



**LEXICON BASED OPINION MINING ON TWITTER DATA BY USING
HADOOP**

MOHAMMED RAAED MAHMOOD ALKSSO

August 2017

**LEXICON BASED OPINION MINING ON TWITTER DATA BY USING
HADOOP**

**A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES OF
ÇANKAYA UNIVERSITY**

**BY
MOHAMMED RAAED MAHMOOD ALKSSO**

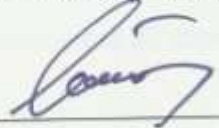
**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF MATHEMATICS
INFORMATION TECHNOLOGY PROGRAM**

August 2017

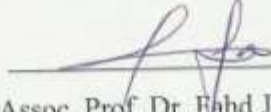
Title of the Thesis: **LEXICON BASED OPINION MINING ON TWITTER
DATA BY USING HADOOP.**

Submitted by **MOHAMMED RAAED ALKSSO**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.


Prof. Dr. Can ÇOĞUN
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.


Assoc. Prof. Dr. Fahd JARAD
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.


Assist Prof. Dr. Abdül Kadir GÖRÜR
Supervisor

Examination Date: 17.08.2017

Examining Committee Members

Assist. Prof. Dr. Abdül Kadir GÖRÜR (Çankaya Univ.)

Assist. Prof. Dr. Bülent Gürsel Emiroğlu (Kırıkkale Univ.)

Assist. Prof. Dr. Reza Hassanpour (Çankaya Univ.)



STATEMENT OF NON-PLAGIARISM PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : MOHAMMED ALKSSO

Signature : 

Date : 17.08.2017

ABSTRACT

LEXICON BASED OPINION MINING ON TWITTER DATA BY USING HADOOP

MOHAMMED RAAED MAHMOOD ALKSSO

M.Sc., Department of Mathematics

Information Technology program

Supervisor: Assist Prof.Dr. Abdül Kadir GÖRÜR

August 2017, 55 pages

In this thesis, we will highlight findings of the assumptions obtained by using the methodologies of machine learning with Hadoop by Virtual Machine.

The practical setup was started to carry out the experiments to study and find tweets by specific words and these tweets are to be collected only within a specific domain and data is to be saved in Hadoop. Then, training data such as the pre-processing operations is to remove all things which are not necessary and extract the features. After that, the classification of tweets using machine learning algorithms (supervised and unsupervised) with the ability to analyse the texts of tweet microblog is to detect emotions by different types of the lexicon. Furthermore, the cluster in Mahout was used to collect data at same polar to know what is best service or product which was expressed positively.

Finally, we prove the objectives which were collected from the achieved results based on accuracy of the classification.

Keywords: Hadoop, Lexicon, Opinion mining, Sentiment.

ÖZ

HADOOP KULLANARAK TWITTER VERİLERİ ÜZERİNDEKİ GÖRÜŞ MADENCİLİĞİ TABANLI VERİ SÖZLÜĞÜ

MEHMET RAAET MAHMUT ALKSSO

Yüksek Lisans, Matematik Bölümü

Bilgi Teknolojisi programı

Tez Yöneticisi: Assist Prof. Dr. Abdülkadir GÖRÜR

Ağustos 2017, 55 sayfa

Bu tezde, Hadoop tarafından Sanal Makine ile makine öğrenme metodolojilerini kullanarak elde edilen varsayımların bulgularını vurgulayacağız. Pratik kurulum, belirli kelimeleri kullanarak twitler incelemek ve bulmak için deneyler gerçekleştirmek üzere başlatıldı ve bu twitler sadece belirli bir alan içerisinde toplanacak ve veriler Hadoop'ta saklanacaktır. Ardından, ön işleme işlemleri gibi eğitim verileri, gerekli olmayan her şeyi kaldırıp, özellikleri ayıklamaktır. Bundan sonra, twit microblog'un metinlerini analiz etme yeteneğiyle makine öğrenme algoritmaları (gözetim altında ve denetlenmemiş) kullanarak twitlerin sınıflandırılması, farklı türdeki sözlüğün duygularını algılamaktır. Ayrıca Mahout'daki kümesi, aynı kutupta veri toplayıp olumlu ifade edilen veya en iyi hizmetin ne olduğunu bilmek için kullanılmıştır. Sonunda, sınıflamanın doğruluğuna dayanarak elde edilen başarılı sonuçlardan toplanan hedefleri kanıtıyoruz.

Anahtar Kelimeler: Hadoop, veri sözlüğü, görüş madenciliği, duygusallık

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Assist. Prof. Dr Abdül Kadir GÖRÜR, who has guidance, suggestions, and encouragement me through the development of this thesis.

I would like to express my deep gratitude to my father, mother and my brothers for their endless and continuous encourage and support throughout the years.

It is a pleasure to express my special thanks to my family my wife (Sevgili karım) and my sons (Mustafa & Mohanad) with whom we have shared good and bad times many years.

TABLE OF CONTENTS

STATEMENT OF NON-PLAGIARISM.....	iii
ABSTRACT.....	iv
ÖZ.....	v
ACKNOWLEDGEMENTS.....	vi
TABLE OF CONTENTS.....	vii
LIST OF FIGURES.....	xi
LIST OF TABLES.....	xii
LIST OF ABBREVIATIONS.....	xiii
CHAPTERS	
1. INTRODUCTION	
1.1. Introduction	1
1.2. Problem.....	2
1.3. Research Question.....	3
1.4. Objective.....	3
2. RELATED WORK AND BACKGROUND	
2.1. Overview of background.....	6
2.2. Related Work.....	6
2.3. Big Data.....	8
2.3.1. Characteristics of Big Data.....	9
2.3.2. Twitter.....	10
2.4. Twitter 4J.....	11
2.5. Apache Flume.....	12
2.5.1. Features of Flume.....	12
2.6. Apache Hadoop.....	13
2.6.1. Hadoop Architecture	13
2.6.2. Hadoop Distributed File System (HDFS)	14
2.6.2.1. Logical View of HDFS	15

2.6.3. MapReduce	15
2.6.3.1. Logical View of MapReduce	16
2.6.4. Hadoop Common Utilities	16
2.6.5. Hadoop YARN Framework	16
2.7. Apache Hadoop Ecosystem	17
2.8. Apache Mahout	17
2.9. Sentiment Analysis in Twitter	18
2.10. Opinion Mining	18
2.11. Types Opinion Mining	19
2.12. Polarity Detection	20
2.13. Emotion Mining	20
2.14. Java language	21
2.15. Oracle VM Virtual Box	21
3. RESEARCH METHODOLOGY AND FRAMEWORK	
3.1. Overview of Methodology	22
3.2. Framework	22
3.3. Data collection/Collecting Tweets	24
3.3.1. Track Parameter value	25
3.3.2. Locations Parameter	25
3.4. Training Corpora	26
3.5. Data Pre-Processing	26
3.5.1. URL/ HTTP Removal	26
3.5.2. Repetition of Letters Removal / Exaggerated word shortening	26
3.5.3. Stop words Removal	27
3.5.4. Hashtag Removal	27
3.5.5. Punctuations/ Digital words Removal	27
3.5.6. Tokenization	27
3.5.6.1. Whitespace tokenized	27
3.5.6.2. Stemming and lemmatization	28
3.5.7. Text Correction	28
3.5.8. Neutral Labels Removal	28
3.6. Feature Selection	29

3.6.1. Negation Features	29
3.6.1.1. Unigrams model with Negation Features	29
3.6.1.2. Bigrams model with Negation Features	30
3.6.2. Part of Speech Features	30
3.6.2.1 Split words or conjunctions	30
3.6.2.2. Interjection	30
3.6.3. Emoticon Features	31
3.6.4. Opinion classifier features	32
3.6.4.1. Positive opinions	32
3.6.4.2. Negative opinions	32
3.6.4.3. Neutral opinions	32
3.7. Machine learning	33
3.8. Supervised Learning Algorithms	33
3.9. Unsupervised Learning Algorithms	33
3.10. Lexicon based approach	34
3.11. Binary models Lexicon	35
3.11.1. Bing Liu Lexicon	35
3.11.2. Lexicon Developer	36
3.12. Rules testing data	36
4. SETUP, FINDINGS and RESULTS	
4.1. Overview of SETUP	37
4.2. Data FRAMEWORK	37
4.3. Hadoop Implementation	39
4.3.1. A Flume Connection	40
4.4. Data streaming at real time	41
4.5. Training data	42
4.5.1. Polar Extraction	45
4.5.2. Negation Extraction	46
4.6. Lexicon Developer and Statistics	48
4.7. Accuracy Measure and Evaluation	49
4.8. Target and making decision	51

5. DISCUSSION, CONCLUSION AND FUTURE WORK	
5.1. Discussion and CONCLUSION.....	54
5.2. Future Work.....	55
REFERENCES.....	R1
APPENDICES.....	A1



LIST OF FIGURES

FIGURES

Figure 1	Illustrates Characteristics of Big Data	9
Figure 2	Showing work of Twitter4J.....	11
Figure 3	Illustrates Architectural Flume in Hadoop (Apache Hadoop).....	12
Figure 4	Illustrates timeline of Hadoop	13
Figure 5	Illustrates Hadoop Architecture without details.....	14
Figure 6	Illustrates Architectural Hadoop cluster include MapReduce and HDFS.....	16
Figure 7	Illustrates Apache Hadoop Ecosystem.....	17
Figure 8	Overview of the framework	23
Figure 9	Illustrates emotions in text.....	31
Figure 10	Illustrates the workflow of opinion mining.....	34
Figure 11	Illustrates Key words in system.....	38
Figure 12	Illustrates Tasks choose key word with location.....	38
Figure 13	Illustrates create NameNode in Hadoop.....	39
Figure 14	Illustrates connect between hadoop, flume and twitter.....	40
Figure 15	Illustrates data streaming in VM.....	41
Figure 16	Illustrates data storage in hadoop.....	41
Figure 17	Illustrates input data in path.....	42
Figure 18	All pre-processing steps.....	43
Figure 19	Illustrates Polar Extraction and Negation Extraction.....	45
Figure 20	Illustrates comparison between numerical ratio of True and False	49
Figure 21	Chart comparison between numerical ratio of True and False	50
Figure 22	Illustrates types of accuracy ratios.....	50
Figure 23	Query tags created to clustering by keywords.....	51
Figure 24	Illustrates analysis only (1200) tweets.....	52
Figure 25	Illustrates analysis all tweets.....	53

LIST OF TABLES

TABLES

Table 1 Illustrates contents examples key words.....	24
Table 2 Illustrates Track examples.....	25
Table 3 Illustrates stemming and lemmatization.....	28
Table 4 Some of emotions Positive and Negative.....	31
Table 5 Includes some words from Bing Liu Lexicon.....	35
Table 6 Illustrates the testing data.....	36
Table 7 Illustrates comparison between before and after Pre-Processing.....	44
Table 8 Illustrates Equation of polarization.....	47
Table 9 Illustrates Lexicon Statistics.....	48
Table 10 Illustrates analysis only (1200) tweets.....	52
Table 11 Illustrates analysis all tweets.....	53

LIST OF ABBREVIATIONS

HDFS	Hadoop Distributed File System
VM	Virtual Machine
BD	Big Data
OM	Opinion Mining
ML	Machine learning
GI	General Inquirer
RDBMS	Relational database management systems
URL	Uniform Resource Locater
HTTP	Hypertext Transfer Protocol
NLP	Natural Language Processing
LTJ	Language Tools Java
IBSG	Cisco Internet Business Solutions Group

CHAPTER ONE

INTRODUCTION

1.1 Introduction

Growth in the volume of big data limits the ability of companies to manage this big data and control it effectively and it is having problems retrieving vast data, hard to deal with by employing relational database management systems with statistics desktop and simulation packages, were simultaneously working programs and large-scale work is required on the tens or hundred or even thousands of servers to handle the increase of big data. [1]

When we explore the modern technology, it is simple to conclude that it is almost all about the increasing amount of data being manufactured. however, the ratio of escalation is accelerating as well. Every second of our life, vast data sets flow, such as Emails, Facebook posts and messages or Twitter posts, is increasing. The most challenging task is obtaining the most beneficial aspects from this data. [2]

With the emergence of such social media as Twitter, exceeding 500 million users registered and exchanging more than 400 million messages daily, has turned priceless for companies to track their feedbacks and retrieve information in many fields, sentiment field being one of them, by extracting, sentiment analyzing and opinion mining of the posted is playing public opinion about them.[3]

One of the important steps towards processing and the most significant challenge is Big Data on Twitter. The performance of analysis tools has become to work in current various fields like services or products based on an implementation of machine learning algorithm with supervised learning and unsupervised learning, so it needs careful studying on whether it is an appropriate method to retrieve information accurately.

One promising solution is an adoption of Hadoop MapReduce framework to enable machine learning algorithms to scale up on large-scale datasets in distributed systems. Apache Mahout which is a collection of algorithms of machine learning applied to perform clustering, classification and recommendation. Mahout is becoming more and more popular because it is a new open source project that is able to run on top of the Hadoop framework. [4]

This system is developed to deal with big data in Twitter and opinion mining, so when we need service while comparing two types of services having same characteristics to be chosen and not follow-up on advertising because organizations always advertise their goods and brands by demonstrating only positive sides and characteristics of the product and hide the negative ones. So the best way of selecting the service is to consider the opinion of the people who have already used that service. This system is developed to deal with big data in Twitter and opinion mining with a sentiment analysis.

This task aims to use a lexicon to find the opinion mining and predict three axes (positive, negative and neutral) that could appear in the future. It is clear, therefore, that the prediction of a relation back to benefit on many events that may affect either an individual or an entire community preference like choosing product or service.

1.2 Problem

The level of technology got to the point, where humans communicate through social media day-to-day and are capable of sharing their lives using social network applications. At that point, relating to data storing structures, having scalability properties and efficient processing algorithms, some problems have occurred. A perfect potential of data mining analysis can be seen in detecting meaningful insights in social networks data. A potential solution to this difficulty is to encourage using opinion mining to find the sentiment that is integrated into diagnosis of future approach, and through analyzing the current big data, full use of big data in computing, especially after the appearance of big data such as Twitter.... etc. [5]

When it comes to the big amount of data and loads live all the data into Hadoop, this process is unsatisfactory of big data collection and filtering only, in real world computing environment, this information is not complete and finds opinion and actual decision. [6]

New demands towards efficiency, accuracy and scalability rise up in data classification by implementing advanced algorithms such as machine learning algorithms (Mahout) and opinion mining, to make a possible examination of the moods of an individual. It can support to choose the positive, negative or neutral views of an individual based on his behavior on tweets by the lexicon, to get benefits flexibility, speed and predictability, and know what users think or how they feel about products or service.

1.3 Research Question

- What is the big data in cloud computing?
- How is it possible to classify Big Data in Twitter and how collect it?
- What is Flume and Twitter 4j?
- What are the characteristics of Hadoop?
- How can apply algorithms in Mahout with Hadoop for opinion mining on Twitter?
- What is the difference between supervised learning and unsupervised learning?
- How can sentiments be found?
- How can the lexicon be developed?
- How can make a decision?

