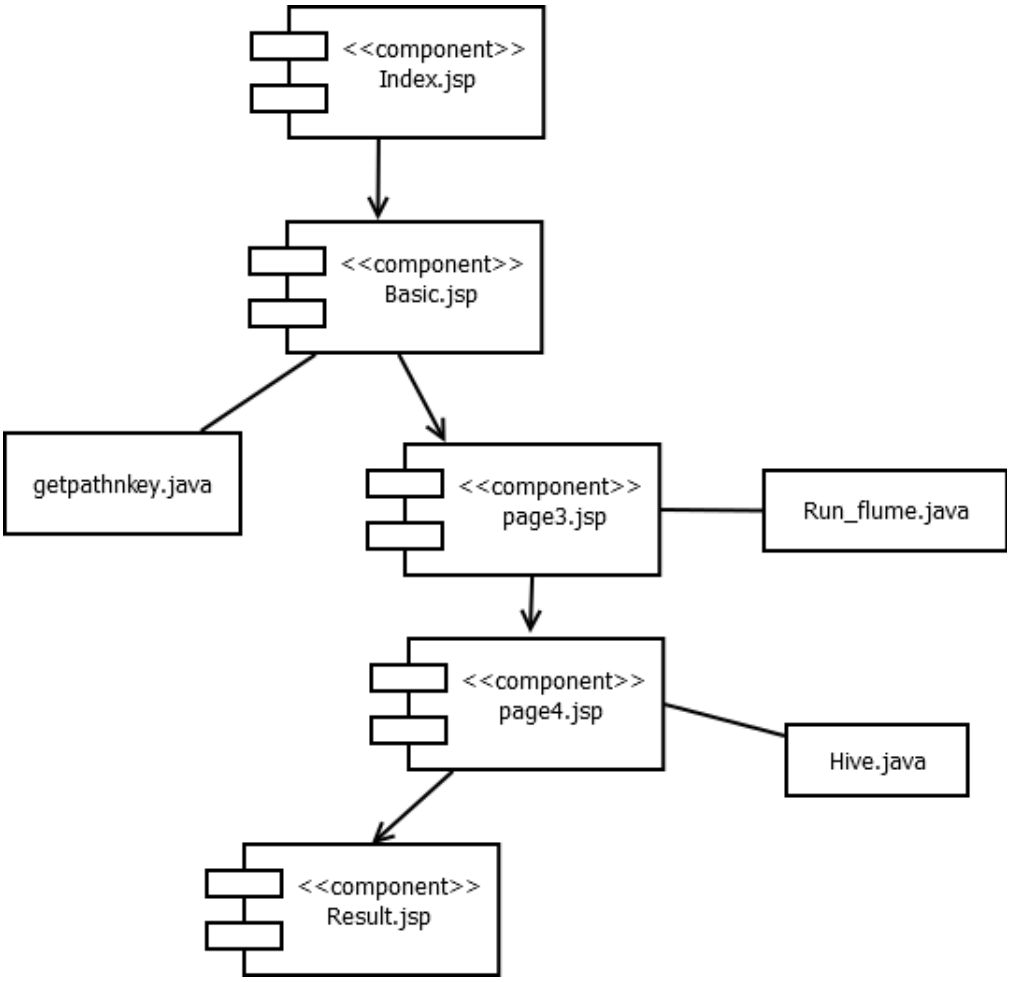


Figure 6.5: Component Diagram

6.3.2 Deployment diagram

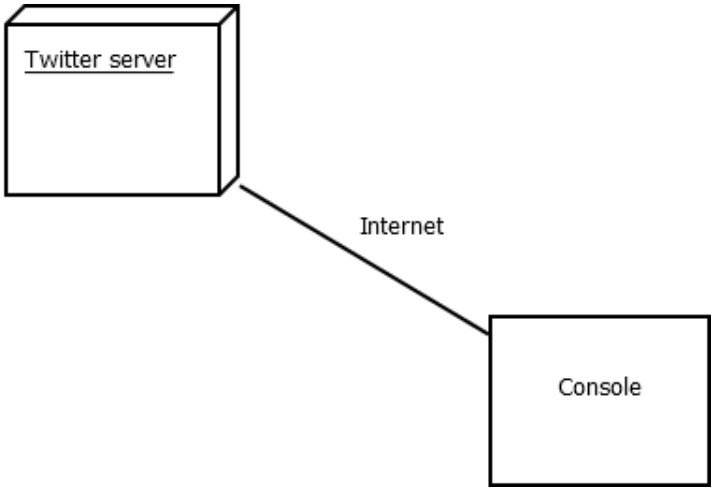


Figure 6.6: Deployment Diagram

7.

Implementation

7.1 Introduction

7.2 Tools and Techniques Used

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 has got two important components for distributed storage and computation:
  1. HDFS (Hadoop Distributed File System)
  2. MapReduce

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 stores schema in a database and processed data into HDFS.
- It provides SQL type language for querying called HiveQL or HQL.

Flume:

- Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events (etc...) from various social networking sites like Facebook and Twitter to a centralized data store.
- In our project we have used flume for collecting data from twitter using streaming API and it also injects data into HDFS.

7.3 Algorithm Implemented

After collecting twitter data we are preprocessing it using using Hive which includes structuring and tokenization. In tokenization we are breaking each tweet into individual tokens and these tokens are stored in mdata array which is of type string.

```
A. Map(key,value,context)
        1.for i from 1 to mdata.length
            a. If key is positive
                    1.compare key with positive dictionary
                    2.Assign value 1
            b. else If key is negative
                    1.Comapare key with negative dictionary
                    2.Assign value -1
            c. else assign value 0
        d. End if
        2. End for
```

This function implements mapping in which it compares each token from mdata with positive and negative dictionary and assigns value accordingly

```
B. Reduce(key,value,context)
        1. Calculate sum of positive words
        2. Calculate sum of negative words
```

7.4 Verification And Validation For Acceptance

- **Verification :**

  Verification helps answering ”Are we building the product right?.

  In design and development, verification concerns the process of examining the result of a given activity to determine conformity with the stated requirement for that activity.

  - Verification determines if something has been built according to specifications.

- Validation determines if something works as intended in the user's environment and meets their needs.

Verification concerns requirements coverage, and if work products (not just code) meet what has been planned and designed (including standards and processes).In our project Geographical prediction of cruelty attacks using data forums , it almost has been working. Goals of our projects meets what we have planned to develop project except alert system.

- **Validation :**

  Validation is trying to answer "Are we building the right product?"

  In design and development, validation concerns the process of examining a product to determine conformity with user needs. Validation is normally performed on the final product under defined operating conditions. It may be necessary in earlier stages. Validation ensures that products meet users expectations (that may not be exactly what is defined in requirement specification). Therefore, validation usually requires the final user involvement and production environments assessments. In our project Geographical prediction of cruelty attacks using data forums , GUI made by us is simple yet intuitive, so that users can easily interact with GUI and can get appropriate result.

8.

Software Testing

8.1 Type of testing Used

We are performing three types of Testing:

- Black-Box Testing
- White-Box Testing
- Usability Testing

1. **Black-box Testing**
    - We have performed black box testing to ensure that whether it correctly accepts inputs and generate expected output
    - Our application accepts keywords only in English language and generate classification of sentiments that is positive and negative sentiments
    - If user enters keywords like integer and special characters then in that case it still generates classification of sentiments, if integers or special characters are there in tweets but 0 polarity will be assigned to integers and special characters
2. **White-Box Testing:**
    - In White-Box testing we check whether all the loops and statements correctly execute
    - Here first there is for loop,if condition under for loop is true then it executes if statement
    - Here we check based on condition whethet it correctly comapres words with positive and negative dictionaries
    - After comparison whether correct polarites are assigned to the words based on condition
3. **Usability Testing:**
    - Usability testing is a black-box technique and is used to identify any error(s) and improvements in the software by observing the users through their usage and operation.
    - In our system we have created user friendly GUI which is easy and efficient to use, user needs to enter keyword and based on that keyword tweets are collected and they are processed and result is displayed

8.2 Test Plan

This test plan describes the testing approach and overall framework that will drive the testing of sentiment analysis of twitter data using hadoop

8.2.1 Functions to be tested

1. Tokenization: We perform white box testing to ensure that whether mapper function correctly splits the tweets into individual tokens and these tokens are stored in mdata
2. Assigning polarities: We perform white box testing to ensure that whether mapper function correctly compares the words with positive and negative dictionaries and assign polarites to each word based on certain condition
3. Classification: We perform white box testing to ensure that whether it correctly calculates the sum of positive and negative words

8.2.2 Functions not to be tested

Structuring:- There is no need to perform any test to see whether the unstructured data is converted into structured data correctly Because it is automatically converted into structured data using hiveQL

8.2.3 Test Stratergy

Test Objective

- The objective of the test is to verify that the functionality of Sentiment analysis of Twitter data using Hadoop module works according to the specifications.
- The test will execute and verify the test scripts, identify and fix and retest all high and medium severity defects per the entrance criteria
- The final product of the test is two fold:

1. Twitter Sentiment Analysis application
2. A set of stable test scripts that can be reused for Functional test execution

Test Principles

- Testing will be focused on meeting the business objectives, cost efficiency, and quality.
- There will be common, consistent procedures for all teams supporting testing activities.
- Testing processes will be well defined, yet flexible, with the ability to change as needed.
- Testing activities will build upon previous stages to avoid redundancy or duplication of effort.
- Testing environment and data will emulate a production environment as much as possible.
- Testing will be a repeatable, quantifiable, and measurable activity.
- Testing will be divided into distinct phases, each with clearly defined objectives and goals.
- There will be entrance and exit criteria.

8.2.4 Test approach

1. **Unit Testing:** We are performing testing on two modules mapper and reducer
   - Mapper unit We perform white box testing to ensure that it correctly compares the words with positive and negative dictionaries and assign polarities to words based on the given condition
   - Reducer unit We perform white box testing to ensure that it correctly calculates the sum of positive and negative words
2. **Integration Testing:**
   - In this testing we combine mapper and reducer units and we check whether the tweet is tokenized into split words correctly
   - Whether split words are compared with positive and negative dictionary correctly
   - Whether the correct values have been assigned to the words and finally whether reducer unit correctly calculates the sum of positive and negative words

3. **Black-Box Testing:**

   - We have performed black box testing to ensure that whether it correctly accepts inputs and generate expected output
   - Our application accepts keywords only in English language and generate classification of sentiments that is positive and negative sentiments
   - If user enters keywords like integer and special characters then in that case it still generates classification of sentiments, if integers or special characters are there in tweets but 0 polarity will be assigned to integers and special characters

4. **Usability Testing:**

   - Usability testing is a black-box technique and is used to identify any error(s) and improvements in the software by observing the users through their usage and operation.
   - In our system we have created user friendly GUI which is easy and efficient to use, user needs to enter keyword and based on that keyword tweets are collected and they are processed and result is displayed

8.2.5 Item pass/fail criteria

1. If the keywords entered by user in English language then it produces the same output as expected output so it is passes
2. If the keyword entered by user is integer or special characters then it does not produce the same output as expected so it fails
3. If condition under for loop is true in mapper function then it produces the same output as expected output so it passes
4. If condition under for loop is false in mapper function then it produces the same output as expected output so it passes
5. If condition under if statement is true in mapper function then it produces same output as expected output so it passes
6. If condition under if statement is false in mapper function then it produces same output as expected output so it passes

7. If condition under else if is true in mapper function then it produces same output as expected output so it passes

8. If condition under else if is false in mapper function then it produces same output as expected output so it passes

8.3 Test cases

We have obtained four test cases by performing testing:

- Black-Box Test Cases
- White-Box Test Cases
- Positive Test Cases
- Negative Test Cases

Test Id	Input	Description	Expected Output	Actual Output	Pass/fail
BV1	Keyword containing English word	Keyword entered by user in english language	Classification of sentiments	Generating Sentiments as positive and negative sentiments	Pass .
BV2	Keyword containing only integers	Keyword entered by user containing only integers	Classificatin of sentiments should not be generated	Generating sentiments	Fail
BV3	Keyword containing only special characters	Keyword entered by user containing only special characters	Classification of sentiments should not be generated	Generating Sentiments	fail .

Table 8.1: Black-Box Test cases

Test Id	Input	Description	Expected Output	Actual Output	Pass/fail
EQ1	Condition under for loop is true	For loop will continue executing until all spilt words from mdata are checked	It should go into if statement	Executing if statement	Pass .
EQ2	Condition under for loop is false	if all words in mdata have been checked	It should come out of for loop	coming out of for loop	pass
EQ3	Condition under if is true	Words are compared from mdata to positive dictionary	It should assign polarity	Assigning polarity	Pass .
EQ4	Condition under if is false	If words are not there in positive dictionary	It should go to else if	Executing else if	Pass .
EQ5	Condition under else if is true	Words are compared from mdata to negative dictionary	It should assign polarity	Assigning polarity	Pass .
EQ6	Condition under else if is false	If words are not there in negative dictionary	It should go to else part	Executing else part	Pass .

Table 8.2: White-Box test cases

Test Id	Input	Description	Expected Output	Actual Output	Pass/fail
PO1	Valid input	Keyword entered by user in english language	Classification of sentiments	Sentiments are classified into positive and negative sentiments	Pass .

Table 8.3: Positive Test cases

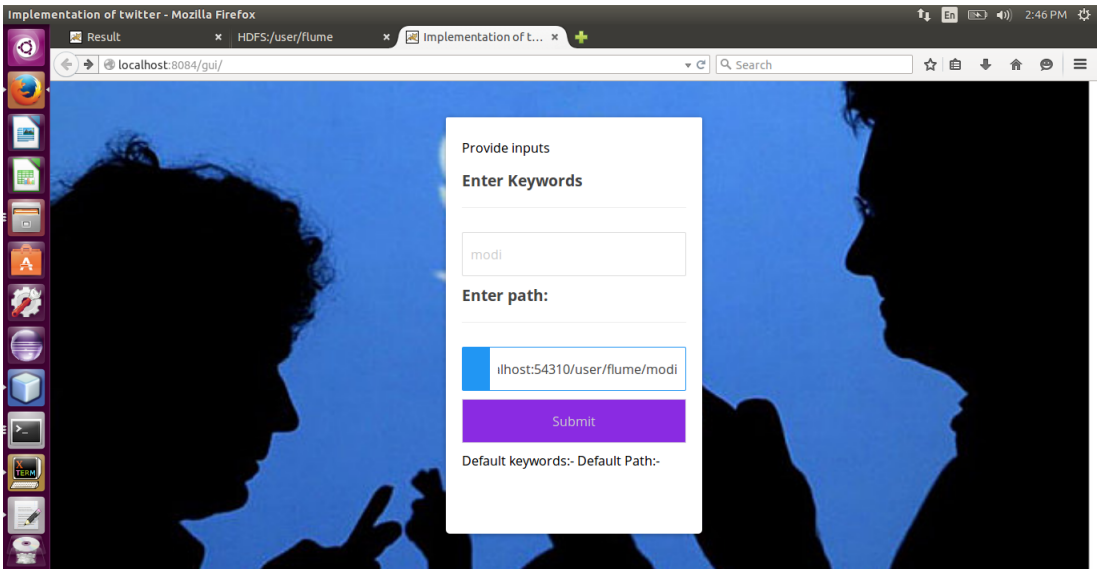
Test Id	**Input**	**Description**	**Expected Output**	**Actual Output**	**Pass/fail**
NO1	Invalid Input	Keyword entered by user are special characters	It should not generate sentiments	Generating Sentiments as positive and negative sentiments	Fail .
NO2	Invalid Input	Keyword entered by user containing only integers	It should not generate sentiments	Generating sentiments	Fail

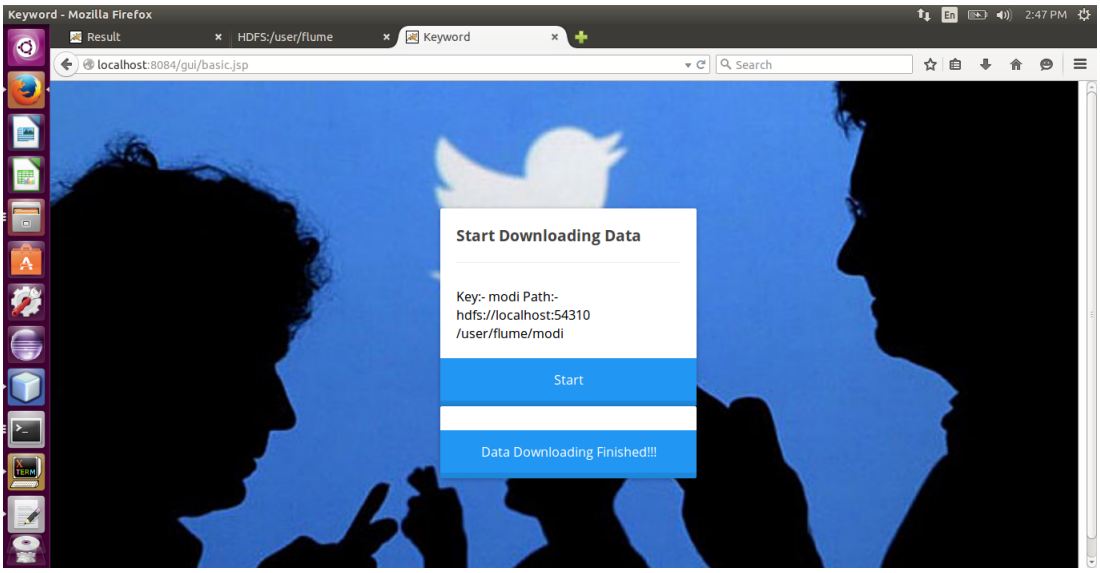
Table 8.4: Negative test cases

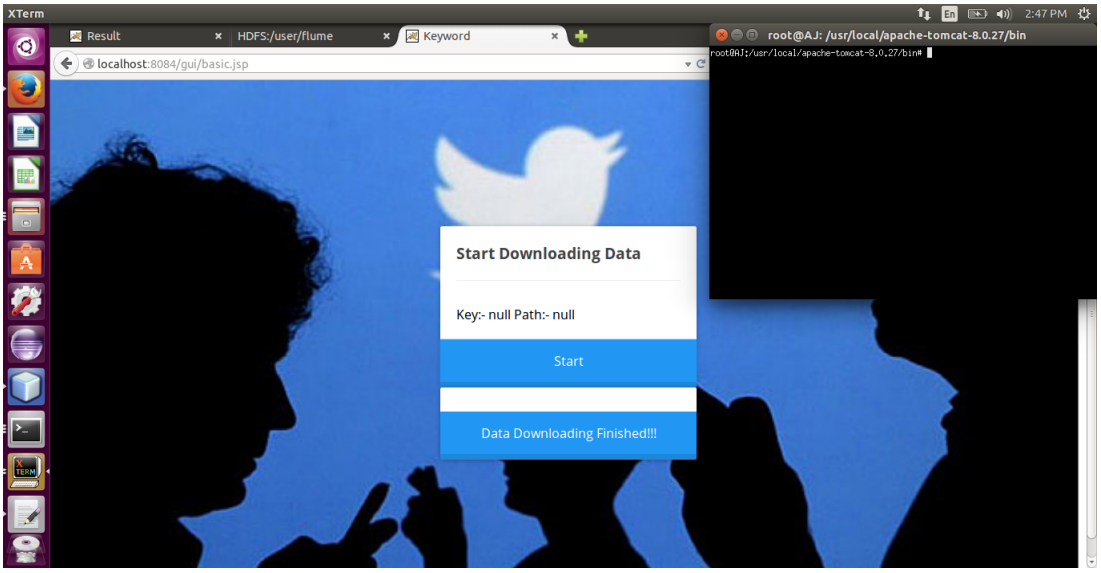
9.