1.4 Objective

Twitter is an online social network, where users can share short messages (tweets). Its deployment gained worldwide popularity with more 320 million active users. Users signed in to service can share their moments with friends, include comments and re-tweet other tweets. Therefore, Twitter became an interesting platform for opinion mining and finding sentiment. [7]

As a consequence, the aim of this study is to focus on opinion mining using modern techniques cloud computing open source like Hadoop (virtual machine) that supports distributed applications for big data and it supports Modern query methods by using Hadoop distributor file system (HDFS) with architectural MapReduce.[8]

Studying modern and unique applications of Hadoop in big database and big data classification system, such as Mahout in real time or batch offline and work to set up a cluster Hadoop with Mahout on the Internet, allows us to free storage with fast processing, and implement smart systems an algorithm of machine learning with supervised learning and unsupervised learning to get the speed and accuracy in the diagnosis of sentiment and give the ability to predict choice of any service or a product, e.g. airlines preference.

This thesis proposes opinion mining for finding sentiment in Big Data like Twitter based on a machine learning with lexicon and Mahout in Hadoop about product or service information and given the accuracy of opinion to (positive, negative or neutral).

The remainder of this thesis is organized as follows:

- **CHAPTER ONE:** Explaining Introduction, Problems, Research Questions and Objective.
- **CHAPTER TWO:** The overview of previous studies and explaining all parts to be used in thesis like Big Data (Twitter), Twitter 4j, Apache Flume, Apache Hadoop, Apache Mahout, Java and virtual machine.
- **CHAPTER THREE:** Proposed framework and method of collecting Tweets and training data sets like Data Pre-Processing and Feature selection and use Lexicon.
- **CHAPTER FOUR:** Experimental setup showing experimental results developed on real data sets, divided into positive or negative and performed by machine learning.
- **Finally, CHAPTER FIVE:** Concluding work with results overview, discussing the results and concluding respectively.

CHAPTER TWO

RELATED WORK AND BACKGROUND

2.1 Overview of Background

This chapter contains an explanation view on previous studies related to this study, then the explanation of the general background of the different related technologies, and explanation of all parts to be used in this thesis, such as big data (Twitter), Twitter 4j, Apache Flume, Apache Hadoop, Apache Mahout, Java and virtual machine.

The next step will be discussing the field of studying people's opinions (opinion mining) by lexicon to find emotions, sentiments and, subsequently, evaluate towards many entities, such as topics, keywords, events, products, individuals and services. final that, expressing opinion mining which gives a positive or negative statement about an objects or keywords.

2.2 Related Work

Gowtamreddy [9] Many opinions on some goods are posted by internet users using review sites and social networks, so Opinion Mining becomes important in e-trade, online shopping and tourism for finding the best way reviews, views, emotions and opinion analysis must be used automatically from a text, big data and a speech by implementing Apriori frequent algorithm which is the best to be applied for mining reviews based on reviews which consumers post. Customer opinion plays an extremely important part in everyday life, especially when it comes to decision making. By implementing this method an opinion analyzing system gets created, which is the key for implying a judgment of various consumer products.

Also, say Vijayaraghavan. [10] A range of Microblogging services, particularly Twitter, has become commonly chosen pathway for displaying the opinion and being an active contributor to revelation on each part of broad spectrum of subjects, and developing a method to identify the specific ideas and sentiments that represent the overall conversation surrounding a topic or event as reflected in collections of tweets and evaluating a large-scale data analytics framework, based on recent advances in deep neural networks for identifying and analyzing election-related conversation on Twitter on a continuous longitudinal basis in order to identify representative tweets across prominent election issues.

Rahnam [11] Sentiment analysis on Twitter public stream and real-time performance of the routine task provides us with a perception of Twitter users public opinion on current time events which were subjected to analysis. Therefore, it is viable to use SAMOA as a framework providing the support for series of scalable online learning algorithms, e.g. Vertical Hoeffding Tree. SAMOA's VHT learner (including Apache Storm) was applied as the Stream Processing Engine.

Houen [12] Opinion mining has faced the challenge of determining an opinion of the author on a subject from the information contained within a text in the original language. Some form of machine learning approach is commonly employed to those and efficiency level of which varies. Consequently, the effectiveness of the semantic frame-based analyzer Frame Net has been explored by the researchers, Having the purpose of improvement of current techniques of sentiment analysis through delivering automatically labelled semantic information to be proceeded by sentiment classifying algorithm. To reach this result, semantic information as a property of a machine learning-based classifier should be applied. In addition, the semantic analysis in a bootstrapping process is to be used to produce a set of sentiment lexicon.

2.3 Big Data

The amount and diversity of data produced within cloud computing have been remarkably increasing during recent years. As it has been noted by “the Digital Universe Study” performed by IDC (being a representative of EMC Corporation) the global data gets twice higher in each two-year time span, and by 2020 it is forecasted that it would get as far as 40ZB (Digital Universe Study). [13] Such data elevation is frequently also called a “data tsunami” and is triggered by the escalating number and popularity of social media sites as well as the elevation of networked and mobile devices i.e. the Internet of Things, funds and online trade and also physical and life sciences fields advancement. Proving the stated above, about 12 TB of data is daily processed by the online Microblogging service Twitter, whereas publishing over five hundred million tweets every day (McKinsey Global Institute). [14] Moreover, as has been forecasted by the Cisco Internet Business Solutions Group (IBSG) by 2015, the number of internet-connected devices will have reached 25 billion, and by 2020 to 50 billion (IBSG). [15] Thus, expansive datasets are generally called “Big Data”. One of the main characteristics of the Big Data is not just the content size, but also a wide range (variety) of formats and types of data. Consequently, present-day Big Data has Twitter as one of its most major sources.

2.3.1 Characteristics of Big Data

Big Data (BD) has also got other properties besides data Size or Volume, which are: data Velocity, Veracity, Value and data Variety. Subsequently, the 5 V's of the Big Data is established based on these five compounds. [16]

- **Volume:** represents an amount of information which is accumulated by a corporation. The amount of various data, which could simply sum up to terabytes or in some cases even petabytes of information, has been comprised by corporations.
- **Velocity:** represents Big Data processing time span. Swift processing increases effectiveness and since some crucial activities require instant responses, it is necessary to consider proceeding time span.
- **Veracity:** represents the extent to which a manager can rely on a piece of information for efficient decision-making. Moreover, for the future of a company, it is crucial to find the appropriate parallels within the Big Data.
- **Value:** represents the key feature of a piece of data determined by the added value which can be brought to the targeted activity, process or prognostic hypothesis/analysis by the data gathered.
- **Variety:** represents the types of the data which can be constituted by the Big Data. This kind of data has also got a structure. It can be structured, semi-structured or unstructured, and Figure (1) illustrates Characteristics of Big Data. [16]

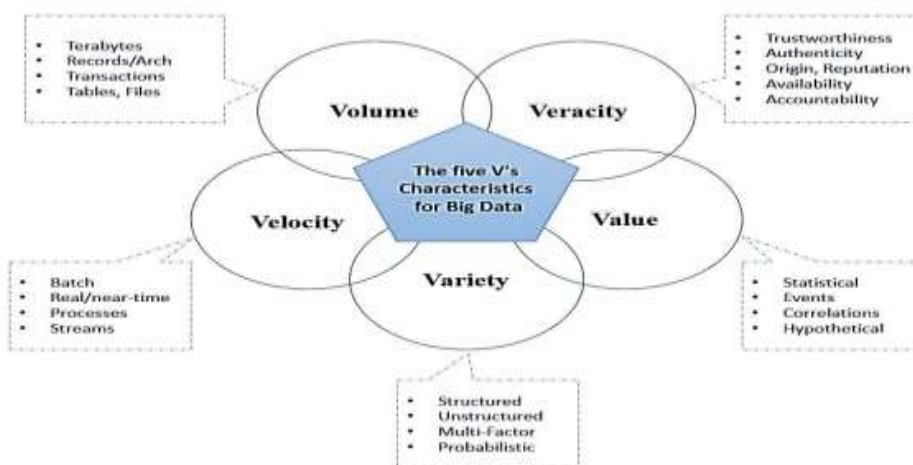


Figure (1): illustrates Characteristics of Big Data.

2.3.2 Twitter

March 21, 2006, 9:50 PM Pacific Standard Time is the date of birth of the first tweet. While working, the founder of the idea of Twitter Jack Dorsey published tweet "just setting up my twttr", four main people established Twitter (the company) Evan Williams, Jack Dorsey, Noah Glass and Biz Stone. Twitter is a Microblogging and online social networking platform that avails users to publish and read short (140 characters) text messages, named "tweets". Alexa's web traffic analysis has included it to the top 10 of most attended sites for previous two years and one of the important resources of Big Data. It comprises about over 250 million users who are monthly active and post over 500 million tweets daily. 78% of the access to Twitter is performed using mobile devices by users, and 77% of the access is performed by users who don't reside in the US. More than 35 languages are offered by this platform. The reason behind using Twitter as an unprecedented platform is to analyze and recognize trends and events framing the world both on an international and a local level due to its broad global reach and usage. A base reality of world events (e.g. natural disasters) can be understood through the use of this great media unit since even prior to news crews reaching the event areas users are able to post their impressions on Twitter. The reach, companies, people of politic and celebrities use it to vend their messages views and opinions to their fan-followers. Thus, according to the specification of the Big Data (volume, velocity, variety,...) of data which is contained on Twitter, it fits the category. [5].

As online social networking becomes more and more popular many studies have been done to discover valuable information from it. upon which social influence has drawn a lot of attention. Influence has long been studied in many fields and the findings of influence contribute a lot to advertising marketing as well as the airline's aspect. On social networks, such as Twitter, the influence represents the ability of a user to have an effect on others or the capacity to drive action. The influence can also be interpreted as the response of one user to the activity of another user on a social network. On Twitter, a small group of users who excel in spreading information is called influential users include a larger number of audiences and higher activities. In addition, the influential tweets might be retweeted more often than those of the others, for instance, President Obama is ranked as No.3. [17]

2.4 Twitter 4J

Informal Java library used for the Twitter API (unofficial Twitter API) and developing Twitter-based Java applications is called Twitter4J. Deploying Twitter4J, your Java application with the Twitter service can be effortlessly integrated.

Download of Twitter4J is available through <http://twitter4j.org/en/index.html>. Twitter4J offers several APIs allowing access to its services.

It goes from the operations of consultation of accounts (tweets, lists of friends and followers, etc.) to the operations of modification filter in order to remove what you do not need (to filter tweets based on a specific keyword). [18] Figure (2) illustrates the role of Twitter4J.

The features of Twitter4J are as follows:

1. 100% Pure Java - is able to operate on a Java Platform having version 1.4.2 or newer.
2. Independence - it does not require any additional jars.

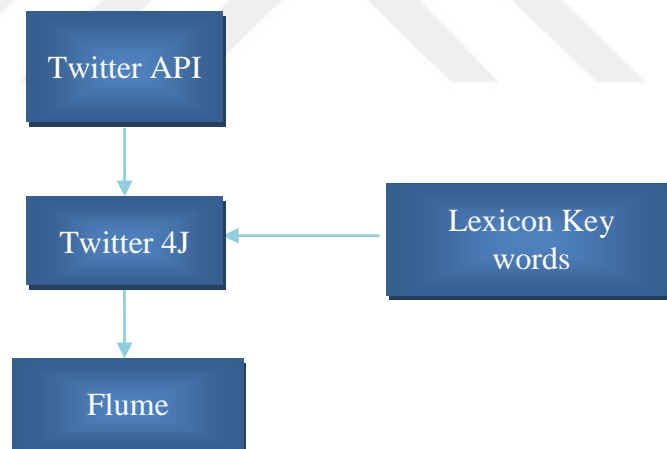


Figure (2) Showing work of Twitter 4J

2.5 Apache Flume

Flume [19] refers to a distributed, trusted, and accessible servicing used for efficient gathering, aggregation and transfer of a vast amount of data logged. Its structure is plain, adjustable and built upon data flows being streamed. Its structure is sturdy and faults tolerant, having mechanisms equipped with tunable reliability. a lot of failovers along with mechanisms for recovering. The plain extendable data model is used which allows big data flows in an application online such as Twitter that includes many aspects related to airlines or other aspects, and Figure (3) illustrates the flow of big data from Twitter into Hadoop (HDFS).

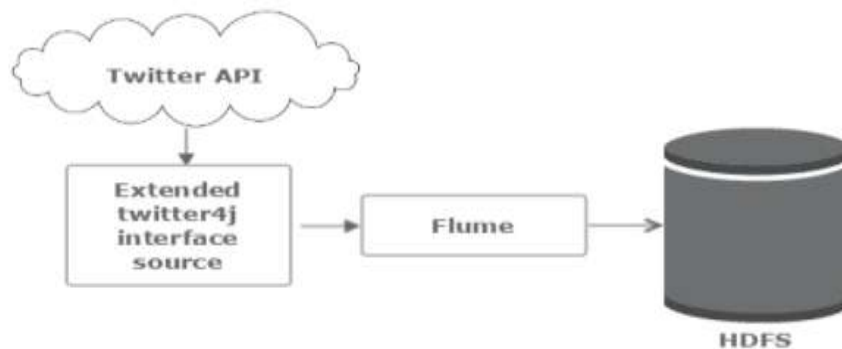


Figure (3) Illustrates Architectural Flume in Hadoop (Hadoop Ecosystem).[19]

2.5.1 Features of Flume

It has many advantages and we will glance over some of those in brief. [20]

- Log data is efficiently ingested by Flume from numerous web servers to a centralised store (HDFS).
- The data from numerous servers can be instantly transferred to Hadoop using Flume.
- In addition to the log files work, importing a vast amount of event data that social networking sites produce, can be performed by Flume as well.
- A wide range of destinations types and sources are supported by Flume with the possibility of horizontal scaling.

2.6 Apache Hadoop

In 2002, the common method of storing and organizing data was executed by applying Relational Database Management Systems (RDBMS). SQL language was used to access these. However, internet search engine storage and retrieval were not compatible with some SQL and stores. [21]

On the contrary, Doug Cutting and Mike Cafarella needed to access the internet, so Nutch project was launched. Hadoop was initially configured as the Apache Nutch project framework and in 2004. MapReduce was published by Google. The implementations were divided resulting in the creation of autonomous sub-project called Hadoop in February 2006. In January 2008, it evolved into Apache project of a high rank and its name derived from “Doug” referring to a game of his son (Yellow Elephant). [18] Design and operation of distributed applications which are processing the vast quantity of data can be performed by this open source framework. Distributed computing is a vast and diverges domain, and Figure (4) illustrates the stages of Hadoop. [8]

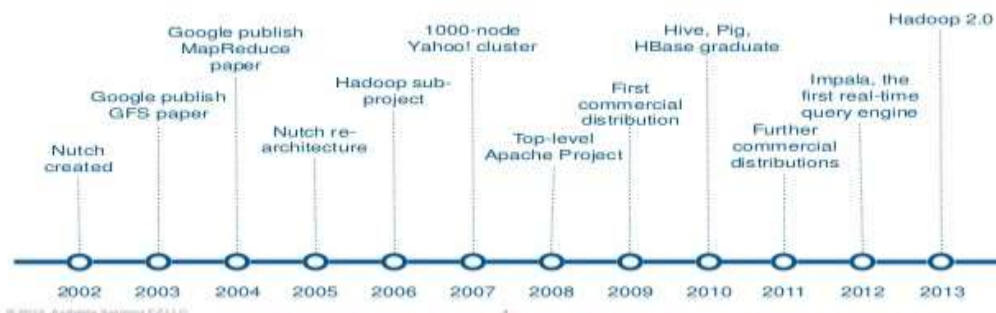


Figure (4) Illustrates Timeline of Hadoop

2.6.1 Hadoop Architecture

Hadoop has major four modules [4] and as illustrated in Figure (5):

- Storage: Hadoop Distributed File System (HDFS).
- Processing: MapReduce.
- Hadoop Common Utilities.
- Hadoop YARN Framework.