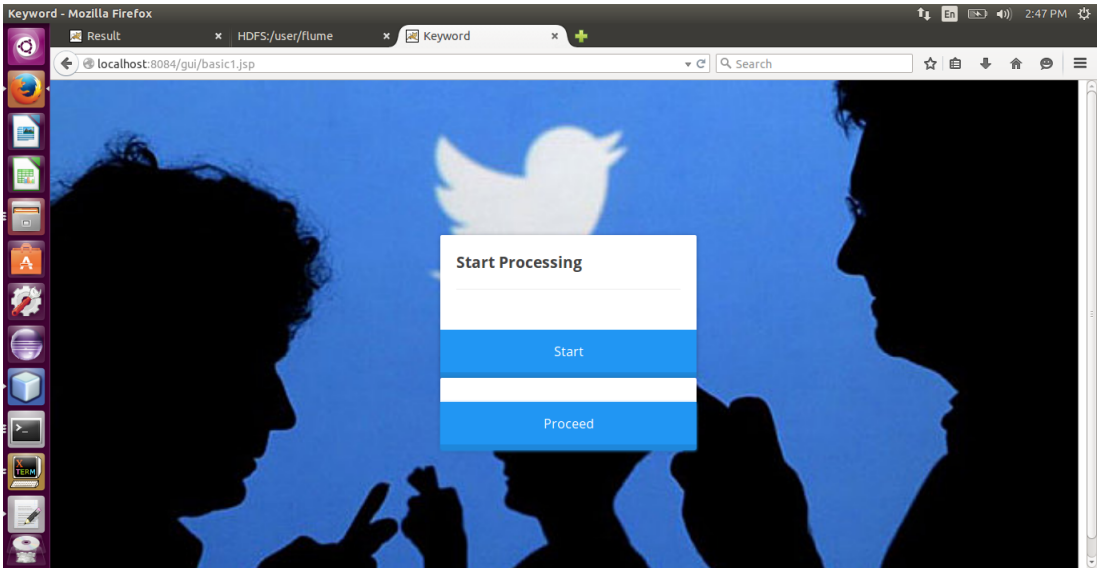
Results

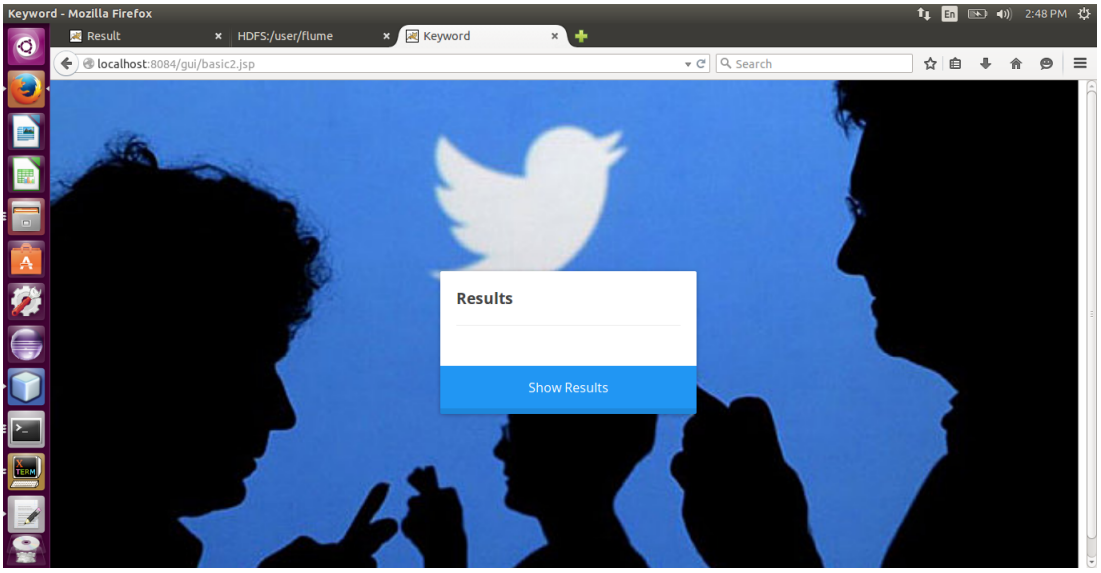
9.1 Screenshots

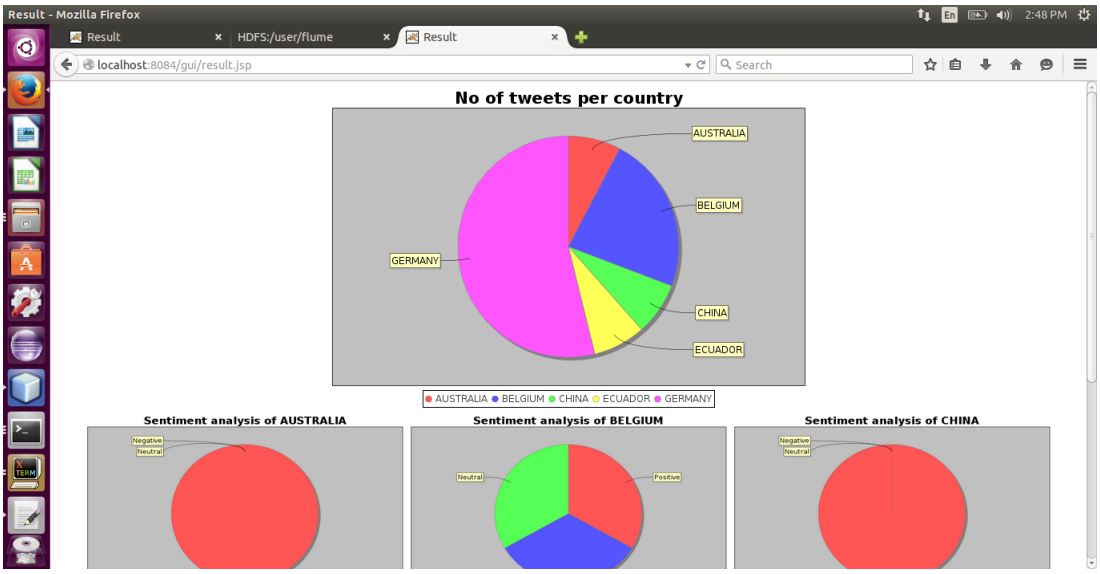


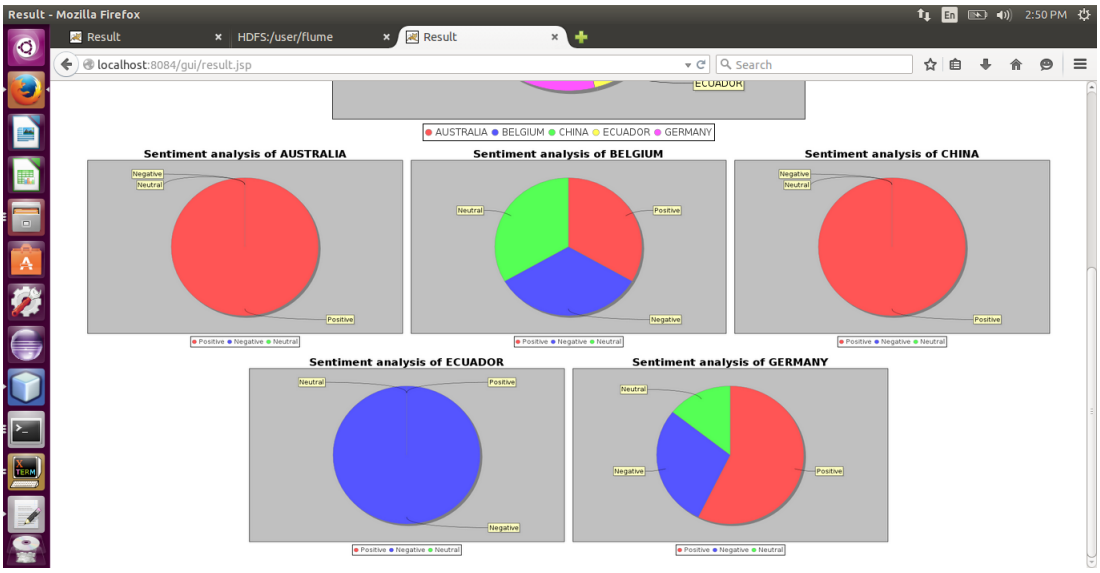



XTerm
Result
HDFS:/user/flume
Keyword
localhost:8084/gui/basic.jsp
root@AJ: /usr/local/apache-tomcat-8.0.27/bin
root@AJ:/usr/local/apache-tomcat-8.0.27/bin#
Start Downloading Data
Key:- null Path:- null
Start
Data Downloading Finished!!!


Keyword - Mozilla Firefox
Result
HDFS:/user/flume
Keyword
localhost:8084/gui/basic1.jsp
Search
Start Processing
Start
Proceed


Keyword - Mozilla Firefox
Result
HDFS:/user/flume
Keyword
localhost:8084/gui/basic2.jsp
Search
Results
Show Results


Result - Mozilla Firefox
2:48 PM
Result
HDFS:/user/flume
Result
localhost:8084/gui/result.jsp
Search
No of tweets per country
AUSTRALIA
BELGIUM
CHINA
ECUADOR
GERMANY
AUSTRALIA BELGIUM CHINA ECUADOR GERMANY
Sentiment analysis of AUSTRALIA
Negative
Neutral
Sentiment analysis of BELGIUM
Neutral
Positive
Sentiment analysis of CHINA
Negative
Neutral


Result - Mozilla Firefox
2:50 PM
Result
HDFS:/user/flume
Result
localhost:8084/gui/result.jsp
Search
ECUADOR
AUSTRALIA BELGIUM CHINA ECUADOR GERMANY
Sentiment analysis of AUSTRALIA
Negative
Neutral
Positive
Positive Negative Neutral
Sentiment analysis of BELGIUM
Neutral
Positive
Negative
Positive Negative Neutral
Sentiment analysis of CHINA
Negative
Neutral
Positive
Positive Negative Neutral
Sentiment analysis of ECUADOR
Neutral
Positive
Negative
Positive Negative Neutral
Sentiment analysis of GERMANY
Neutral
Negative
Positive
Positive Negative Neutral

10.

Deployment and Maintenance

10.1 Installation Steps:

Installation steps of Hadoop 2.6.3:

- Install latest version of Java
- Install ssh
- Create and Setup SSH Certificates
- Install latest version of hadoop
- Setup Configuration Files

The following files will have to be modified to complete the Hadoop setup:

- /.bashrc
- /usr/local/hadoop/etc/hadoop/hadoop-env.sh
- /usr/local/hadoop/etc/hadoop/core-site.xml
- /usr/local/hadoop/etc/hadoop/mapred-site.xml.template
- /usr/local/hadoop/etc/hadoop/hdfs-site.xml

1. Format the New Hadoop Filesystem
2. After making all changes in respective files and after switiching to respective hadoop user as a root user start hadoop with following command on command prompt. start-all.sh or (start-dfs.sh and start-yarn.sh)
3. To check whether hadoop framework is running or not use following command: jps
4. If its running correctly then it gives following output: hduser@laptop:/usr/local/hadoop/sbin jps
   9026 NodeManager
   7348 NameNode
   9766 Jps

8887 ResourceManager

7507 DataNode

5. To stop hadoop framework :stop-all.sh

Installion Steps of hive 1.2.1:

- Install latest version of hive
- Set Hive envirnoment variables
- Set HADOOP PATH in hive confg.sh
- Create hive directories within HDFS
- set READ or WRITE permission for table
- Make respective changes in hive configuration files
- launch hive using command hive.
- Now we can create databases,tables and work on those by running different hive queries.

Installation steps of Flume 1.6:

PART 1 :

- Download latest version of flume
- Extract the tar file and store it to respective directory.
- Remove unnecessary jar files from lib directory of bin
- download flume-sources-1.0-SNAPSHOTS.jar from following link
- https://drive.google.com/file/d/0B-Cl0IfLnRozUHcyNDBJWnNxdHc/view usp=sharing
- Move the flume-sources-1.0-SNAPSHOT.jar file from Downloads directory to lib directory of apache flume

- Edit flume-env.sh file and set java HOME and FLUME CLASSPATH.

PART 2 :

1. login into twitter app.
2. Create new application and enter all the details in the application:
3. Get following Keys:
    - Access Tokens
    - Access Token Secret
    - Consumer Key
    - Consumer Secret key
4. download flume.conf file using following link: https://drive.google.com/file/d/0B-Cl0IfLnRozdlRuN3pPW usp=sharing
5. Edit flume.conf
6. Add Consumer Key, Consumer Secret, Access Token, Access Token Secret ,keyword into flume.conf file.
7. Use following command to strat flume so as to fetch data from twitter app. ./bin/flume-ng agent -n TwitterAgent -c conf -f /usr/lib/apache-flume-1.4.0-bin/conf/flume.conf

Installation steps for Netbeans IDE:

1. Download the installer using following links netbeans.org/downloads/.
2. Give executable permission.
3. Start the installer:
4. Netbeans IDE installed and now we can create projects.

11.

Future Scope

At this moment, the code can handle the analysis part with a very good accuracy. But there are a few areas which have a lot of scope in this aspect. Sarcastic comments are the ones which are very difficult to identify. Tweets containing sarcastic comments give exactly opposite results owing to the mindset of the author. These are almost impossible to track.

Also depending on the context in which a word is used, the interpretation changes. For ex: the word unpredictable in unpredictable plot in context of a landplot is negative whereas unpredictable plot in context of a movies plot is positive.

So its important to relate the interpretation with the context of the tweets. Also the use of native language combined with English usage is difficult to interpret. We can also analyze emoticons and images with twitter sentiment analysis

12.

Conclusion

12.1 Conclusion

This project will give us hands on experience of handling and parallel processing of huge amount of data. Data collection process will introduce us to Java twitter streaming API. We will get exposure to work with prominent parallel data processing tool: Hadoop.

Apache Hadoop framework is gaining significant momentum from both industry and academia as the volume of data to analyze growth rapidly. This project will help us not only to gain knowledge about installation and configuration of hadoop distributed file system but also map reduce programming model. Amongst the many fields of analysis, there is one field where humans have dominated the machines more than any the ability to analyze sentiment, or sentiment analysis.

The future of this data analysis field is vast. This project not only analyses the sentiments of the user but also computes other results like the user with maximum friends/followers, top tweets etc. hence hadoop can also be effectively used to compute such results in order to determine the current trends with respect to particular topics. This can be very useful in the marketing sector.

References

[1] *SentiView: Sentiment Analysis and Visualization for Internet Popular Topics*; IEEE Transactions On Human-Machine Systems, Vol. 43, No. 6, November 2013,Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang, SentiView:

[2] *Twitter sentiment analysis: The good and bad the omg*; Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media Rushabh Mehta, Dhaval Mehta, Disha Chheda, Charmi Shah and Pramila M. Chawan

[3] *Sentiment Analysis and Influence Tracking using Twitter*;in International Journal of Advanced Research in Computer Science and Electronics Engineering, Vol 1, Isuue 2, May 2012

[4] *Sentiment Analysis of Twitter Data*;Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau

[5] *TwitterRank, Finding Topic-sensitive Influential Twitterers*; WSDM'10, February 46,2010, New York City, New York, USA Copyright 2010 ACM

[6] *Apache Hadoop Available*; wiki.apache.org

[7] *Facebook Lexicon*; www.facebook.com, lexicon

[8] *Hadoop MapReduce tutorial at*; hadoop.apache .org

[9] *Hadoop pig available at*; Hadoop.apache.org,R. Chaiken, et. al. Scope: Easy and Efficient Parallel Processing

[10] *Hadoop hdfs user guide at*;hadoop.apache.org

Appendices

APPENDIX A

A.1 Algorithm Implemented

After collecting twitter data we are preprocessing it using using Hive which includes structuring and tokenization. In tokenization we are breaking each tweet into individual tokens and these tokens are stored in mdata array which is of type string.

```
A. Map(key,value,context)
        1.for i from 1 to mdata.length
            a. If key is positive
                    1.compare key with positive dictionary
                    2.Assign value 1
            b. else If key is negative
                    1.Comapare key with negative dictionary
                    2.Assign value −1
            c. else assign value 0
        d. End if
        2. End for
```

This function implements mapping in which it compares each token from mdata with positive and negative dictionary and assigns value accordingly

```
B. Reduce(key,value,context)
        1. Calculate sum of positive words
        2. Calculate sum of negative words
```

A.2 Mathematical Model

Let 'S' be the system defined as:-

$S = \{I, A, F, O\}$

1) 'I' is the set of input:-

$I = \{I_1, I_2, I_3\}$

I_1 = keywords

I_2 = path

I_3 = datasets

2) 'A' is the set of algorithms:-

$A = \{A_1\}$

Where A_1 is map-reduce algorithm

3) 'F' is the set of functions:-

$F = \{F_m, F_r\}$

Where F_m is mapper function and F_r is reducer function

Let $S_1 = \{I_3 \cap A_1\}$
Where I_3 is datasets
A_1 is an algorithm used for map-reduce operation

Functions:-

1) Let 'A' is a set of positive words
$f(A) = \{f(a) \mid a \in A\}$
Where 'a' is a token

2) Let 'B' is a set of negative words
$f(B) = \{f(b) \mid b \in B\}$
Where 'b' is a token

3) Let 'C' is a set of words

f(C) = {f(c) | c ∉ A,B }

Where 'c' is a token

Where f(A) , f(B) , f(C) are subsets of F_m

Let 'P' is a set

f(P) = { $\sum_{i=1}^{n} ai$ }

Let 'N' is a set

f(N) = { $\sum_{i=1}^{n} bi$ }

Let 'K' is a set

f(K) = { $\sum_{i=1}^{n} ci$ }

f(P) , f(N) , f(K) are subsets of F_r

4) Let 'O' is the set of output:-

O = {O_1, O_2}

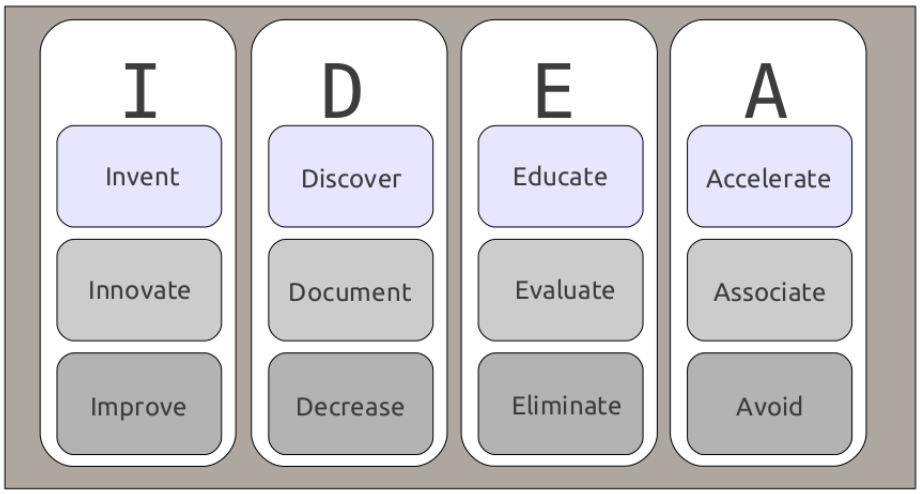
O_1 is pie chart of tweets per country

O_2 is pie chart of sentiments of 5 countries

APPENDIX B

B.1 Assignment 1

Problem Statement: To develop the problem under consideration and justify feasibility using concepts of knowledge canvas and IDEA Matrix.



1. Invent, Innovate, Improve:

   - Invention leads to newer ideas, innovation leads to the effective development of those ideas and improvement is the key aspect of success of every single product.
   - Our idea is to invent a system that can reduce the cruelty attacks in any place. Looking a the present scenarios, it is very important to be cautious about crimes. Increase in the usage of social media has been a boon and a curse for the world. We are developing a system that will use the data from these social media websites and help to reduce the occurrence of crimes, cruelty attacks, riots, etc.

- The innovativeness in our system is that we want to find how the views (whether good or bad) about a particular event are distributed geographically. This will lead us to calculate if there is a chance of a riot and accordingly alert authorities to increase safety.
- There are many other systems that perform sentiment analysis and detect trends in behavior and moods of the people. We want to use those techniques, improve them wherever necessary and make use of those trends.

2. Discover, Document, Decrease:

   - Initially the idea was to do something about the crimes due to social media. Later on we discovered the methods to reduce them. More and more opportunities are discovered during the course of design of the software.
   - Documentation is necessary for the systematic development of any software. Documentation can help several people in different sections, offices or even countries communicate regarding the development of the system. We include various UML diagrams, SRS report to show the necessary changes taking place during the various stages various stages of system development.
   - We intend to decrease the irrelevant work to make the development more effort and cost-effective. We also study the various alternate methods so as to use the one which suits our requirements thereby decreasing the unnecessary work.

3. Educate, Evaluate, Eliminate:

   - Education is very necessary. At every stage we need to finalize which method to use for that we need to be educated about all the possible options. Moreover someone who knows more than the others needs to spread knowledge so that all the members need to be on the same page.
   - At every stage we need to evaluate all the options and the various conditions. Our project involves a lot of safety as well as security parameters, hence utmost care needs to be taken not to violate those measures. Taking into consideration all the parameters we need to evaluate the methods.
   - During the design phase we consider our system to have all the possible features we can think of. After the feasibility study we need to eliminate the redundant features so as to reduce the complexity for the user thereby making it more usable. By eliminating certain features we are trying to make the system more user-friendly so as to increase the scope of the system.

4. Accelerate, Associate, Avoid :

   - Acceleration, that is the process of speeding up, as soon as our design was ready we can focus on the implementation due accelerate the software development.
   - Our system uses association of domains like data mining and its sub-domains like sentiment analysis and trend analysis.
   - The main idea behind the system is to develop a system that protects the people from cruelty attack that can happen regarding certain comments on on-line forums. We want to avoid the crime scenes as much as we can using this system. This includes decision making and alerting the officer in-charge(user).

B.2 Assignment 2

Problem Statement: Perform feasibility assessment of the problem statement using NP- Hard, NP-Complete or satisfiability issues using modern algebra and/or relevant mathematical models.

Theory :

P Problems : Problems that can be solved in polynomial time. The polynomial time is the time expressed in terms of polynomial.

NP Problems : NP stands for "Non-deterministic polynomial time". Problems that can't be solved in polynomial time are nothing but NP problems.

NP-Hard : NP-hard is nothing but non deterministic polynomial hard. In computational complexity theory, NP-Hard is a class of problems that are, informally, "at least as hard as the hardest problems in NP". More precisely, a problem H is NP-hard when every problem L in NP can be reduced in polynomial time to H. As a consequence, finding a polynomial algorithm to solve any NP-hard problem would give polynomial algorithms for all the problems in NP, which is unlikely as many of them are considered hard.

We say that a decision problem Pi is NP-hard if every problem in NP is polynomial time reducible to Pi.
In symbols,
Pi is NP-hard if, for every Pj belongs to NP,
Note that this doesnt require Pi to be in NP.
Highly informally, it means that Pi is as hard as all the problems in NP.
If Pi can be solved in polynomial-time, then so can all problems in NP.
Equivalently, if any problem in NP is ever proved intractable, then Pi must also be intractable.

NP-complete: NP-complete problem is a class of problems which contains the hardest problems in NP. Each NP-complete problem has to be in set of NP problems.

We say that a decision problem Pi is NP-complete if

- it is NP-hard and
- it is also in the class NP itself.

In symbols, Pi is NP-complete if Pi is NP-hard and Pi belongs to NP Highly informally, it means that Pi is one of the hardest problems in NP.

A problem is said to be Non-deterministically Polynomial (NP) if its solution comes from a finite set of possibilities, and it takes polynomial time to verify the correctness of a candidate solution. In case of Twitter Sentiment analysis using Hadoop, there may be following conditions to determine whether the problem is NP hard or NP complete as follows : -

1. If wrong keyword as integers or special character is given to the system then system will collect tweets containing these integers and special character and classification sentiments will be generated and displayed
2. If keyword containing english word is given then then system will successfully generate classification of sentiments

Therefore the problem may be defined as NP - Complete problem by considering above cases. The system will never enter into halting state and will generate the output in any condition, so the system is ”NP Complete” problem.