Figure (5) Illustrates Hadoop Architecture without Details

2.6.2 Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS) represents a distributed file system which was developed to be applied to hardware additionally, another key underlying concept is that of "divide and conquer" breaking a Big Data into chunked pieces or splitting compound Big Data problems into smaller tasks and fixing those is far quicker. [16,22] A single problem gets broken into multiple parts. Each node within Hadoop model contains Node (NameNode) individually, and the block has many Nodes (DataNodes) to assemble HDFS block and doing each data node to access data over the network using TCP/IP for the connection and customers used to make the connection distance to communicate with each other. [23]

One of the main advantages of using HDFS is the provision of awareness of data between JobTracker and TaskTracker with MapReduce, which we will talk about at the next point, HDFS is designed for most of the files of non-change and may not be appropriate for systems that require concurrent processes in writing, but HDFS is the best for work that is generally of the write-once and read many types. [24]

2.6.2.1 Logical View of HDFS: HDFS has two types of node:

- **NameNode:** operations within this type store all files in HDFS as a guide-tree, i.e. a hierarchy of files and directories. NameNode represents files and registers in nodes, and supporting a big quantity of files may create bottlenecks. [25]
- **DataNode:** DataNodes is executing data storing in the HDFS, during startup, each DataNode connects to the NameNode and performs a handshake. That provides a necessary strength of processing power for performing the analysis of several hundreds of terabytes on the go to the petabyte. [25]

2.6.3 MapReduce

The MapReduce programming models are working with the purpose of supporting applications containing intensive data and are operating on parallel computers as commodity clusters and have two crucial function programming basics in MapReduce which are Map and Reduce. The Map function applies to the HDFS on input data which is specific to each application to (key or JobTracker, value or task tracker) [27]

JobTracker work schedule maps or reduces jobs for the task tracker with HDFS being aware of the data location, as an instance, if node A contains information (X, Y, Z) and Node B is containing the data (a, b, c) a Node B will be scheduled by Job Tracker to implement a map or lower tasks (a, b, c) and Node A may get a schedule to implement a map or lower tasks (X, Y, Z) and this lowers the quantity of data passing through the network and averts transfer of data that is not necessary. This could have a big impact on the time of achieving the job, which is proved when making jobs. [24, 26]

2.6.3.1 Logical View of MapReduce

In the map, a status division the input data set into a large number of parts and assigns each part to a map job. The Map in the regular relation database, which sorts the data out into rows and columns which are preserved in charts, but MapReduce applies the following major elements:

- **JobTracker or master servers:** The purpose of this process is to interact with applications on a client. Another duty is a distribution of MapReduce tasks between specific nodes within each cluster. [27]
- **TaskTracker or slave servers:** That is an operation in the cluster that is able to receive tasks (Map, Reduce included) from a JobTracker. Figure (6) illustrates the work Logical View of MapReduce with HDFS in Hadoop. [27]

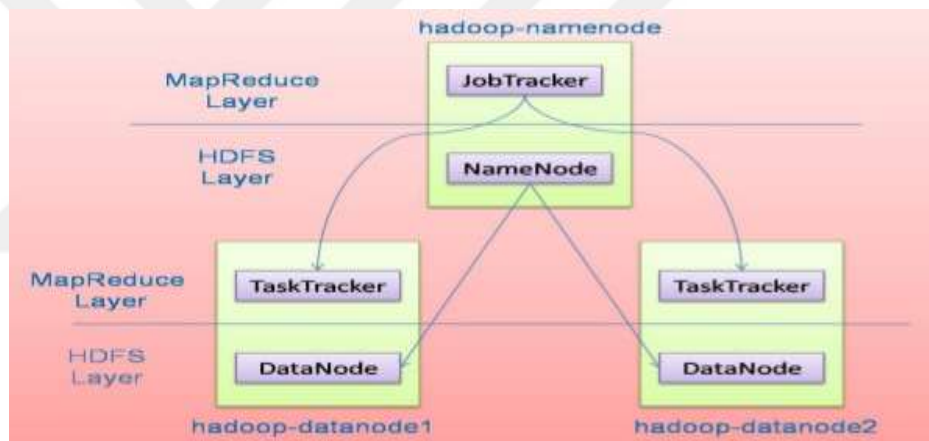


Figure (6): illustrates Architectural Hadoopinclude MapReduce and HDFS.

2.6.4 Hadoop Common Utilities: the typical utilities which are backing the other Hadoop modules when Flume needs to access HDFS or any other part of Hadoop Ecosystem. [28]

2.6.5 Hadoop YARN Framework: this name stands for Yet Another Resource Negotiator (YARN) and this is a structure to schedule job and manage the resource of cluster, and an updated way of handling resources for MapReduce jobs, and can be a more advanced way of carrying out MapReduce jobs such as classifying data by Mahout within the Hadoop. [29]

2.7 Apache Hadoop Ecosystem

Apache Hadoop Ecosystem is software that runs above of or side by side Hadoop and that achieve high-level Apache project status, but in this thesis, only two applications will be used (Flume and Mahout) with Hadoop, and Figure (7) illustrates Hadoop Ecosystem.

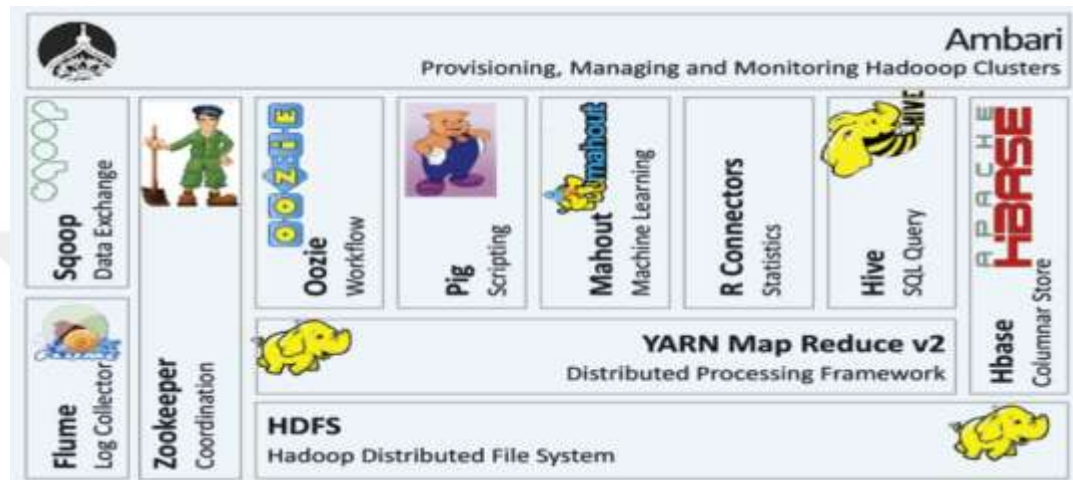


Figure (7) Illustrates Apache Hadoop Ecosystem

2.8 Apache Mahout

Project Apache Mahout belongs to Apache Software Foundation that has applied algorithms for machine learning. Apache Mahout is written in Java and open source library built on top of Hadoop (a person who rides and controls an elephant). [30] It mainly integrates many algorithms to implement three use cases collaborative filtering, classification and clustering, and provision and implementation of scalable machine learning algorithms. [31] The aim of Mahout is to make the choice for machine learning problems in which the data is too large to fit into a single machine. [30] meaning that the data cannot be processed on a single machine, Mahout fulfils the need for a machine learning tool that can scale beyond a single machine by using the MapReduce paradigm. [31]

2.9 Sentiment Analysis in Twitter

Using Twitter data for the task of sentiment analysis presents additional challenges as well as unique opportunities that must be incorporated. Tweets are informal, unstructured messages lacking many features often found in the professionally written text (grammatical structure and use of a standard lexicon). Furthermore, Tweets are limited in length by 140 characters making them exceptionally brief and therefore the message of the user is very compressed and often contextual. These challenges, however, are balanced by certain unique features about Tweets that most standard specifically the use of emoticons and hash tags. What is special about Tweets including emoticons is that the sentiment is often explicitly specified in the Tweet itself by the emotion making it ready to classify. [32]

Therefore, we can use this feature to overcome one of the largest obstacles of analyzing Twitter sentiment, and by using this method of labelling Tweets based on their emotions we can query Twitter and create a large database of different Tweets.

2.10 Opinion Mining (OM)

One of the major applications of machine learning is opinion mining. In this field, computer programs attempt to predict the emotional content or opinions of a collection of articles, blogs and comments. This becomes useful for organizing data such as finding positive and negative reviews, or extracting person's opinion while diminishing the need for a human effort to classify the information. With the purpose of identification and extraction of subjective information from the materials' source opinion mining applies processing of original language, computational linguistics, and analytics of text as well. [33]

2.11 Types Opinion Mining

Where the goals include enabling computers to recognize and express emotions. Classification of the oppositions of a text given is considered to be an essential task of opinion mining types to find sentiments whether the expressed opinion in (positive, negative, or neutral) and classify those to three levels accordingly. [34, 35]

- **At the Document Level:** the entire document is contemplated to be a sole item accordingly, the whole document gets analyzed at a time in this method. Sometimes the outcome given by this approach is inappropriate. A document that is positively opinioned about an entity doesn't implicate that the author has only positive reviews about all the feature of that particular entity likewise, a document that is negatively opinioned about an entity doesn't signify that the author is wholly negative for all the features of that entity. In the conventional opinioned text the author expresses both positive and negative opinions about the entity and its attributes. [4]
- **Sentence level:** in this method, the document is broken into sentences and then each sentence is treated as a single entity and a single sentence is analyzed at a time. The results generated by this approach are better than those of document level and are more refined. The majority of present techniques, attempt to establish the overall polarity of document, paragraph, sentence or text regardless of the entity being expressed. [35]
- **Entity Level / Aspect based:** the main advantage of this type of sentiment analysis is that there is a possibility to get the opinions about the interesting feature of the product. Different features may have different sentiment responses. For instance, a phone has good camera quality, but the screen is small. Sentiment analysis is useful in many areas like in social media, etc. [34]

2.12 Polarity Detection

The polarity detection (also called polarity Classification or sentiment polarity Classification) is one of the most important tasks in Sentiment Analysis. Assuming that the overall opinion in an input document is about one single topic, polarity detection can be defined as (two, three, five) class Classification task and many types depending on the way the lexicon is categorized and sentiments such as:

- **Two-Class Classification:** The task is defined as a binary Classification which assigns one of two possible labels to each document of the input data (positive or negative). [33]
- **Three Class Classification:** The number of classes also considers the neutral such as class which is characterized by the absence of opinion on the input document or by the presence of opposite opinions which return a neutral polarity value when they are aggregated. [36]
- **Five-class Classification:** This task considers two degrees of negative, the neutral, and two degrees of positive as classes. This type of work can also be considered as rating prediction task in which the classes: very negative (-2), negative (-1), neutral (0), positive (1), and very positive (2). [37]

2.13 Emotion Mining

It has been believed by theorists who study basic emotions that human beings possess a narrow range of primary individual emotions. It has been attempted by various researchers (in particular, Paul Ekman and his coworkers) to determine a set of primary emotions being general for all people and there is an important difference between them, they came to conclusion that there exist six primary emotions (surprise, disgust, anger, happiness, fear, and sadness). [30] Nevertheless, the theorists have not come to the agreement regarding the contents of those primary emotions set. Furthermore, distinguishing one emotion from the other is a matter of dispute within emotions researching field. For example, since “surprise” can obtain positive, negative or even neutral sentiment, it is not clear whether it should be classified as an emotion. [38]

2.14 Java language

Java® is a general purpose, circumstantial and objects oriented language used for programming. The syntax of that is akin to C and C++, but it leaves out a lot of the properties which cause C and C++ to be complex, perplexing, and insecure. At the development stage, this platform has basically had a different purpose - to fix the flaws of building software for devices of a consumer being networked. It was designed for support in various host architectures and allowing software elements to be delivered safely. Compliance with such conditions requires a compiled code to ensure transportation through networks, be able to run on any client and also authenticate the operating safety of that client. The aforementioned attributes became even more appealing after internet popularization. Net surfing and accessing media enriched content became available for millions of people through a web browser window. [39]

2.15 Oracle VM Virtual Box

Virtual Box is a strong cross-platform program which has a virtualization software based system. The term "Virtualization Software" implies the ability to initiate and operate several Virtual Machines, which can run distinct operating systems, at the same time on the same computer. For example, it installs on your current Intel or Cyrix-based computer, regardless of their operating system: Windows, Mac, Linux or Solaris, and it enables you to operate Linux and Solaris on your Windows PC, Windows on you Linux systems, or operate Windows and Linux on your Mac. [40]

These units are called Virtual Machine (VM) and in a window of the "host" operating system, it is possible to develop and operate a "guest" operating system. A self-contained environment is provided by the VM which provides an opportunity to experiment with new software without a risk of making changes that can damage the host operating system.

CHAPTER 3

RESEARCH METHODOLOGY AND FRAMEWORK

3.1 Overview of Methodology

We designed and implemented Flume on Hadoop HDFS to enable immediate flow data Twitter to HDFS in real time, [19] and also apply Mahout to a top of Hadoop by Java that has a scalable classification of content on Twitter. The main design requirement for diagnoses of opinion mining is to construct highly accurate models as well as enable efficient classification on large-scale datasets. Finally, we describe the machine learning algorithm (supervised learning and unsupervised learning) that has been used in extracting the experiment and in detecting the sentiment polarity of the tweets for the test data set after applying them to the processed training corpora by using the supervised machine learning and unsupervised machine learning algorithm with lexicon for text categorization.

3.2 FRAMEWORK

In In the design shown in figure (8), we see the technical structure of this project. Basically, we have a list of tweets obtained from the Twitter API and a dictionary (what I need). From this point, we know which tweets have words from this dictionary and calculate a tweet score that will be the sum of the scores of each of those words that include a sentiment. After this started, Training data sets then Pre-Processing, Feature Selection will be done.

We have a lexicon for all word and all word have different values, To know the tweet is positive, negative and neutral and Figure (8) Illustrates Overview of the framework.