Mathematical Model:

Let 'S' be the system defined as:-

$S = \{I, A, F, O\}$

1) 'I' is the set of input:-

$I = \{I_1, I_2, I_3\}$

I_1 = keywords

I_2 = path

I_3 = datasets

2) 'A' is the set of algorithms:-

$A = \{A_1\}$

Where A_1 is map-reduce algorithm

3) 'F' is the set of functions:-

$F = \{F_m, F_r\}$

Where F_m is mapper function and F_r is reducer function

Let $S_1 = \{I_3 \cap A_1\}$
Where I_3 is datasets
A_1 is an algorithm used for map-reduce operation

Functions:-

1) Let 'A' is a set of positive words
$f(A) = \{f(a) \mid a \in A\}$
Where 'a' is a token

2) Let 'B' is a set of negative words
$f(B) = \{f(b) \mid b \in B\}$
Where 'b' is a token

3) Let 'C' is a set of words

f(C) = {f(c) | c ∉ A,B }

Where 'c' is a token

Where f(A) , f(B) , f(C) are subsets of F_m

Let 'P' is a set

f(P) = { $\sum_{i=1}^{n} ai$ }

Let 'N' is a set

f(N) = { $\sum_{i=1}^{n} bi$ }

Let 'K' is a set

f(K) = { $\sum_{i=1}^{n} ci$ }

f(P) , f(N) , f(K) are subsets of F_r

4) Let 'O' is the set of output:-

O = {O_1, O_2}

O_1 is pie chart of tweets per country

O_2 is pie chart of sentiments of 5 countries

B.3 Assignment 3

Problem Statement : Use of divide and conquer strategies to exploit distributed/parallel/concurrent processing of the above to identify objects, morphisms, overloading in functions (if any), and functional relations and any other dependencies (as per requirements).

Divide And Conquer Strategy : In computer science, divide and conquer (D and C) is an al-

gorithm design paradigm based on multibranched recursion. A divide and conquer algorithm works by recursively breaking down a problem into two or more sub-problems of the same (or related) type (divide), until these become simple enough to be solved directly (conquer). The solutions to the sub-problems are then combined to give a solution to the original problem.

Diagram explaining divide and conquer :

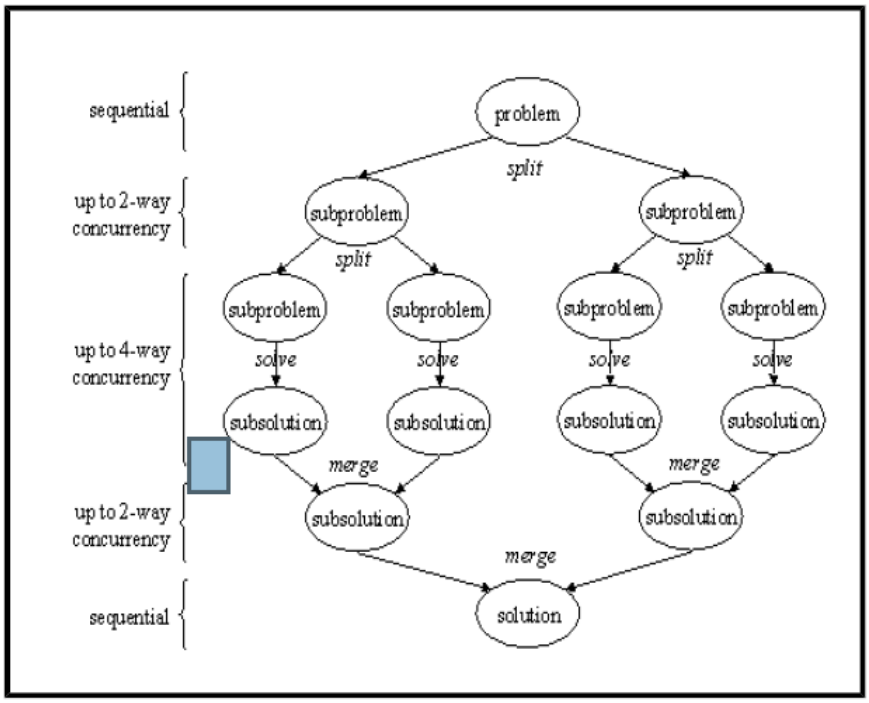


Figure B.1: Diagram explaining divide and conquer

Need Of Divide And Conquer In Sentiment Analysis:- In sentiment analysis , we often try to solve a problem in a general fashion and in many cases based on simplistic views. In the context of aspect extraction and aspect sentiment classification it is not always sentiment word and aspect word pairs are important. The real world is more complex than that,Hence we need divide and conquer strategy in sentiment analysis

Divide And Conquer Used In Project:- The main problem of classification of sentiments is done using divide and conquer technique. In our project, classification of sentiment data is done. Data is collected using Twitter Streaming API.

Main Problem Using Divide And Conquer:- Instead of doing classification in simplistic

manner, data dictionary of polarities is built. Polarities are divided into positive, negative and neutral data.Here the Twitter data is compared with this dictionary where classification is done and we get classified data as an output.We can divide the problem of classification in following way:

1. We can form three types of polarity dictionaries.
2. First will be the dictionary containing positive words. Second will be the dictionary containing negative words. And third will be the dictionary containing neutral words.
3. In first step we can separate out positive words, in second step we can separate out negative words and in third step we can separate out neutral words. Likewise, problem is divided into 3 steps of data classification..
4. In fourth step, we can combine the results of first three steps where we can get the whole output as classified data.

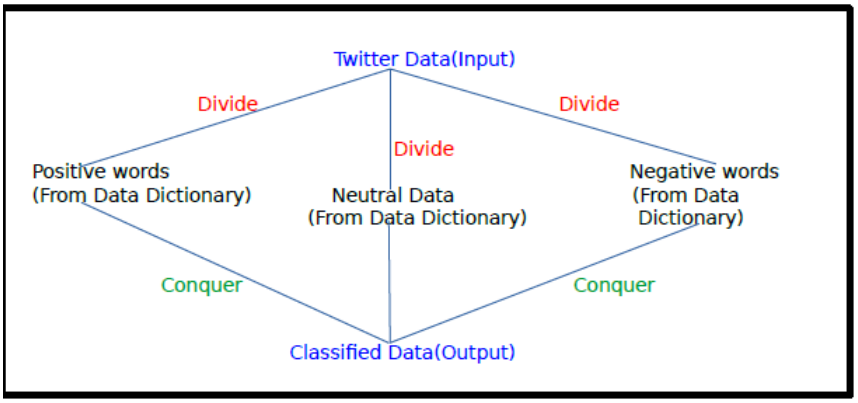


Figure B.2: Functional Dependies of different components

Functional Dependencies Of Different Components: The whole project is implemented in Hadoop Framework.Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of commodity computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each providing computation and storage.

Components of Hadoop:-

1. HDFS(storage)
2. MapReduce (processing)

3. Apache Hive

HDFS (storage) and MapReduce (processing) are the two core components of Apache Hadoop. The most important aspect of Hadoop is that both HDFS and MapReduce are designed with each other in mind and each are co-deployed such that there is a single cluster and thus provides the ability to move computation to the data not the other way around. Thus, the storage system is not physically separate from a processing system

Hive is a data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems.It provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. Hive also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL

In our project,

- Twitter data sets are stored in HDFS.
- Twitter data sets are huge in amount where MapReduce is used to process those data sets.
- It is done with the help of Hive Tool. Hive provides SQL like interface. Where queries can be designed to mold the data according to our need.
- In our project, the main problem of classification is solved with the help of Hive.
- Where the queries are designed such that, the processing is done on data sets where we get classified data as output. Classified data is in the form of positive, negative and neutral tweets.

B.4 Assignment 4

UML diagrams using appropriate tool: Usecase diagram:

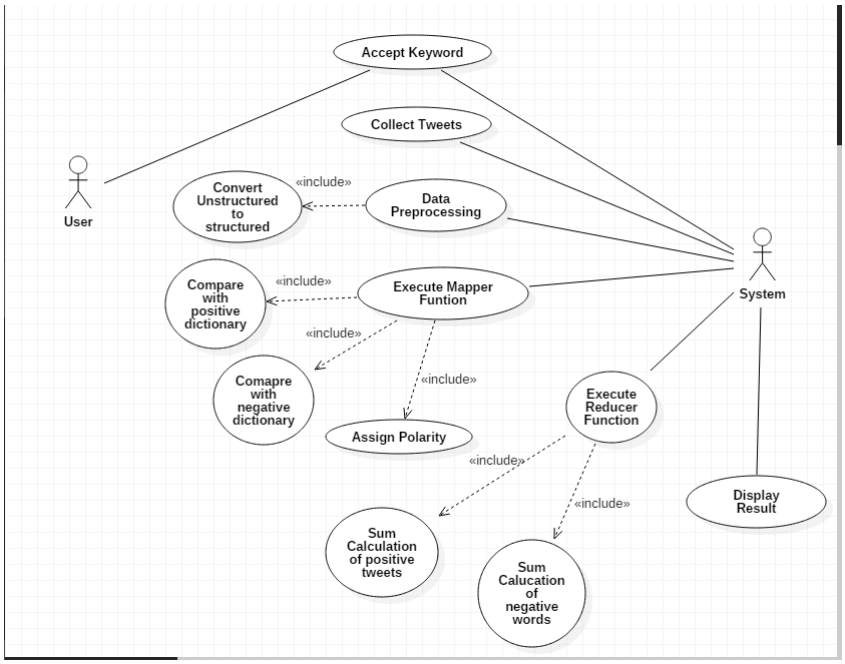


Figure B.3: Use Case Diagram

Activity diagram:

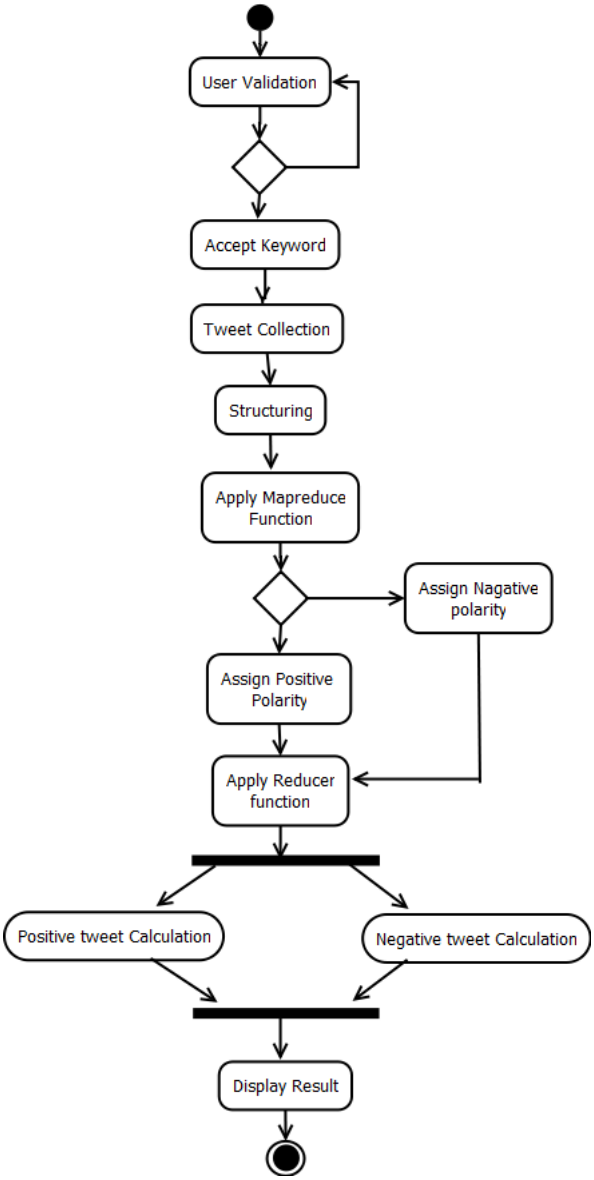


Figure B.4: Activity Diagram

B.5 Assignment 5

Problem Statement : Test the defined problem statement using generated test data

Test Cases generated :

Test CaseId	Input	Description	Expected output	Actual output	Result
Id1	check whether keyword contain english words	keyword containig words in english language	Classification of sentiment are genrated	Classifying tweets into positive,negative and neutral	Pass
Id2	check whether keyword contain words in other language	keyword containing words in language other than english	classification of sentiments should not be generated	error message	Fail
Id3	check whether keyword is a integer	keyword containing positive and negative numbers	classification of sentiments should not be generated	Classification of sentiments are still generated	fail
Id4	check keyword contains special characters	keyword entered by user are special characters	classification of sentiments should not be generated	classification of sentiments are generated	fail
Id5	Check whether the keyword contain stop words	stop words are words which do not give any emotion like a,an the,are	classification of sentiments should not be generated	classification of sentiments are generated	fail

B.6 Assignmnet 6

Problem Statement : Details of Changes made in design if any after term-I assessment or any review in competitions/conferences. Prepare document mentioning:

- Previous design
- Modified Design
- Explanation

Changes made:

Before: At first we we used only Hive queries for assigning polarity to the tokens and also for classifying tweets into positive , negative and neutral sentiments

After: In the latter part we included Map Reduce Algorithm in the design mode which mapper function is used to compare words with positive and negative dictionaries and based on that it assigns polarites to words and Reducer function is used to calculate sum of toatal number of positive and neagtive sentiments

Old Architecture :

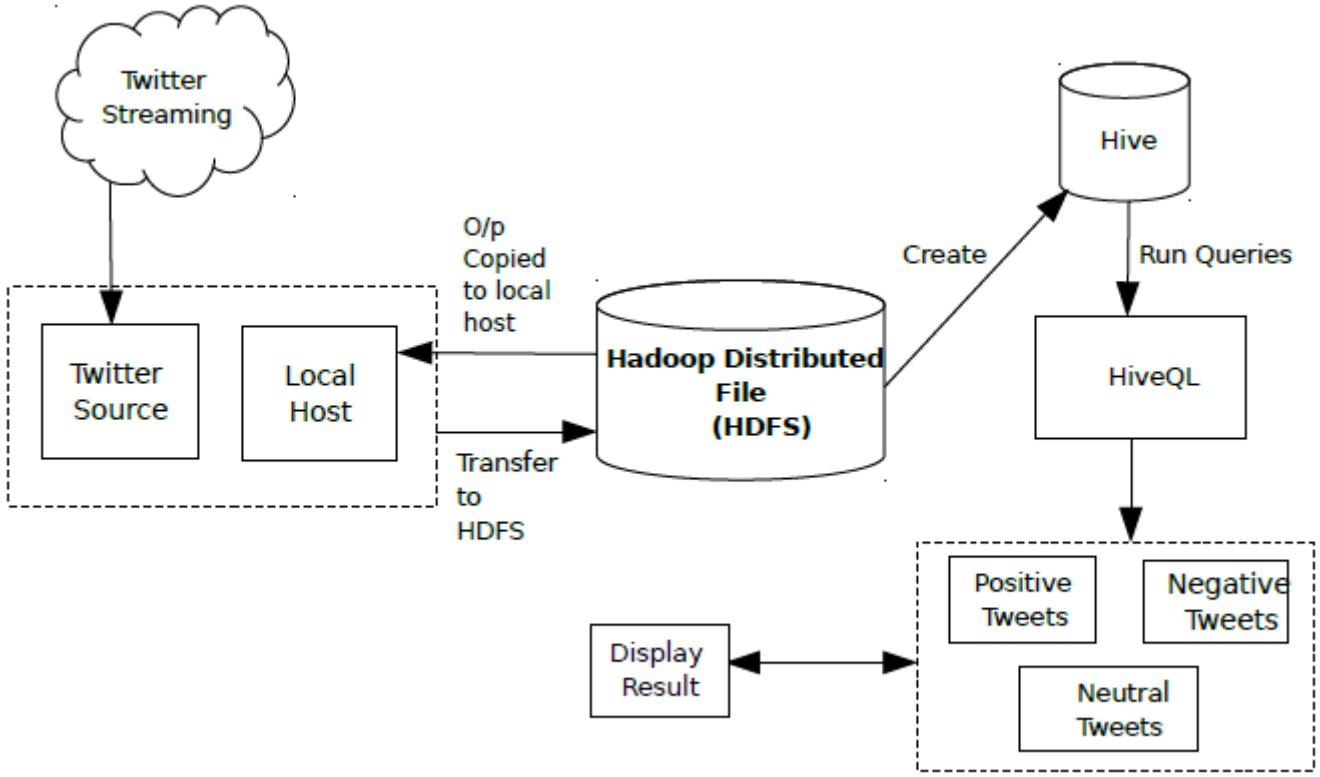


Figure B.5: Old Architecture

New Architecture :

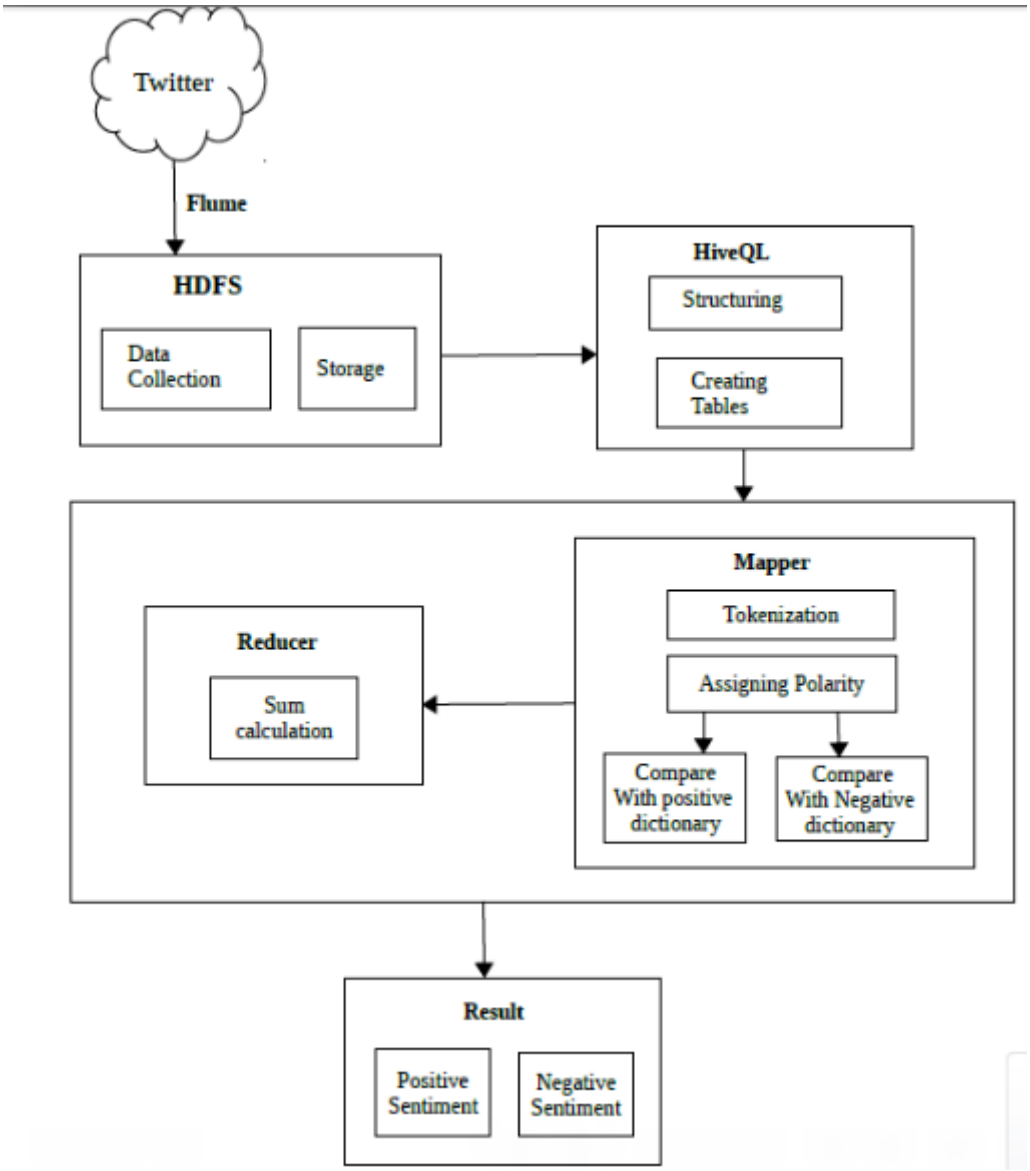


Figure B.6: New Architecture

B.7 Assignment 7

Problem Statement : Prepare document mentioning Project workstation selection, installations along with setup and installation report preparations.

Project workstation Selection:

Softwares used:

- Netbeans IDE 8
- Virtual Box
- HiveQL

Tools used:

- Flume 1.6
- Hive 1.2.1

Framework used:

- Hadoop 2.6.3

Packges Used:

- JDBC Hive Connector
- hadoop-commom-2.6.3.jar
- hadoop-mapreduce-client-core-2.6.3.jar
- hadoop-mapreduce-client-common-2.6.3.jar

- commons-cli-1.2.jar

Operating System Used:

- Ubuntu 15

Hardware Used:

- RAM 8 GB
- Hard Disk Space: 100 GB (Recommended)
- CPU Processor: intel i5

Installation Steps:

Installation steps of Hadoop 2.6.3:

- Install latest version of Java
- Install ssh
- Create and Setup SSH Certificates
- Install latest version of hadoop
- Setup Configuration Files

The following files will have to be modified to complete the Hadoop setup:

- /.bashrc
- /usr/local/hadoop/etc/hadoop/hadoop-env.sh
- /usr/local/hadoop/etc/hadoop/core-site.xml
- /usr/local/hadoop/etc/hadoop/mapred-site.xml.template
- /usr/local/hadoop/etc/hadoop/hdfs-site.xml

1. Format the New Hadoop Filesystem
2. After making all changes in respective files and after switiching to respective hadoop user as a root user start hadoop with following command on command prompt. start-all.sh or (start-dfs.sh and start-yarn.sh)
3. To check whether hadoop framework is running or not use following command: jps
4. If its running correctly then it gives following output: hduser@laptop:/usr/local/hadoop/sbin jps
   9026 NodeManager
   7348 NameNode
   9766 Jps
   8887 ResourceManager
   7507 DataNode
5. To stop hadoop framework :stop-all.sh

Installion Steps of hive 1.2.1:

- Install latest version of hive
- Set Hive envirnoment variables
- Set HADOOP PATH in hive confg.sh
- Create hive directories within HDFS
- set READ or WRITE permission for table
- Make respective changes in hive configuration files
- launch hive using command hive.
- Now we can create databases,tables and work on those by running different hive queries.