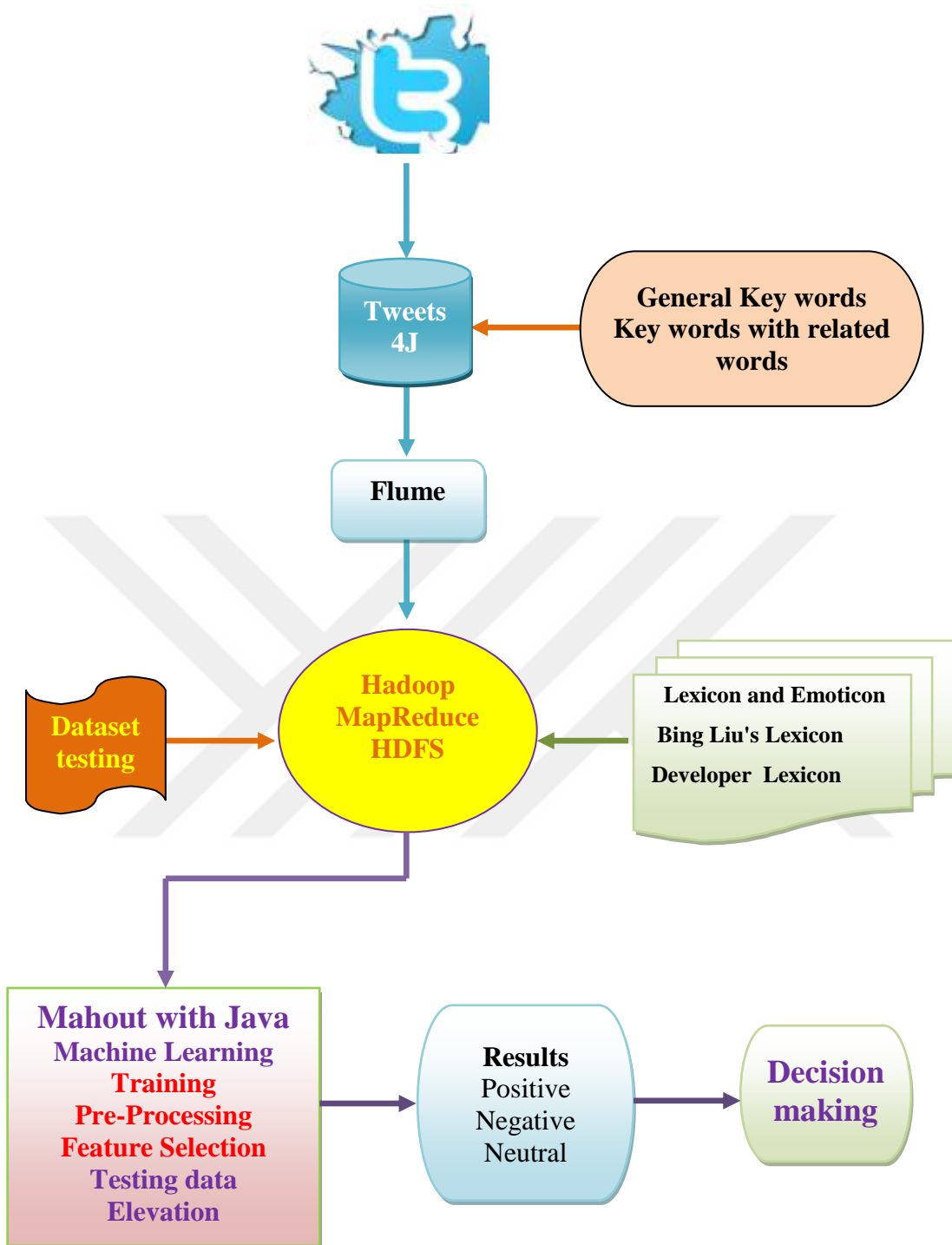


Figure (8) Overview of the framework

3.3 Data collection/Collecting Tweets

Before starting to write code it was necessary to decide which tool to use. Something able to process big quantities of data without wasting too much memory is needed. Saving millions of tweets with related and only tweets with related are needed, so the approach works as follows. [41]

- **Keywords general:** this set contains potentially relevant tweets. They are identified using the general related terms shown in Table (1). Intuitively, it includes any tweet, which matches one or several of the given keywords.
- **Keywords Related:** the positive tweets are tweets related to airlines and these are selected by having a precise Boolean regular expression such as the plane, passenger etc., more examples can be seen in the table (1).

Table (1) Illustrates contents examples key words

Key words	General 1	Related 1	Related 2	Related 3	Related 4
Virgin America	@Virgin America	Virgin America Airlines	Virgin America travel	Virgin America flight	Virgin America passenger
United	@United	United Airlines	United travel	United flight	United passenger
Southwest	@Southw est	Southwest Airlines	Southwest travel	Southwest flight	Southwest passenger
Delta	@Delta	Delta Airlines	Delta travel	Delta flight	Delta passenger
US Airways	@US Airways	US Airways Airlines	US Airways travel	US Airways flight	US Airways passenger
American	@Americ an	American Airlines	American travel	American flight	American passenger

3.3.1 Track Parameter value

A comma-separated list of keywords which will be used to specify which Tweets will be taken on the stream, Keywords may be one or more words separated by spaces, and keywords will match if all of the words in the keywords are present in the Tweet, regardless of order and ignoring case, we can follow the tables (2) to take exactly what is needed. [42]

Tables (2) Illustrates Track examples

Parameter value	Will match...	Will not match...
Twitter	“Twitter” twitter. #twitter @twitter	TwitterTracker, #newtwitter
Twitter’s	I like Twitter’s	I like @Twitter’s office
twitter api, twitter streaming	The Twitter API is awesome The twitter streaming service is fast Twitter has a streaming API	I’m new to Twitter
Delta.com	I will travel in Delta.com	I will travel in Delta.com/ foobarbaz
Delta.com/ foobarbaz	Delta.com/foobarbaz www. Delta.com/foobarbaz	Delta.com

3.3.2 Locations Parameter

We can find separated by commas list of longitude-latitude pairs to specify a group of bounding boxes for Tweets filtering, whether Tweets which fall within bounding boxes requested include:

Each bounding state or city is to be described as a combination of longitude and latitude pairs, this will help to identify tweets accurately and according to the city by using Flickr in website <https://www.flickr.com/places/info/1> in use own code to get started with using geo-information. [43]

3.4 Training Corpora

After checking and running the program, it was found that a dataset is necessary to ensure that the data collected is accurate and comparable to ensure reliability. therefore, data has been taken from the website: <https://www.crowdfunder.com/data-for-everyone/>. The total of number (11000) tweets and these tweets are reliable and tested (positive, negative and neutral).

3.5 Data Pre-Processing

Pre-processing in general, once the Twitter data has been fetched and stored in the directory on HDFS, but this data was raw and unstructured. When research became more focused on Twitter, the need of pre-processing increased proportionally because Pre-processing of Twitter data is very important and many tweets are not properly formatted or contain spelling errors. [44] Subsequently, before starting with the analysis, the data must have been cleaned in order to remove unnecessary information and make the data less noisy.

3.5.1 URL/ HTTP Removal

It is necessary to clean Twitter's data as tweets hold a couple of syntactic properties which might not be used for analysis. [45] In this process, any kind of unnecessary text from tweets being not substantial for our analysis, were removed these include, removing the Uniform Resource Locator (URL) that is commonly used for Hypertext Transfer Protocol (HTTP).

3.5.2 Repetition of Letters Removal / Exaggerated word shortening

Remove repetition – remove all duplicates words or letters from the text so that there will be no repetitions. [45] Thus, words containing the same letter more than twice and absent from the lexicon get shortened to the word that contains the repeated letter just once. For instance, the hyperbolic word “YESSSSSSSSS” gets shortened to “YES”. [46]

3.5.3 Stop words Removal

Java Natural Language Toolkit has a library containing a stop word dictionary to remove the stop words from each text. [47] We check each word in the tweets with words against the dictionary. If a word is a stop word, we filter it out. The list of stop words contains some prepositions and other words that add no sentimental value and which provides no meaning for classification like ‘a’, ‘the’, ‘to’, ‘at’, ‘he’, ‘what’, ‘or’, ‘able’.... etc.

3.5.4 Hashtag Removal

Removes the Hashtag or remove # in tweets, because the Hashtag is not useful at the stage of data analysis and opposite stage in the process of data collection where it has a key role and effectively parallels like the keywords.

3.5.5 Punctuations/ Digital words Removal

Some tweets have numbers or date and these numbers are like ‘1990’, ‘4:00pm’, but these do not have a meaning in the analysis stage and the better to give up to this to make results better and more accurate.

3.5.6 Tokenization

Given input as a character sequence, tokenization implies to a task of dividing it into parts named “tokens” as well as removing specific characters.

3.5.6.1 Whitespace tokenized

Used a string tokenizer to construct words from a sequence of characters, the input for this activity is the actual review whitespace tokenized simply breaks the sentences and divides the text on any sequence of whitespace, tab, or newline characters. For example “I saw a movie, it is nice and songs are very good”. After applying whitespace tokenized output is “I, saw, a, movie, it, is, nice, and, songs, are, very, good”. [48]

3.5.6.2 Stemming and lemmatization

While doing opinion mining, it was found out that due to grammatical reasons the sentences may contain different forms of a word and in addition to these, there are a lot of derivationally related words which possess similar meaning and these types of words can be analyzed as the same word. The iterated Lovins stemmer algorithm is considered to be the first published stemming algorithm and it contains a broad range of endings, conditions and transformation rules. [49]

Therefore, the aim of both lemmatization and stemming is to reduce these types of forms and reduce the word from its derivational form to its original base form by using Stand Natural Language Processing (NLP), and Table (3) illustrates stemming.

Table (3) Illustrates stemming and lemmatization

	Derivative word1	Derivative word2	root word
1	girls	girl's	Girl
2	decorates	decorating	Decorate
3	automaticity	automatically	Automatic
4	happiest	happier	Happy

3.5.7 Text Correction

First, every word in a tweet is evaluated by Language Tools Java (LTJ). If the evaluated word is misspelled, then the function gives probably suggestions for a correction. Finally, a suggested word with the highest probability will be selected as the corrected word.

3.5.8 Neutral Labels Removal

Remove @ - implements an option to remove @ symbol, remove @ including the user name, or substitute the @ and the username using letters: 'AT_USER' and add it to stop words.

3.6 Feature Selection

Each tweet has up to 140 characters, and these characters are a set of words, each word has a value and features and this leads to a specific polarity for each tweet and this is important in opinion mining and some properties must be extracted into opinion mining process, including:

3.6.1 Negation Features

Handling negation can be an important concern in an opinion-related analysis. The only differing token in the tweet, attaching “not, neither, nor...ECT” to tweets occurring to the negation term. [50] Negation plays a vital role in altering the polarity of the accompanying adjective and hence the polarity of the complete text. [51]

3.6.1.1 Unigrams Model with Negation Features

A unigram is simply an N-gram of size one or a single word and extracting unigrams from a tweet is the easiest way to get a feature. For each unique word in a tweet is processed, we simply parse a tweet into words, and every single word is a unigram, a unigram feature is created for the classifier. [47] For instance, when a negative tweet contains the word (not), a feature for classification would be found or not a tweet contains them a word (not). Since the word came from a tweet, the classifier would be to classify tweets containing the word as negative like the example. [52, 53]

The book is good (Pos.).

The book is bad (Neg.).

3.6.1.2 Bigrams Model with Negation Features

Use bigrams to parse from our tweets as features to develop our classifier. [52] Bigrams are evaluation consecutive words and bigrams are more eligible to handle negated phrases which add to the accuracy of the classifier by distinguishing words. [53] One possible solution to handle negations is to reverse the polarity of the adjective happening after a negation word. [51] Also choosing the right dimension ability to capture the opinion expression patterns. [47] For instance, if parsing such phrases with unigrams, we can't accurately get their opinions.

Today is not a bad day (Pos.).

Today is not a good day (Neg.).

3.6.2 Part of Speech Features

We use Part of Speech Tagging in to increase degrees of success of opinion mining, and it is a method of marking up a word in a text corresponding to a particular part-of-speech, also known as word class or lexical category, within a text, all words are classified to their corresponding lexical category. [54] We use this information; certain features can give more best weighted like.

3.6.2.1 Split Words or Conjunctions

Using some the words for splitting sentences into clauses like (and, but, because, so.... etc.). [55] The split words list consists of conjunctions words that are given as extra to best analysis. For instance, the complex sentence "The camera is good, but the battery is bad." contains two clauses:1. "The camera is good" and 2. "The battery is bad".

3.6.2.2 Interjection

In parts of speech exist short exclamation words which come at the beginning of the sentence like Oh, Ouch, Hey, My God or Ah! They have no real grammatical value, but these have a strong value and we use them often in speaking more than in writing for example, "Ouch! Camera is good! Or My God! Battery is bad".

3.6.3 Emoticon Features

Emoticons are useful features for emotion Classification and are known as the polar in a text message since they portray a writer’s emotion through icons. Figure (9) shows how it is displayed inside the text.

These properties seem to be commonly used for opinion mining by taking emotions from tweets and replacing them with their specific values. [45] There are lots of emoticons which are able to express sadness or happiness. To visualize, in many of emoticons in Table (4).



Figure (9) Illustrates Emotions in Text

Table (4) Some of Emotions Positive and Negative

Positive Emotions	:): D =D =):] =] :-):-D:-]); D;] ;-):-D;-] –
Negative Emotions	:(=(:[=[:-(-[:?([D,)-: , , =(,]: , :[, :(

3.6.4 Opinion Classifier Features

At the beginning of opinion mining for finding features in tweets, the tweets are broken down and assigned a polarity, there are three opinions in classes of sentiments, i.e. positive, negative and neutral sentiments. [56, 57, 58]

3.6.4.1. Positive Opinions

Positive tweets are the tweets which show a good or positive response towards something and this refers to a positive attitude of the speaker about the text. Emotions with positive opinion reflect happiness and smile, etc. In the case of airlines reviews, if the positive reviews/opinions about the passenger or customer prevail, it means people are happy with their services.

3.6.4.2 Negative Opinions

Tweets have a negative opinion or show a negative response or oppose towards something refer to a negative attitude of the speaker about the text. Emotions with negative sentiments reflect hate and sadness, and so forth. In the case of airlines reviews, if the negative reviews/opinions about the passenger or customer prevail, it means people are not happy with their services.

3.6.4.3 Neutral Opinions

Tweets don't have any opinions and neither oppose or support, depreciate or appreciate anything if there are no emotions reflected in the text. It is neither neglected nor preferred. Although this class doesn't mean anything, it is important for better isolation between positive and negative classes.

3.7 Machine Learning

Machine Learning (ML) is one of the types of artificial intelligence due to its use of computer learned experience with the program. Machine learning approach uses the data to determine the patterns within user's data and adjust the program actions accordingly. [59] In addition, ML approach makes use of standard ML algorithms to classify Sentiment Analysis and Opinion Mining as to a standard text classification, like tweets by making use of syntactic and linguistic characteristics. [60] It consists of two type's techniques: supervised learning and unsupervised learning.