Installation steps of Flume 1.6:

PART 1 :

- Download latest version of flume
- Extract the tar file and store it to respective directory.
- Remove unnecessary jar files from lib directory of bin
- download flume-sources-1.0-SNAPSHOTS.jar from following link
- https://drive.google.com/file/d/0B-Cl0IfLnRozUHcyNDBJWnNxdHc/view usp=sharing
- Move the flume-sources-1.0-SNAPSHOT.jar file from Downloads directory to lib directory of apache flume
- Edit flume-env.sh file and set java HOME and FLUME CLASSPATH.

PART 2 :

1. login into twitter app.
2. Create new application and enter all the details in the application:
3. Get following Keys:
    - Access Tokens
    - Access Token Secret
    - Consumer Key
    - Consumer Secret key
4. download flume.conf file using following link: https://drive.google.com/file/d/0B-Cl0IfLnRozdlRuN3pPW usp=sharing
5. Edit flume.conf
6. Add Consumer Key, Consumer Secret, Access Token, Access Token Secret ,keyword into flume.conf file.

7. Use following command to strat flume so as to fetch data from twitter app. ./bin/flume-ng agent -n TwitterAgent -c conf -f /usr/lib/apache-flume-1.4.0-bin/conf/flume.conf

Installation steps for Netbeans IDE:

1. Download the installer using following links netbeans.org/downloads/.
2. Give executable permission.
3. Start the installer:
4. Netbeans IDE installed and now we can create projects.

B.8 Assignment 8

Problem Statement : Test tool selection and testing of various test cases for the project performed and generate various testing result charts, graphs etc. including reliability testing

We have performed following testing on our project:

Black-Box testing

- We have performed black box testing to ensure that whether it correctly accepts inputs and generate expected output
- Our application accepts keywords only in English language and generate classification of sentiments that is positive and negative sentiments
- If user enters keywords like integer and special characters then in that case it still generates classification of sentiments, if integers or special characters are there in tweets but 0 polarity will be assigned to integers and special characters

White-box Testing:

- White box testing is performed to test all the loops and statements in the main code of the project
- First it checks for loop which contains condition for the no of words present it mdata ,for loop will continue executing till all the words in mdata are finished so we check whether condition under for loop is true and if it is true it executes if statement
- If condition in if statement is true then it compares words with positive dictionary and assigns polarity 1, else it compares with negative dictionary and assigns polarity -1

Usability Testing:

- Usability testing is a black-box technique and is used to identify any error(s) and improvements in the software by observing the users through their usage and operation.
- In our system we have created user friendly GUI which is easy and efficient to use, user needs to enter keyword and based on that keyword tweets are collected and they are processed and result is displayed

Test Id	Input	Description	Expected Output	Actual Output	Pass/fail
BV1	Keyword containing English word	Keyword entered by user in english language	Classification of sentiments	Generating Sentiments as positive and negative sentiments	Pass .
BV2	Keyword containing only integers	Keyword entered by user containing only integers	Classificatin of sentiments should not be generated	Generating sentiments	Fail
BV3	Keyword containing only special characters	Keyword entered by user containing only special characters	Classification of sentiments should not be generated	Generating Sentiments	fail .

Table B.1: Black-Box Test cases

Test Id	Input	Description	Expected Output	Actual Output	Pass/fail
EQ1	Condition under for loop is true	For loop will continue executing until all spilt words from mdata are checked	It should go into if statement	Executing if statement	Pass .
EQ2	Condition under for loop is false	if all words in mdata have been checked	It should come out of for loop	coming out of for loop	pass
EQ3	Condition under if is true	Words are compared from mdata to positive dictionary	It should assign polarity	Assigning polarity	Pass .
EQ4	Condition under if is false	If words are not there in positive dictionary	It should go to else if	Executing else if	Pass .
EQ5	Condition under else if is true	Words are compared from mdata to negative dictionary	It should assign polarity	Assigning polarity	Pass .
EQ6	Condition under else if is false	If words are not there in negative dictionary	It should go to else part	Executing else part	Pass .

Table B.2: White-Box test cases

Test Id	Input	Description	Expected Output	Actual Output	Pass/fail
PO1	Valid input	Keyword entered by user in english language	Classification of sentiments	Sentiments are classified into positive and negative sentiments	Pass .

Table B.3: Positive Test cases

Test Id	Input	Description	Expected Output	Actual Output	Pass/fail
NO1	Invalid Input	Keyword entered by user are special characters	It should not generate sentiments	Generating Sentiments as positive and negative sentiments	Fail .
NO2	Invalid Input	Keyword entered by user containing only integers	It should not generate sentiments	Generating sentiments	Fail

Table B.4: Negative test cases

APPENDIX C

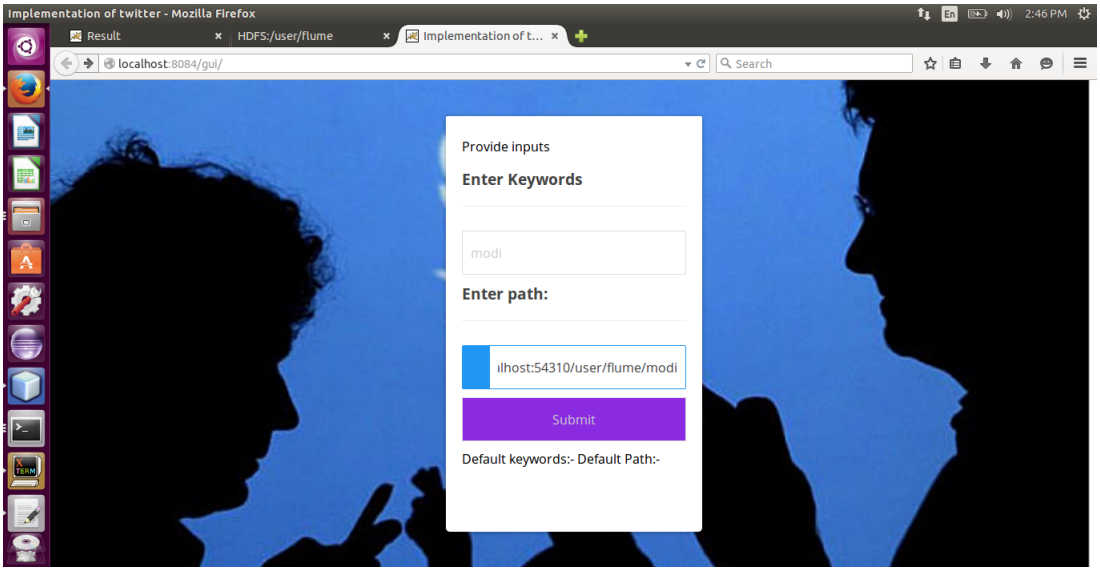
C.1 Project Quality and Reliability Testing

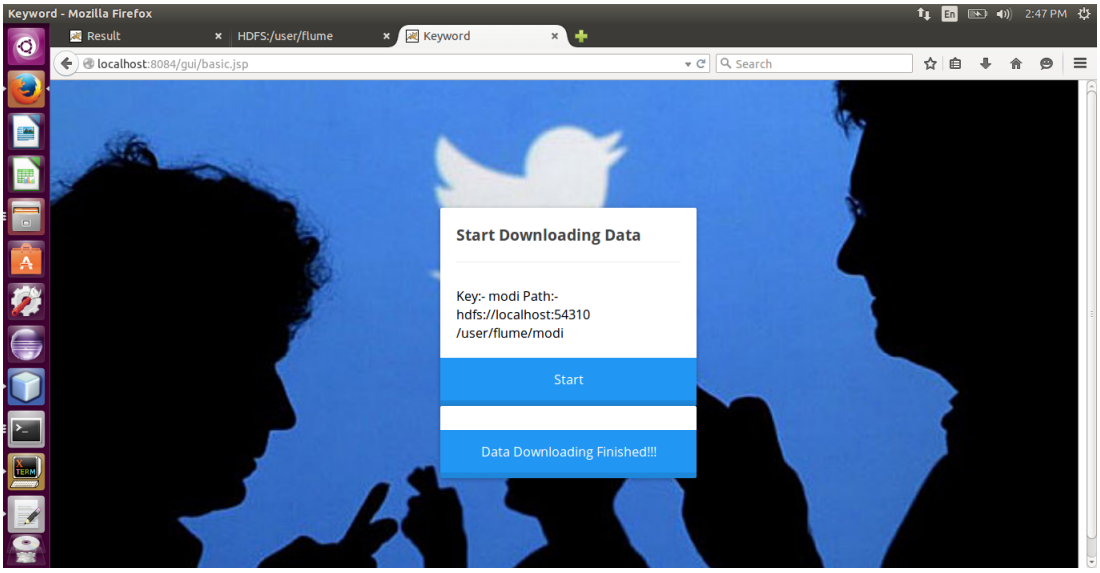
C.1.1 GUI testing

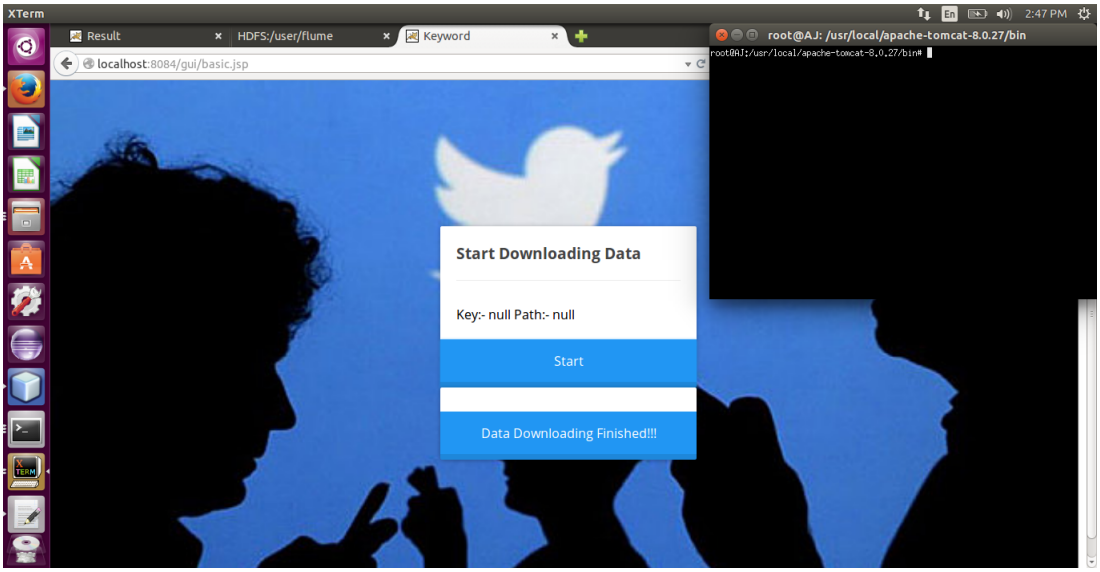
Test Id	Input	Description	Expected Output	Actual Output	Pass/fail
BV1	Keyword containing English word	Keyword entered by user in english language	Classification of sentiments	Generating Sentiments as positive and negative sentiments	Pass .
BV2	Keyword containing only integers	Keyword entered by user containing only integers	Classificatin of sentiments should not be generated	Generating sentiments	Fail
BV3	Keyword containing only special characters	Keyword entered by user containing only special characters	Classification of sentiments should not be generated	Generating Sentiments	fail .