3.8 Supervised Learning Algorithms

Supervised learning is an approach to learning from an available data set and making predictions for future cases. [61] Such models are produced using an array of training data. The classifier is then trained using one of the standards supervised learning techniques algorithms, e.g. Naive Bayes Classifier, Maximum Entropy and Support Vector Machines (SVM). The tweets are classified according to polarity features. [37, 60] To perform supervised learning there must be the idea in the connection between the input and the output. [58, 61] After training completion, the classifier predicts an instance label of a class according to the chosen features, so it is difficult to deal with Twitter data because there is no unified context. In this case, unsupervised learning algorithms need to be used to better determine the polar.

3.9 Unsupervised Learning Algorithms

Unsupervised learning algorithms can draw conclusions from data in tweets. [27] Structure is derived by clustering the data based on relationships among the variables in the data. Therefore, unsupervised opinion classification techniques are based on lexical. [59] Each word of a given document is examined for its polarity by looking it up in a special lexicon. [37] Also, classification polar from the text using an unsupervised algorithm, then the degree of polar for that lexicon can be measured by the indicators in a text. Therefore, unsupervised methods such as clustering in Mahout, it collects relevant data in clustering within a specific area depending on the type of relationship. This classification is useful when training data is not available or challengeable to be found; [27] which is then used as an input for supervised learning based algorithm.

3.10 Lexicon based approach

Lexicon-based is unsupervised learning method. Such approach relies only on opinion lexicon for identification, the lexicon approach assumes that opinion is related to the presence of certain phrases or words in the document that is a concept of dividing the opinion into more than one type by a dictionary and determine tweets, i.e. positive, negative or maybe neutral. [62]

Opinion Lexicons represent dictionaries containing opinion terms as well as semantic orientation related to these. The semantic polarity (or orientation) represents the way of the word changing its way from the other words into its group. Consequently, words expressing inside of a group hold a positive orientation, while words expressing outside of a group have a negative orientation or inside of a negative group, meaning this the opinion using these words from the lexicon that are present or are not present in the document. [63]

There are various lexical types available for opinion mining and these describe into this section to find sentiment. Figure (10) illustrates the workflow of opinion mining.

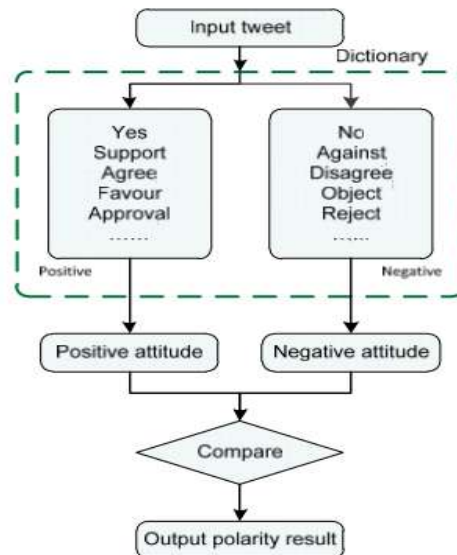


Figure (10) Illustrates the Workflow of Opinion Mining

3.11 Binary models Lexicon

A Defined task is a binary classification which assigns one of two to each tweet of the input data. [64] A lexical binary devised to support opinion analysis. It provides an annotation based on two numerical opinion scores (positivity or negativity). Some researchers consider this classification to be within three-class classification, by considering the neutral as the third class. [65] class which is characterized by the absence of opinion on the input document. [37] or by finding the opposite opinion in tweets which return a neutral polarity when they collect value, these lexicons will be adopted.

3.11.1 Bing Liu Lexicon

Liu is more attuned to sentiment expressions in tweets because it has undergone manual extraction from opinion sentences contained in people's reviews. [37] It holds some words that are strange or do not exist in other dictionaries like misspellings and other tweets expressions.

Bing Liu Lexicon is publicly available containing nearly 6,800 words (2,006 words positive and 4,783 words negative). [66] The dictionary has been developed by researchers over the past few years and it is considered to be an important resource used for identifying emotions as positive and negative terms in English. [67] In addition, table (5) includes some words explained by Bing Liu Lexicon.

Table (5) Includes some words from Bing Liu Lexicon

	word	Polar according to Liu	word	Polar according to Liu
1	awesome	Positive	absence	Negative
2	clear	Positive	bad	Negative
3	cool	Positive	broke	Negative
4	cute	Positive	weak	Negative
5	good	Positive	upset	Negative

3.12 Rules Testing Data

After applying the process of polar extraction of the lexical, it is important to know better accuracy and lower error rate, this is the best way to measure performance for Data Sets, and Table (11) illustrates testing data.

Table (6) Illustrates the testing data

Polarity	Positive Prediction	Negative Prediction
Positive Dataset	True Positive (A)	False Negative (B)
Negative Dataset	False Positive (C)	True Negative (D)

1- The error ratio for Lexicon is calculated by the following equations:

- Recall OR Error Rate for all Features = $\frac{(C+B)}{(A+B+C+D)}$
- Error Positive = $\frac{A}{(A+C)}$
- Error Negative = $\frac{D}{(B+D)}$

2- The accuracy of Lexicon is measured by the following equations:

- Precision OR Accuracy for all Features = $\frac{(A+D)}{(A+B+C+D)}$
- Positive Predictive Accuracy = $\frac{A}{(A+B)}$
- Negative Predictive Accuracy = $\frac{D}{(C+D)}$

CHAPTER 4

SETUP, FINDINGS and RESULTS

4.1 Overview of SETUP

This chapter describes findings of the assumptions obtained by using the methodologies machine learning demonstrated in the previous chapter to find the results. Firstly, the practical setup to carry out the experiments to study and classify tweets by specific words and these tweets are collected only by a specific domain.

The second, implement of training data such as the pre-processing operations to remove all not necessary things and extract features. Then, classify tweets using machine learning algorithms (supervised and unsupervised) with the ability to analyze the texts of tweet microblog to detect emotions by different types of the lexicon.

After that, provide the objectives which were collected from the achieved results by an accuracy of classification. Finally, use the cluster in Mahout to collect data at same polar to determine the best service or product.

4.2 Data FRAMEWORK

The focus of this thesis is opinion mining in the context of airlines, As Twitter accommodates an infinite number of opinionated tweets from a vast and varied users' area containing tweets on almost any topic conceivable. The fundamental purpose of selecting Twitter as the source of information for completing this thesis was to determine whether it is possible to use Twitter data for developing airlines system of recommendations. Thus, it is necessary to apply the steps to get tweets accurately, step one a lexicon of keywords that are commonly used words and related words in twitter about airlines. The lexicon was compiled from personal experience by studying around airlines as was explained in Chapter 3. Figure (11) illustrates keywords in the system.

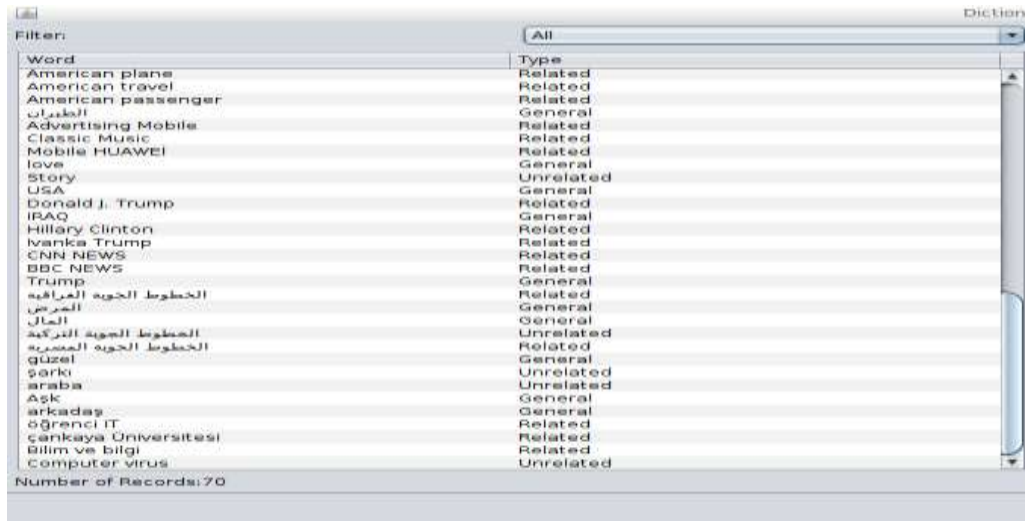


Figure (11) Illustrates Key words in system

Before beginning to combine, we note that there is a need to determine the geographical location of the tweets we want to get, as was mentioned in Chapter3, we also determine for each group of keywords and location that is called tasks and the user can store multiple tasks with many keywords in many subjects and different locations. When the user wants to perform previous tasks he returns to the address and executes these tasks, and Figure (12) illustrates tasks choosing a keyword with location.

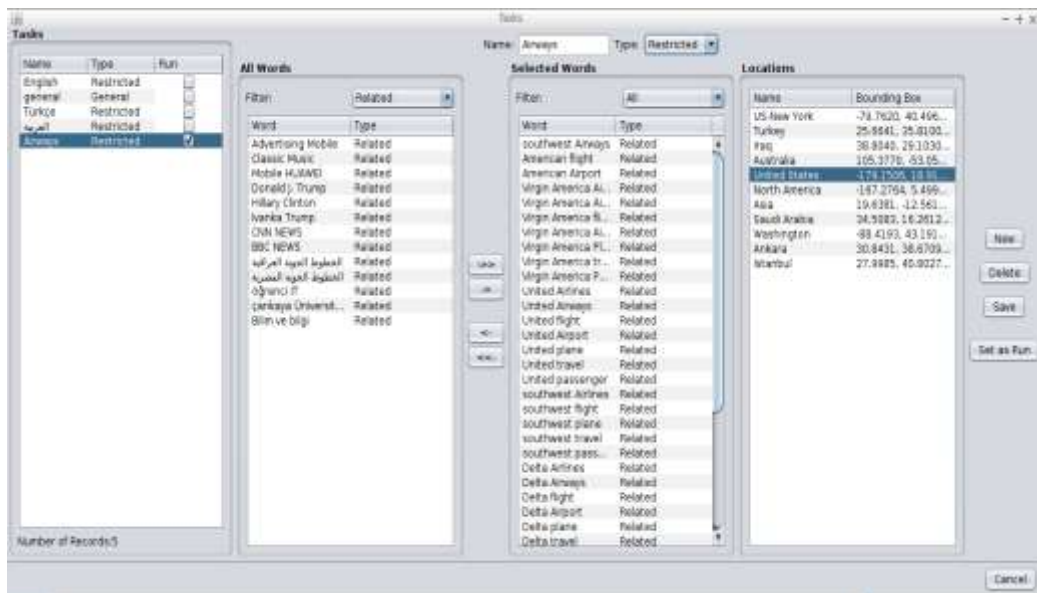


Figure (12) Illustrates Tasks choose key word with location

4.3 Hadoop Implementation

This describes the required procedure for setting up a NameNode Hadoop cluster backed by the Hadoop Distributed File System, operating on Ubuntu Linux, and our own DataNode to store data in Hadoop. [69] That will represent the core part of the work and data warehouse.

To create Hadoop by the virtual machine in Ubuntu by using some commands in MapReduce, we need writing below commands to build Hadoop cluster, and Figure (13) illustrates how to create NameNode in Hadoop.

```
./sbin/start-dfs.sh  
./sbin/start-yarn.sh  
./bin/hdfs dfs -mkdir /user  
./bin/hdfs dfs -mkdir /user/Hadoop
```

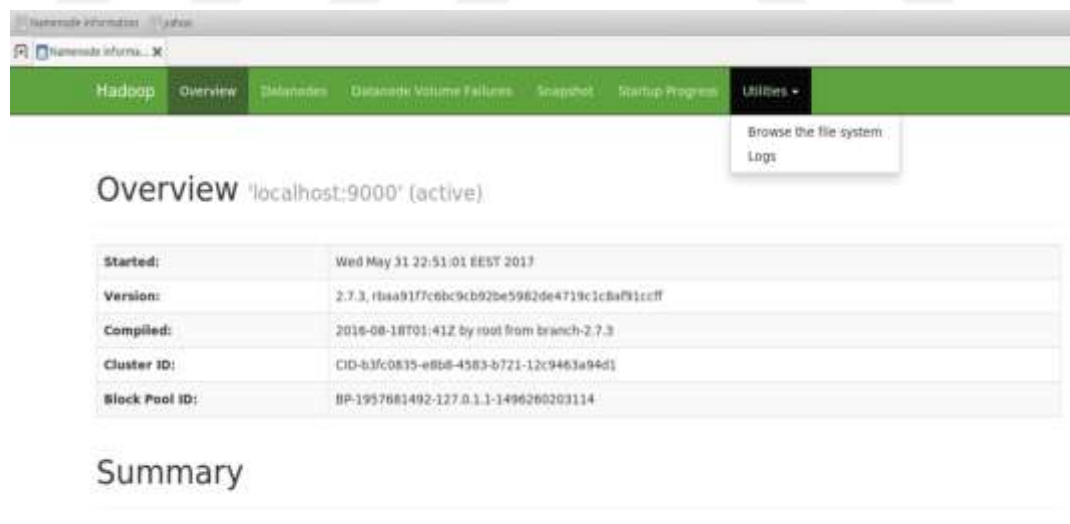


Figure (13) Illustrates create NameNode in Hadoop

4.3.1 A Flume Connection

By a virtual machine (VM) with MapReduce working that hosts the tweets stream a large amount of real-time data that will come inside Twitter through and then flow to HDFS real time [69], we need these commands below to build Hadoop and figure (14) explains the result of connection between Twitter and Hadoop.

```
./bin/hdfs dfs -mkdir /user/Hadoop/twitter_data
```

```
./bin/hdfs dfs -copyToLocal /user/Hadoop/twitter_data /home/hduser/Downloads
```

```
cd /home/hduser/ProgramFiles/apache-flume-1.7.0-bin
```

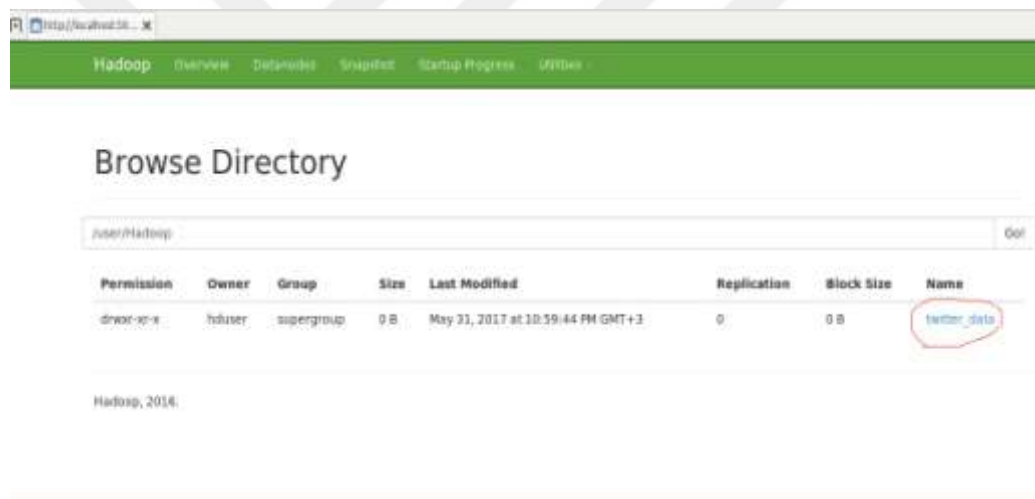


Figure (14) Illustrates connect between hadoop, flume and twitter.