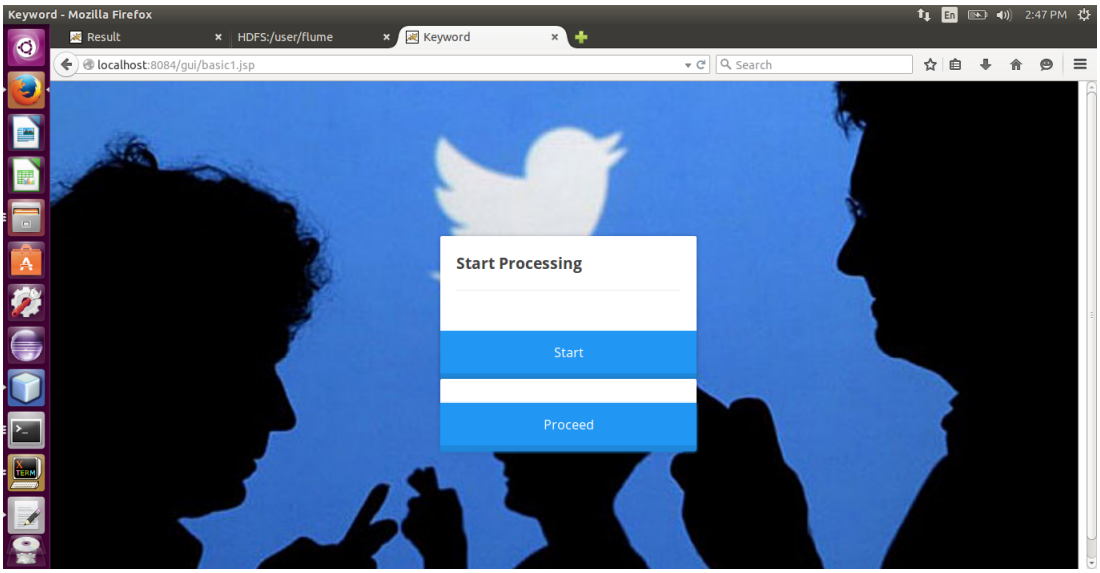
Table C.1: GUI Test cases

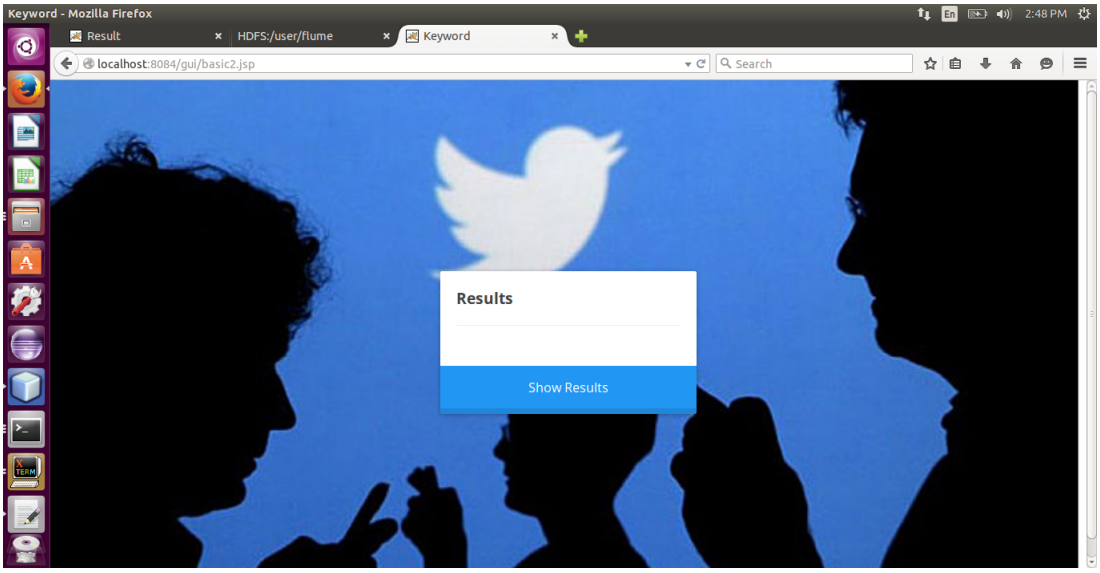
C.2 GUI Screenshots




Keyword - Mozilla Firefox
Result
HDFS:/user/flume
Keyword
localhost:8084/gui/basic.jsp
Search
Start Downloading Data
Key:- modi Path:-
hdfs://localhost:54310
/user/flume/modi
Start
Data Downloading Finished!!!


XTerm
Result
HDFS:/user/flume
Keyword
localhost:8084/gui/basic.jsp
root@AJ: /usr/local/apache-tomcat-8.0.27/bin
Start Downloading Data
Key:- null Path:- null
Start
Data Downloading Finished!!!


Keyword - Mozilla Firefox
Result
HDFS:/user/flume
Keyword
localhost:8084/gui/basic1.jsp
Search
Start Processing
Start
Proceed



APPENDIX D

C.3 Published Paper

Figure C.1: First paper certificate

Figure C.2: Second paper certificate

Figure C.3: First paper certificate

Figure C.4: Second paper certificate

Figure C.5: First paper certificate

Figure C.6: Second paper certificate

Figure C.7: First paper certificate

Figure C.8: Second paper certificate

C.3.1 Reviewers comments of paper submitted

Paper Title

1. Sentiment Analysis of Twitter Data using Hadoop

2. A Survey on Twitter Sentiment Analysis Using Hadoop

Name of the conference/journal

1. International journal of Engineering Research and General Science(IJERGS)
2. International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE)

Paper accepted or rejected

1. IJERGS accepted
2. IJIRCCE accepted

Review comments by reviewer

- Subject content is good
- Technical content is good
- Domain of the paper is good
- Contribution to the field is Satisfied
- Depth of research is good
- Presentation is good

Recommendation

Accepted

corrective actions if any

- References must be of font size 8.
- Strictly convert as per the IJIRCCE Single column MS word format

APPENDIX E

C.4 Plagarism Report

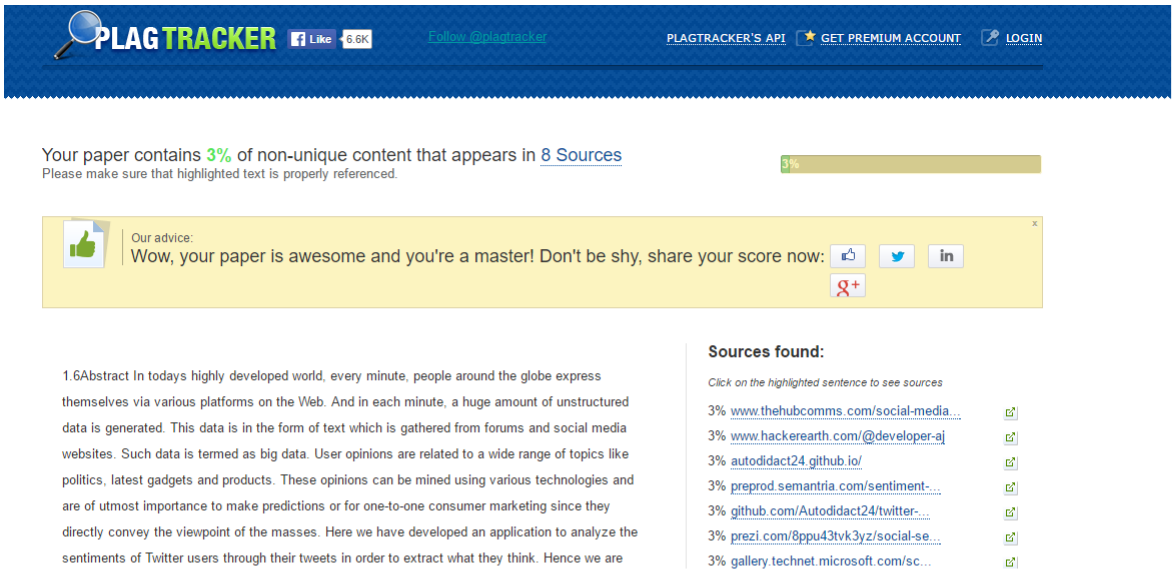


Figure C.9: Plagarism Report

APPENDIX F

D.1 Information of each project member

- Name: Anjali S Kante
- Date of Birth: 10/05/1995
- Gender: Female
- Address: Kante tel udyog,new adarash colony,ausa road Latur
- Email id: anjali.kante@gmail.com
- Contact No: 9890487176
- Paper Published: Sentiment analysis of twitter data using Hadoop(IJERGS)and A surver on twitter sentiment analysis using Hadoop(IJIRCCE)

- Name: Anita Kumari
- Date of Birth: 07/07/1994
- Gender: Female
- Address: A-3/32 jai shiv shankar society opposite home guard on alandi road pune-03
- Email id: anielegant123@gmail.com
- Contact No: 8975963361
- Paper Published: Sentiment analysis of twitter data using Hadoop(IJERGS)and A surver on twitter sentiment analysis using Hadoop(IJIRCCE)

- Name: Shriya Satish Samak
- Date of Birth: 19/10/1994
- Gender: Female
- Address: 10 And 5 sadashiv path Vishal apt Pune-411030
- Email id: shriyasam10@gmail.com
- Contact No: 9420150274
- Paper Published: Sentiment analysis of twitter data using Hadoop(IJERGS)and A surver on twitter sentiment analysis using Hadoop(IJIRCCE)

- Name: Ajinkya R Ingle
- Date of Birth: 05/11/1994
- Gender: Male
- Address: ”samarth” Vijay Housing Society,Gorakshan Road, Akola-444004
- Email id: ajinkyaingle05@gmail.com
- Contact No: 8793646961
- Paper Published: Sentiment analysis of twitter data using Hadoop(IJERGS)and A surver on twitter sentiment analysis using Hadoop(IJIRCCE)