4.4 Data streaming at Real Time

We note after the completion of the connection process begins data streaming by Flume and Twitter4j with same keywords and specific location, and figure (15) explains all operation then automatically store the tweets in Hadoop or HDFS, figure (16) proves it, and below written commands are needed to start.

```
./bin/flume-ng agent --conf ./conf/ -f conf/twitter.conf
Dflume.root.logger=DEBUG,console -n TwitterAgent
```

```

hduser@my-VM-Lubuntu: ~/ProgramFiles/apache-flume-1.7.0-bin
...
ited plane,United travel,United passenger,southwest Airlines,southwest flight,southwest plane,
southwest travel,southwest passenger,Delta Airlines,Delta Airways,Delta flight,Delta Airport,D
elta plane,Delta travel,Delta passenger,US Airways flight,US Airways Airport,US Airways plane,
US Airways travel,US Airways passenger,American Airlines,American Airways,American plane,Ameri
can travel,American passenger.
keywordsNegative:
ListSettings:
Locations:2
MyFlumeTwitter Starting up Twitter filtering...
[Thu Jun 01 00:34:23 EEST 2017]Establishing connection.
[Thu Jun 01 00:34:37 EEST 2017]Connection established.
[Thu Jun 01 00:34:37 EEST 2017]Receiving status stream.
** GEO:GeoLocation(latitude=32.7554883, longitude=-97.3307658)
** PL:PlaceJSONImpl(name='Fort Worth', streetAddress='null', countryCode='US', id='42e46bc3663
a4b5f', country='United States', placeType='city', url='https://api.twitter.com/1.1/geo/id/42e
46bc3663a4b5f.json', fullNames='Fort Worth, TX', boundingBoxType='Polygon', boundingBoxCoor
dinate=[[LLtwitter4j.GeoLocation@1f0af527], geometryType='null', geometryCoordinates=null,
containedWithin=null)
** LAN:en
** SAVED TWEET** <user>=tmj <dfw_eng> <tweet>= Want to work at American Airlines? We're #hiring in
#FortWorth, TX! Click for details: https://t.co/zYyJ2E2sLJ #Aerospace #Job #Jobs
** PLACE ** PlaceJSONImpl(name='Fort Worth', streetAddress='null', countryCode='US', id='42e46
bc3663a4b5f', country='United States', placeType='city', url='https://api.twitter.com/1.1/geo/
id/42e46bc3663a4b5f.json', fullNames='Fort Worth, TX', boundingBoxType='Polygon', boundingBoxCo
ordinates=[[LLtwitter4j.GeoLocation@1f0af527], geometryType='null', geometryCoordinates=null,
containedWithin=null)
** GEO:NULL
** PL:PlaceJSONImpl(name='Providence', streetAddress='null', countryCode='US', id='7b93be1d864
cedbb', country='United States', placeType='city', url='https://api.twitter.com/1.1/geo/id/7b9
3be1d864cedbb.json', fullNames='Providence, RI', boundingBoxType='Polygon', boundingBoxCoor
dinate=[[LLtwitter4j.GeoLocation@714bc85f], geometryType='null', geometryCoordinates=null,
containedWithin=null)
** LAN:en
** SAVED TWEET** <user>=ROJ036 <tweet>= While we're launching things at the sun @NASA can we als
o send the Delta Airlines Baggpipe ad from Snapchat too? (ah, https://t.co/NNte0830FS
** PLACE ** PlaceJSONImpl(name='Providence', streetAddress='null', countryCode='US', id='7b93b
e1d864cedbb', country='United States', placeType='city', url='https://api.twitter.com/1.1/geo/
id/7b93be1d864cedbb.json', fullNames='Providence, RI', boundingBoxType='Polygon', boundingBoxCo
ordinates=[[LLtwitter4j.GeoLocation@714bc85f], geometryType='null', geometryCoordinates=null,
containedWithin=null)

```

Figure (15) Illustrates data streaming in VM

Browse Directory

/user/hadoop/twitter_data

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-r--r--	hduser	supergroup	0 B	May 31, 2017 at 11:16:13 PM GMT+3	1	128 MB	FlumeData.1496261745700
-r--r--	hduser	supergroup	131 B	May 31, 2017 at 11:17:00 PM GMT+3	1	128 MB	FlumeData.1496261789218
-r--r--	hduser	supergroup	95 B	May 31, 2017 at 11:21:59 PM GMT+3	1	128 MB	FlumeData.1496262088489
-r--r--	hduser	supergroup	72 B	May 31, 2017 at 11:25:14 PM GMT+3	1	128 MB	FlumeData.1496262280998

Figure (16) Illustrates data storage in hadoop

4.5 Training data

After the data collection begins with the new stage or the second stage of filtering and removal, to cleaning data, as was mentioned in Chapter3. We start to choose the file content of all tweets which amount to approximately 11,000 tweets, then input file in the path to start the analysis process and pre-process to tweet one by one and Figure (17) illustrates input file in the path that quickly performs analysis process.

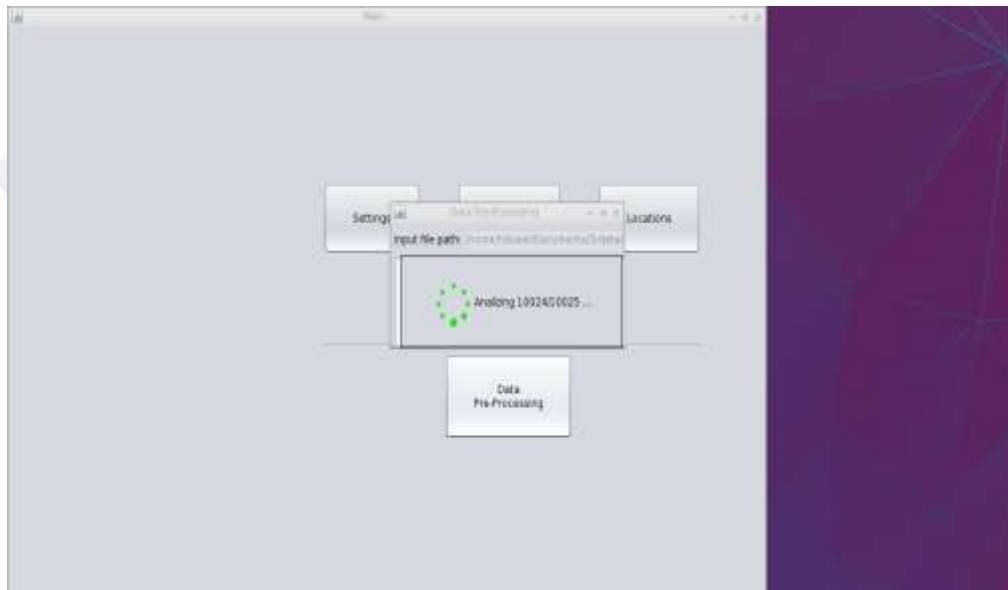


Figure (17) Illustrates input data in path

Then, we find the machine learning is Pre-Processing all data or Pre-Processing all of the tweets and remove something that has no value, and which drains time useless when starting opinion mining process like.

- Hashtag Removal.
- URL/ HTTP Removal.
- Repetition of Letters Removal or Exaggerated word shortening.
- Stop words Removal.
- Punctuations or Digital words.
- Neutral Labels @etc Removal.
- Whitespace tokenized.

Also, the clean data or tweets are not enough, **Text Correction** is also needed because we are dealing with informal information and different expressions of culture and educational level among Twitter users, and this may lead to spelling errors then we may lose important information about opinions. To resolve this obstacle, machine learning modern tools are used, those tools are called **Language Tools Java (LTJ)** it will correct spelling errors inside the tweets.

However, the process of opinion mining for finding sentiment depends on matching, i.e., when the analysis started, there was a difference between Try and Tried in the machine, but these words possess the same value, so we need to save the verb with all derivatives and this needs a lot of effort. To solve this problem, we use **Stanford Natural Language Processing (NLP)**. It turns it back to the verb by root and this operation is called **Stemming and Lemmatization** and steps of this process are shown in figure (18), and table (11) explains more clearly by comparison between first row-Before Pre-Processing tweet (Original data) and second row-After Pre-Processing tweet, then the third row- Removed or modified appear briefly, what was removed or changed is specified inside the parentheses ().

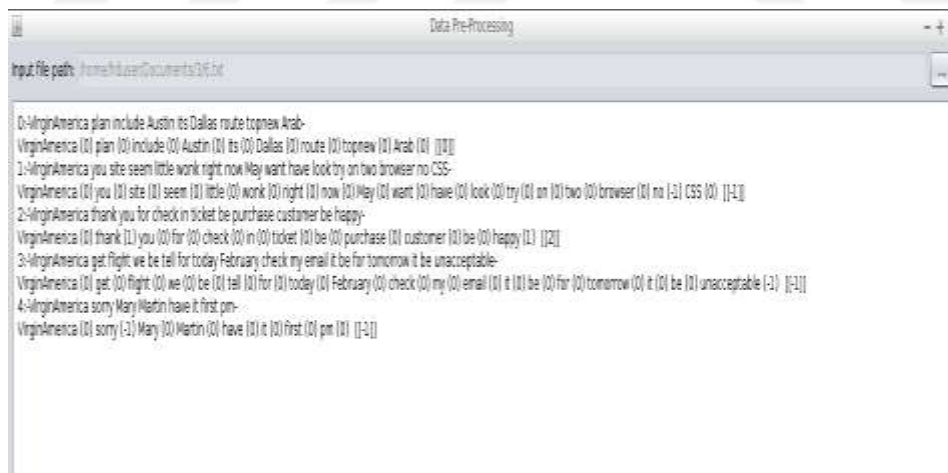


Figure (18) All pre-processing steps.

Table (7) Illustrates comparison between before and after Pre-Processing.

Processing	Tweets
Before Pre-Processing	VirginAmerica Plans to Include Austin to its Dallas Route TopNews Arab #Emirates #flight.
After Pre-Processing	0:-VirginAmerica plan include Austin its Dallas route topnew Arab-
Remove or Modified	Plan(S), to, to, TopNew(S), #Emirates, #flight.
Before Pre-Processing	VirginAmerica Your site seems a little wonked right now. May want to have a look. Tried on two browsers. No CSS? http://t.co/8qsQMM7KF2 .
After Pre-Processing	1:-VirginAmerica you site seem little wonk right now May want have look try on two browser no CSS-
Remove or Modified	You(R), seem(S), (A) little, wonk (ed), to, a, Tr(ied) , ?, http://t.co/8qsQMM7KF2
Before Pre-Processing	VirginAmerica thank you for checking in. tickets are purchased and customer is happppy.
After Pre-Processing	2:-VirginAmerica thank you for check in ticket be purchase customer be happy-
Remove or Modified	Check(ing), ticket(S), are, purchas(ed), and, (is), happ(ppp)y.
Before Pre-Processing	VirginAmerica got a flight (we were told) for today, February, 2017 checked my email and it is for TOMORROW. It is unacceptable.
After Pre-Processing	3:-VirginAmerica get flight we be tell for today February check my email it be for tomorrow it be unacceptable-
Remove or Modified	got, a, (), were, told, 2017, , check(ed), and, (is).
Before Pre-Processing	VirginAmerica @LadyGaga @CarrieUnderwood Sorry, Mary Martin had it farst at 4 pm!
After Pre-Processing	4:-VirginAmerica sorry Mary Martin have it first pm-
Remove or Modified	@LadyGaga, @CarrieUnderwood, have, (farst), at, 4, !

4.5.1 Polar Extraction

After Pre-Processing operations, the machine learning works to extract polar features by reading a value of each word (one by one), and figure (19) illustrates how the machine learning performs this operation.

This process is not complicated, but it is very important, it matches each word with all words within the lexicon to Polar Extraction of opinion.

- If this word matches any Positive word in the lexicon, this word is given **Positive (1)** value.
- If this word matches any Negative word in the lexicon, this word is given **Negative (-1)** value.
- If this word matches neither Positive or Negative word in the lexicon, this word is given **Neutral (0)** value.

Also, the same was applied to the extraction of opinion inside emotions because some emoticons can have positive or negative value, like words and there are some examples in the table (12) about polarity.

```
Input file path: /home/fuser/Documents/3/Negation.txt

0- now am sad-
now (0) am (0) sad (-1) [] -1[]

1- love travel-
love (1) travel (0) [] []

2- didn't pass in exam-
didn't (-1) pass (0) in (0) exam (0) [] -1[]

3- don't like exam-
don't (-1) like (1) exam (0) [] -1[]

4- didn't hate Blue-
didn't (-1) hate (-1) Blue (0) [] []

5- cannot travel in air plane now happy travel in car-
cannot (-1) travel (0) in (0) air (0) plane (0) now (0) happy (1) travel (0) in (0) car (0) [] [] []

6- mobile camera be good, screen be good battery be bad-
mobile (0) camera (0) be (0) good (1), (0) screen (0) be (0) good (1) battery (0) be (0) bad (-1) [] [] []
```

Figure (19) Illustrates Polar Extraction and Negation Extraction.

4.5.2 Negation Extraction

The negation within sentences is a complex process because it comes in multiple forms and methods and this a difficult step in the programming process and requires precision to get the correct results, it can be noted that in machine learning many ways of negation have been applied to take the results, and this is very important because it is possible to reflect the direction of the polar from negative to positive and positive to negative, and figure (19) displays Negation Extraction results with polarization in machine learning.

Several equations were applied to obtain these results and in order to obtain them more accurately, and for a detailed explanation of the processing operations, the table (12) explains all of the Equations and the roots of these equations.

Table (8) Illustrates Equation of polarization

	Tweets After preprocess	Value	Equation	Observation
0	now am sad-	now (0) am (0) sad (-1) [-1]	Unigrams model	Polar based on Dictionary (Negative)
1	love travel-	love (1) travel (0) [1]	Unigrams model	Polar based on Dictionary (Positive)
2	didn't pass in exam-	didn't (-1) pass (0) in (0) exam (0) [-1]	Bigrams model	Negation word does not follow anything (Negative)
3	don't like exam-	don't (-1) like (1) exam (0) [-1]	Bigrams model	Negation follow him Positive (Negative)
4	didn't hate Blue -	didn't (-1) hate (-1) Blue (0) [1]	Bigrams model	Negation follow him Negative (Positive)
5	cannot travel in airplane now happy travel in car	cannot (-1) travel (0) in (0) airplane (0) now (0) happy (1) travel (0) in (0) car (0) [0]	Part of Speech (conjunction)	1- Negation follows him Positive but Positive not dependent direct. 2-Tweet has two value (Negative – Positive) = (Neutral)
4	mobile camera be good , screen be good battery be bad	mobile (0) camera (0) be (0) good (1) , (0) screen (0) be (0) good (1) battery (0) be (0) bad (-1) [1]	Part of Speech (conjunction)	Tweet has number positive bigger than negative 2-Positive – 1 Negative) = (Positive)

4.6 Lexicon Developer and Statistics

It was found that number of words in Bing Liu lexicon is not sufficient, and this is a weakness detained during the process of matching because the lexicon of Liu includes 6800 words (positive and negative) and this number of words is small compared to the field of comment in tweets and under the shortcuts, randomness and new expressions.

Consequently, a new lexicon was developed by us as we decided to collect many lexicons and merge them together: **Bing Liu, AFINN, SemEval 2015, General Inquirer (GI) etc.**

The result, providing a lexicon that contains large amount of words, over 33,000 words for positive and 34,000 words for negative, is shown table (13) as the comparison of the lexicon of Liu and the dictionary developed by me in terms of words.

Table (9) Illustrates Lexicon Statistics.

Lexicon	positive	Neutral	negative
Bing Liu	2005	0	4783
AFINN	878	0	1597
SemEval 2015	776	13	726
General Inquirer	1629	0	2001
Other Lexicon	29345	0	24011
Lexicon Developer	34633	13	33118

4.7 Accuracy Measure and Evaluation

After completing all the stages: Pre-Processing, Polar Extraction, Negation Extraction and Lexicon Developer, the effectiveness and efficiency of the project must be known, this is achieved by comparing results with results obtained in a previous period and this data is checked manually or in a traditional way, as in previous period.

Now, the program will match and display the results when we check 1200 tweets (which represents 10% of the data ratio) we can find this results in figure (20) illustrates Comparison between the percentage of True and False.

- **True positive:** 218
- **False Positive:** 1
- **True Negative:** 634
- **False Negative:** 18

	Positive Prediction	Negative Prediction
Positive Class	218.0	18.0
Negative Class	1.0	634.0

Figure (20) Illustrates comparison between numerical ratio of **True** and **False**.

Also, we can show a comparison between of **True** and **False** like a chart, but this chart displays the data like numerical ratio and Figure (21) displays the details.

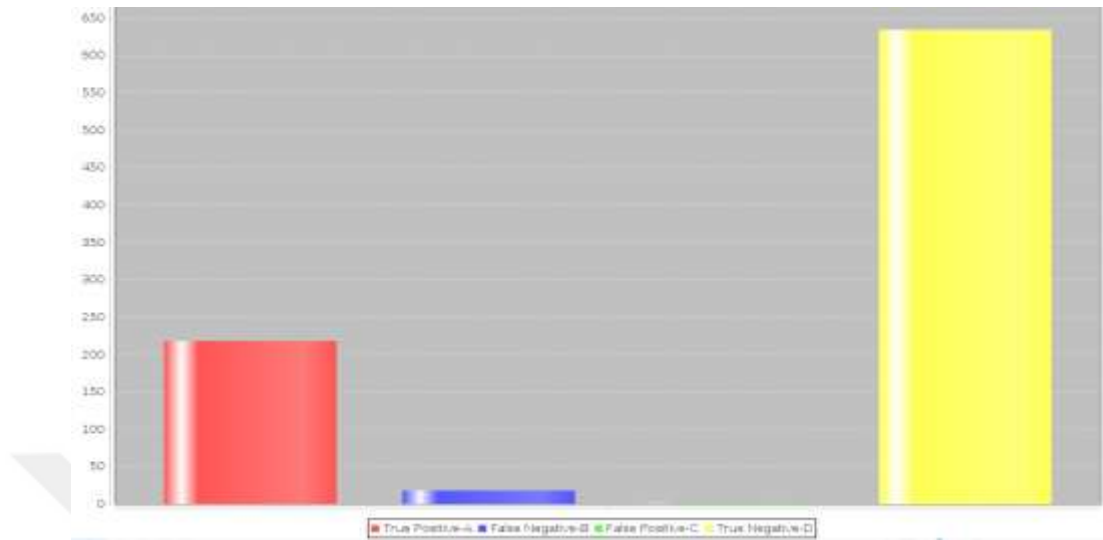


Figure (21) Illustrates chart comparison between numerical ratio of **True** and **False**

Now, after comparison between true and false, we show the **Accuracy, Error rate, Recall (Error Positive and Error Negative), Precision (Positive, Negative)**, we have achieved by applying all the previous steps.

- **Achieved accuracy (0.978186)**
- **Error Rate (0.021814)**

And many different accuracy ratio and figure (22) illustrates this accuracy

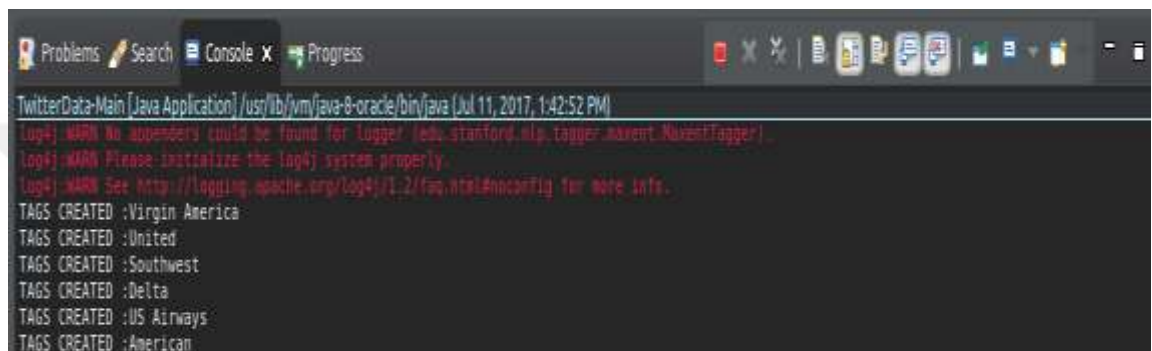
Error Rate E	0.021814
Acc	0.978186
Recall/Error Positive	0.995434
Error Negative	0.972393
Precision/Positive Predictive Acc	0.923729
Negative Predictive Acc	0.998425

Figure (22) Illustrates type of accuracy ratios

4.8 Target and making decision

Using clustering of the Keywords, Figure (23) shows the features in the query (K=6) by clustering all tweets which have the same key words through our experiment.

Feature: Virgin America, United, Southwest, Delta, US Airways, American.



```
TwitterData-Main [Java Application] /usr/lib/jvm/java-8-oracle/bin/java (Jul 11, 2017, 1:42:52 PM)
log4j:WARN No appenders could be found for logger [edu.stanford.nlp.tagger.MaxentNaveitTagger].
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
TAGS CREATED :Virgin America
TAGS CREATED :United
TAGS CREATED :Southwest
TAGS CREATED :Delta
TAGS CREATED :US Airways
TAGS CREATED :American
```

Figure (23) Query tags created to clustering by keywords

After the machine learning performed clustering, we are approaching the target of the thesis, predicting the decision - making process by finding what are the opinions that refer to keywords (product or service) is the best one, it has **a lot of Positive** and **a little Negative**.

When we start to analyze the data and find opinion mining about six keywords, we can easily see that the best service among the six airline companies analyzed in the US is **Delta** because Delta Company has **High positive and Low Negative** value. On the other hand, US Airways Company is the worst company because it has **Low positive and High Negative** value.

This is based on the opinions of customers of these companies and this lets us make the right decision and the right choice, we can show Figure (24, 25) for explaining the resulted tweets data like chart and Table (14,15) explaining the resulted tweets data in numerical value.

When we analyze only (1200) tweets for Decision-making and predicting the future, we can find the result in Figure (24) and Table (14).

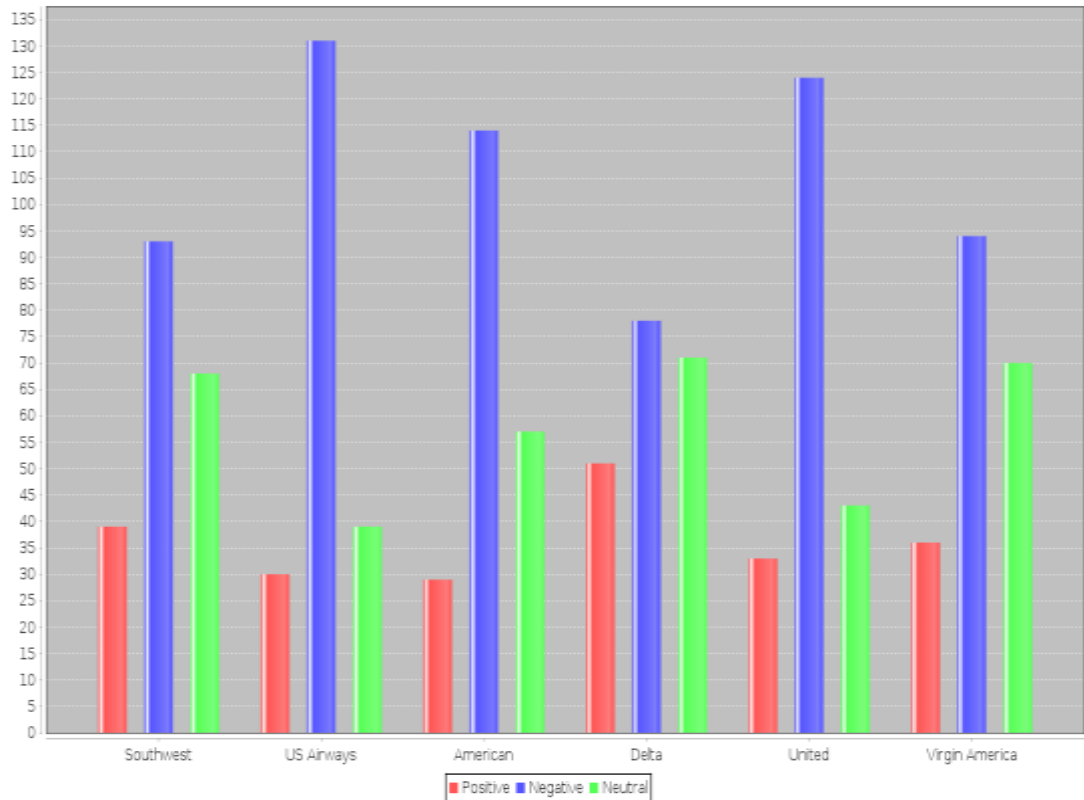


Figure (24) Illustrates analysis only (1200) tweets

Table (10) Illustrates analysis only (1200) tweets

Key word	Positive	Neutral	Negative
Southwest	39	68	93
US Airways	30	39	131
American	29	57	114
Delta	51	71	78
United	33	43	124
Virgin America	36	70	94

After analyzing all tweets, approximately (10025) tweets, for decision-making and predicting the future, we can find the results in Figure (24) and Table (14).

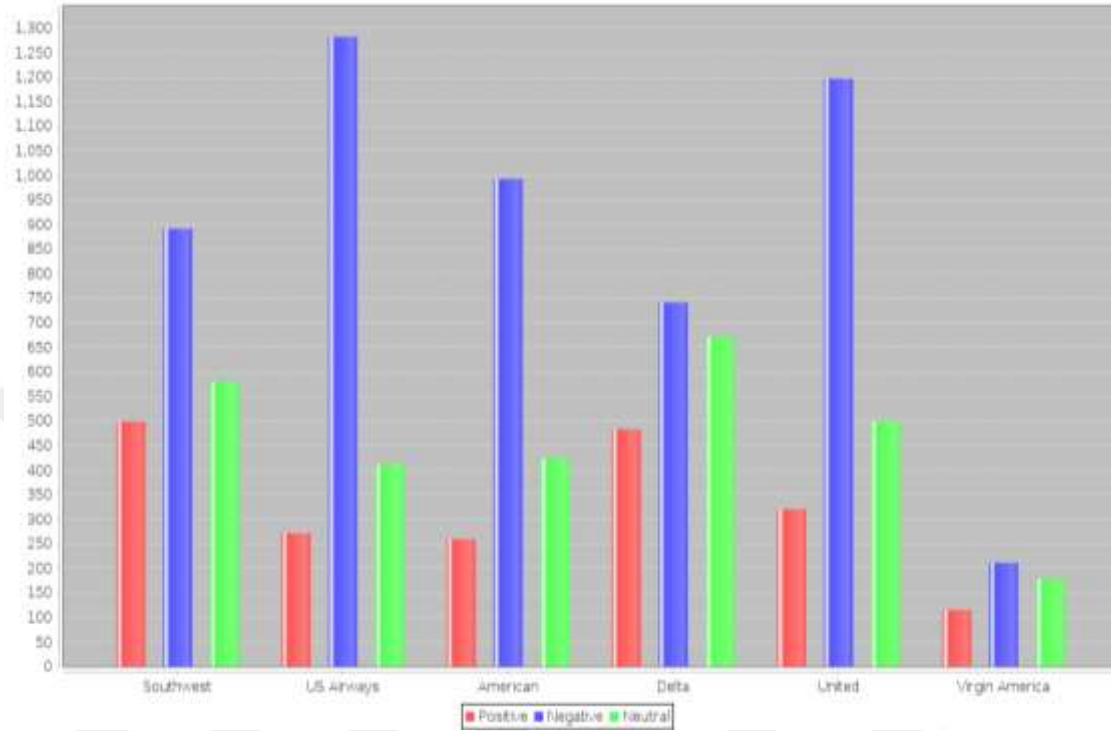


Figure (25) Illustrates analysis all tweets

Table (11) Illustrates analysis all tweets

Key word	Positive	Neutral	Negative
Southwest	499	580	891
US Airways	271	412	1282
American	259	423	993
Delta	489	671	741
United	320	499	1197
Virgin America	115	178	211

CHAPTER 5

DISCUSSION, CONCLUSION AND FUTURE WORK

This chapter consists of **two parts**:

The first part summarizes the findings in Section 5.1.

The second part provides some plan related to this study to improve and assure the prospective course of the given project in Section 5.2.

5.1 Discussion and CONCLUSION

These experiments showed the high potential of Hadoop and JAVA with machine learning by finding opinion mining at real time and predicting future of product or service and finding the best of them.

As mentioned in our objectives, our main goal was to combine accuracy and performance in terms of speed and time, what we want is exactly what we want in terms of data coming through keywords and this reduces us with the collection of millions of tweets of non-value data because all of this data needs a procession and time.

Also, a real-time opinion mining is applied towards different products and services, it helps us to follow and make decisions and choose the best in many areas streams. Twitter helps to understand what users' opinions about experiencing something are, and it can be applied in different areas, such as marketing, news monitoring, customer satisfaction about a product, prevention of disasters such as natural disasters or disease spread, or the expectation of the winner of the elections, etc.

5.2 Future Work

In this thesis, opinion mining was implemented using only one score of a lexicon, in the future lexicon with different scales needs to be used to see the best results because some words are positive and have a strong value and some have weak value, these can be improved.

Also, Hive in Hadoop ecosystem can be used to demonstrate the data like tables, if applicable of this case makes all data easy to handle with office software.



REFERENCES

1. **Eugene G. (2013)**, “*Cloud Computing Models*” Master of Management and Master of Engineering in conjunction with the System Design and Management Program at the Massachusetts Institute of Technology, Chapter 1, Sloan School of Management, Room E62-422 and Massachusetts Institute of Technology Cambridge, pp1-3.
2. **Garry T. (2013)**, “*Hadoop Beginner's Guide, Learn how to crunch big data to extract meaning from the data avalanche*”, Books, Chapter 2, pp36-45 .
3. **Hassan S., Miriam F., Yulan H. and Harith A. (Dec 12, 2013)**, “*Evaluation Datasets for Twitter Sentiment Analysis, A survey and a new dataset, the STS-Gold*”, pp1-7.
4. **Parth G., Dweepan G., Bakul P. (2014)**, “*A Performance Analysis of MapReduce Applications on Big Data in Cloud based Hadoop*”. INSPEC Accession Number: 14916053, Publisher: IEEE, Conference Location: Chennai, India.
5. **Haewoon k., Changhyun L., Hosung P., Sue M. (April 26–30, 2010)**, “*What is Twitter, a Social Network or a News Media*”, Department of Computer Science, KAIST 335 Gwahangno, Yuseong-gu, Daejeon, Korea, p1-2.
6. **Paul E. (1999)**, “*Basic emotions*” Handbook of cognition and emotion”, University of California, San Francisco, CA, USA, pp53-54.

7. "Company | About," Twitter About. [Online]. Available at website: <https://about.twitter.com/company> [Accessed at 15/7/2017].
8. **Chuck L. (2011)**, "*Hadoop in Action, MANNING*" Books, ©2011 by Manning Publications Co. All rights reserved, Printed in the United States of America, pp-27-38.
9. **Patlammagari G. (May 2014)**, "*Opinion Mining of Online Customer Reviews*" for the degree of Master of Technology, National Institute of Technology Rourkela.
10. **Prashanth V. (May 11, 2016)**, "*Automatic Identification of Representative Content on Twitter*" Master of Science in Media Arts and Sciences, MASSACHUSETTS INSTITUTE OF TECHNOLOGY.
11. **Amir H. R. (2015)**, "*Real-time Sentiment Analysis of Twitter Public Stream*" Master of Computer Science, University of Jyväskylä.
12. **Soren H. (February 18, 2011)**, "*Opinion Mining with Semantic Analysis*" Master in Computer Science, Department of Computer Science University of Copenhagen.
13. **Li C., Yangyong Z. (2015)**, "*The Challenges of Data Quality and Data Quality Assessment in the Big Data Era*" Data Science Journal, pp-2.
14. **Siddharth J., Rakesh K., Sunil J. (2014)**, "*A Comparative Study for Cloud Computing Platform on Open Source Software*", An International Journal of Engineering & Technology (AIJET), pp-29-30.
15. **Jisha S. M., Jisha S. (6 June 2013)**, "*A COMPARATIVE STUDY ON OPEN SOURCE CLOUD COMPUTING FRAMEWORKS*" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2, pp- 2025.

16. **HIBA J. H., AMMAR H. S., SARAH H., AZIZAHBT H. A. (Jan.-2015)**, “*BIG DATA AND FIVE V’S CHARACTERISTICS*”, International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835, Volume-2, PP18- 20.
17. "The Twitaholic.com Top 100 Twitterholics based on followers." Twitterholics [.http://twitaholic.com/](http://twitaholic.com/) . (Accessed 8/7/ 2017).
18. Java library for the Twitter API, <http://twitter4j.org/en/index.html> (Accessed: 15/7/2017).
19. Flume - Apache Software Foundation project home page <https://flume.apache.org> (Accessed: 15/7/2017).
20. **Sneha M., Viral M. (2015)**, “*Hadoop Ecosystem: An Introduction*”, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064, PP 561.
21. **Kevin S., Marshall P. (2/12/2014)**, “*A Field Guide to Hadoop, An Introduction to Hadoop, Its Ecosystem, and Aligned Technologies*”, Chapter1 Core Technologies, PP1.
22. **Jens D., Jorge A. (2012)**, "*Efficient big data processing in Hadoop MapReduce*" Information Systems Group, Saarland University Istanbul, Turkey, pp 2014-2015.
23. **John W., Sons C. (2012)**, “*Hadoop® FOR DUMMIES SPECIAL EDITION- Compliments of IBM Platform Computing IBM*” Chapter 2-3 , pp 17- 26.
24. **Tom W. (2015)**, “*Hadoop: The Definitive Guide - STORAGE AND ANALYSIS INTERNET SCALE -FOURTH EDITION*” Chapter 3, PP 43-77.

- 25. Konstantin S., Hairong K., Sanjay R., Robert C. (2010)** “*The Hadoop Distributed File System*”, Publisher: IEEE, Conference Location: Incline Village, NV, USA, USA, PP 2-3.
- 26. Hae-Duck J. J., WooSeok H., Jiyong L., Ilsun Y. (2012)**, “*Anomaly Teletraffic Intrusion Detection Systems on Hadoop-based Platforms: A Survey of Some Problems and Solutions*” Publisher: IEEE, International Conference on Network-Based Information Systems, Conference Location: Melbourne, VIC, Australia, PP 766-769.
- 27. Jiong X., Shu Y., Xiaojun R., Zhiyang D., Yun T., James M., Adam M., Xiao Q. (April 2010)**, “*Improving MapReduce Performance through DataPlacement in Heterogeneous Hadoop Clusters*”, Department of Computer Science and Software Engineering Auburn University, Auburn, PP 2-4.
- 28.** The Apache Software foundation, Hadoop Commands Reference, <https://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-common/CommandsManual.html> (Accessed: 15/7/2017).
- 29.** The Apache Software foundation, Apache Hadoop YARN, <https://hadoop.apache.org/docs/r2.7.2/hadoop-yarn/hadoop-yarn-site/YARN.html> (Accessed: 15/7/2017).
- 30. Ashish G. (2015)**, “*Learning Apache Mahout Classification*” [PACKT] open source Build and personalize your own classifiers using Apache Mahout, pp 64.
- 31. Chandramani T. (March 2015)**, “*Learning Apache Mahout, Acquire practical skills in Big Data Analytics and explore data science with Apache Mahout*” Chapter 2, pp 27.

- 32. Michael K., David A. (December 20, 2012)**, “*Sentiment Classification in Twitter: A Comparison between Domain Adaptation and Distant Supervision*”, Statistical Natural Language Processing Final Project, pp5-12.
- 33. Bo P., Lillian L. (2008)**, “*Opinion mining and sentiment analysis*” Foundations and Trends in Information Retrieval, pp16- 32.
- 34. Michelle de Haaff** “Sentiment Analysis, Hard But Worth It!” CustomerThink, retrieved http://customerthink.com/sentiment_analysis_hard_but_worth_it/ (Accessed: 15/7/2017).
- 35. Kajal S., Vandana P. (6, January 2017)**, “*Opinion Mining: Aspect Level Sentiment Analysis using SentiWordNet and Amazon Web Services*”, International Journal of Computer Applications (0975 – 8887) Volume 158 – No 6, pp 31-33.
- 36. Giovanni A., Georgina C. (2014)**, “*A Hybrid Computational Intelligence Approach for Efficiently Evaluating Customer Sentiments in E-Commerce Reviews*” Publisher: IEEE, Conference Location: Orlando, FL, USA ,pp-2.
- 37. Bing L. (April 22, 2012)**, “*Sentiment Analysis and Opinion Mining*” Synthesis lectures on human language technologies, v. 5, n. 1, pp. 16-29-167.
- 38. Russell J. (1980)**, “*A circumplex model of affect*” Journal of Personality and Social Psychology.
- 39. Tim L., Frank Y., Gilad B., Alex B. (2015-02-13)**, Java SE 8 Edition “The Java® Virtual Machine Specification”.
- 40. Oracle VM Virtual Box**
<http://www.oracle.com/technetwork/servestorage/virtualbox/overview/index.html> [Accessed: at 15/7/2017].

41. **Ahmed A., Walid M., and Stephan V. (August 1, 2013)**, “*A Tool for Monitoring and Analyzing HealthCare Tweets*” Qatar Computing Research Institute
42. Twitter developer documentation, Streaming API request parameters <https://dev.twitter.com/streaming/overview/request-parameters> [Accessed: at 15/7/2017].
43. Flic To lecation <https://www.flickr.com/places/info/1> [Accessed: at 15/7/2017].
44. **Mike T., Kevan B., Georgios P., Di C., Arvid K. (2010)**, “*Sentiment Strength Detection in Short Informal Text*”. In: Journal of the American Society for Information Science and Technology”.
45. **Apoorv A., Boyi X., Ilia V., Owen R., Rebecca P.** “*Sentiment Analysis of Twitter Data*” Department of Computer Science Columbia University New York, NY 10027 USA.
46. **Alec G., Richa B., Lei H. (2009)**, “*Twitter Sentiment Classification using Distant Supervision*” Technical report, Stanford Digital Library Technologies Project
47. **Linhao Z., (April 16, 2013)**, “*Sentiment Analysis on Twitter with Stock Price and Signi_cant Keyword Correlation*”, Department of Computer Science, the University of Texas at Austin, PP 8-11.
48. **Muhammad T. I., Naveed A. B., Muhammad T. A. (20.02.2015)**, “*Open source software adoption evaluation through feature level sentiment analysis using Twitter data*”, Turkish Journal of Electrical Engineering & Computer Sciences, pp4482-4485.

- 49. Lovins J. B., (1968),** “*Development of a Stemming Algorithm*”, Electrical System Laboratory, Massachusetts Institute of Technology, Massachusetts 02139, *Mechanical Translation and Computational Linguistics*, vol.11, nos. 1 and 2, pp. 22-31.
- 50. Sanjiv D., Mike C. (2001),** “*Extracting market sentiment from stock message boards*” In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).
- 51. Neelam D., Amrita K., Sangeeta. (December 2015),** “*LEXICON BASED OPINION MINING SYSTEM USING HADOOP*” *Journal of Network Communications and Emerging Technologies (JNCET)*, Volume 5, Special Issue 2, pp 95 - 99.
- 52. Daniel J., James H. M., (2008),** “*Speech & Language Processing*” An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, pp 83-123.
- 53. William B. C., John M. T. (2014),** “*N-Gram-Based Text Categorization*”, Environmental Research Institute of Michigan.
- 54. Alexander P., Patrick P. (2010),** “*Twitter as a Corpus for Sentiment Analysis and Opinion Mining*”, International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA).
- 55. English grammars there are 8 parts of speech**
<https://www.easypacelearning.com/all-lessons/english-level-2/1347-english-8-parts-of-speech-with-examples>. [Accessed: at 15/7/2017].

- 56. Bholane S. D., Deipali G. (Jun 2016),** “*Sentiment Analysis on Twitter Data Using Support Vector Machine*”, International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 3, pp 336-368.
- 57. Ajinkya I., Anjali K., Shriya S., Anita K. (December, 2015),** “*Sentiment Analysis of Twitter Data Using Hadoop*”, International Journal of Engineering Research and General Science Volume 3, pp 145
- 58. Ishana R., Sourabh G., Parth S., Aishwarya D., Balaji B. (November 2014)** “*Twitter Sentiment Analysis using Apache Storm*”, International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-3 Issue-5, PP 24.
- 59. Mariam A. O., Mohamed M. G., Frederic S.** “*A Survey of Data Mining Techniques for Social Network Analysis*”, School of Computing Science and Digital Media, Robert Gordon.
- 60. Walaa M., Ahmed H., Hoda K. (8 April 2014)** “*Sentiment analysis algorithms and applications: A survey*”, Ain Shams Engineering Journal, pp 1099-1103.
- 61. Lukasz A., Piotr S. N., Tomasz K., Włodzimierz T. (15 December 2015),** “*Comprehensive Study on Lexicon-based Ensemble Classification Sentiment Analysis*”, Academic Editors: J. A. Tenreiro Machado and Kevin H. Knuth, pp3-5.
- 62. Lei Z., Mohamed G., Riddhiman G., Mohamed D., Meichun Hsu, Bing L. (2011),** “*Combining lexicon-based and learning-based methods for twitter sentiment analysis*”, Technical report, HP Laboratories.
- 63. Minqing H., Bing L. (2014),** “*Mining and Summarizing Customer Reviews*” Department of Computer Science University of Illinois at Chicago.

- 64. PANG, B., LEE L. (2008),** *“Opinion mining and sentiment analysis. Foundations and trends in information retrieval”* Foundations and Trends in Information Retrieval, PP27.
- 65. Kumar R., Vadlamani R, (2015),** *“A survey on opinion mining and sentiment analysis”* Institute for Development & Research in Banking Technology, tasks, approaches and applications. Knowledge-Based Systems, Elsevier, PP33-35.
- 66. Felipe B. M., Eibe F., Bernhard P. (2015),** *“Positive, Negative, or Neutral: Learning an Expanded Opinion Lexicon from Emoticon-Annotated Tweets”*, Department of Computer Science, University of Waikato.
- 67. C.J. Hutto, Eric G. (2014),** *“VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text”*, Association for the Advancement of Artificial Intelligence.
- 68. S. Deelters, S. Auwatanamongkol, (2007),** *“Enhancing K- Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance,”* International Journal of Computer Science, Vol. 2, Number 4, pp 518-519.
- 69. SOLUTION GUIDE (2014),** *“Apache Flume™ and Apache Sqoop™ Data Ingestion to Apache™ Hadoop® Clusters on VMware vSphere”*.

APPENDICES A

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: ALKSSO, Mohammed

Date and Place of Birth: 27 November 1983, MOSUL, IRAQ

Marital Status: Married

Phone: +90 535 260 22 57, 00964 770 3333 949

Email: alsabagh.mohammed@yahoo.com.



EDUCATION

Degree	Institution	Year of Graduation
High Diploma (One year)	Mosul Univ., Business Information Technology.	2011
B.Sc.	Mosul Univ., Management Information System.	2006
High School	Alghrbea for Boys	2002

WORK EXPERIENCE

Year	Place	Enrollment
2015- Present	Çankaya Univ. The Graduate School of Natural and Applied Science	Specialist
2012-2014	Technical Collage of Management / Mosul Media Department	Systems Analyst
2011	Mosul Univ. Business Information Technology	Specialist
2007-2010	Technical Collage of Management / Mosul Human Resources Department	Systems Analyst

FOREIN LANGUAGES

Advanced English, Beginner Turkish

SKILLS

- **Programming languages:** JAVA, C++, Object Oriented Programming.
- **Software:** Hadoop with all Ecosystem, Eclipse, visual studio, MS Office Applications, Arc Map.
- **Other skills:** Project Development and Management, Management Information Systems, Introduction to Geographic Information Systems, Networks, Fractional Differential Equations, Analysis (Mathematics).

HOBBIES

Travel, Books, Swimming, Fitness