

REPUBLIC OF TURKEY
FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE

**Sentiment Analysis and Opinion Mining from Big Social
Data Using MapReduce and Machine Learning Methods**

Banan Jamil Awrahman

Master Thesis
Department: Software Engineering
Supervisor: Assoc. Prof. Dr. Bilal Alataş

MARCH – 2017

REPUBLIC OF TURKEY
FIRAT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE

**Sentiment Analysis and Opinion Mining from Big Social Data Using Map
Reduce and Machine Learning Methods**

MASTER THESIS

BANAN JAMIL AWRAHMAN

142137108

Submission Date to the Institute: .../.../2017

Thesis Presentation Date: 31.03.2017

Thesis Supervisor: Assoc. Prof. Dr. Bilal ALATAS (F.Ü.)

Other Juries: Prof. Dr. Asaf VAROL (F.Ü.)

: Asst. Prof. Dr. Özal YILDIRIM (M.Ü.)

MARCH- 2017

DECLARATION

I am presenting the thesis entitled “Sentiment Analysis and Opinion mining from Big Social Data by using MapReduce and Machine Learning Methods” for fulfilling the requirements of Master’s Degree in Software Engineering. I declare that the presented content of this thesis is my own work with all simulations and programming.

Banan Awrahman

ELAZIĞ – 2017

ACKNOWLEDGMENT

After an intensive period, today is the day: writing this note of thanks is the finishing touch on my thesis. It has been a period of intense learning for me, not only in the scientific arena, but also on a personal level. Writing this thesis has had a big impact on me. I would like to reflect on the people who have supported and helped me so much throughout this period.

I would first like to thank my thesis advisor Assoc. Prof. Dr. Bilal Alataş of the software engineering at Firat University. The door to Prof. Alataş office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right the direction whenever he thought I needed it.

I would also like to thank the expert who were involved in the validation survey for this research project, without their passionate participation and input, the validation survey cloud not have been successfully conducted, who have willingly shared their precious time during the process of interviewing.

Finally, I must express my very profound gratitude to my parent for giving birth to me at the first place and supporting me spiritually my life, and all friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and witting this thesis. This accomplishment would not have been possible without them.

Thank you.

TABLE OF CONTENTS

	<u>Page NO</u>
DECLARATION	I
ACKNOWLEDGMENT	II
ÖZET	V
ABSTRACT.....	VI
LIST OF FIGURES	VII
LIST OF TABLES	IX
ABBREVIATIONS.....	X
1. INTRODUCTION	1
1.1. Aim of Study.....	3
1.2. Overview of Thesis	3
2. RELATED WORKS.....	4
2.1. Sentiment Analysis	4
2.2. Social Big Data Processing Using Open Systems	4
2.3. Analysis of Sentiment Using Hadoop.....	5
2.4. Analysis of Sentiments Using MapReduce.....	6
3. THEORETICAL FRAMEWORKS.....	7
3.1. Big Data	7
3.2. Social Media	8
3.3. SA for Big Data Problems	9
3.4. Analysis of Sentiments Using Big Data Framework	10
3.5. Sentiment Analysis General Approach.....	13
3.6. Features of Sentiment Analysis	13
3.6.1. Explicit Characteristics	13
3.6.2. Implicit Characteristics	14
3.7. Konstanz Information Miner (KNIME).....	16
3.8. Classification of Sentiments via Machine Learning	16
4. EXPERIMENTAL RESULTS AND DISCUSSIONS	20
4.1. System Design	20
4.2. Dataset.....	21
4.3. Data Preparation.....	21
4.4. Preprocessing	24
4.5. Saving Files.....	26
4.5.1. MapReduce	26
4.6. Prior Polarity Scoring	29
4.7. Text Classification	32
4.8. Experimentation and Results	33

5. CONCLUSIONS AND RECOMMENDATION.....	48
6. REFERENCES.....	49

ÖZET

MapReduce ve Makine Öğrenmesi Yöntemleri ile Büyük Sosyal Veride Duygu Analizi ve Fikir Madenciliği

Duygu analizi; doğal dil işleme, metin analizi, hesaplama dilbilimi ile metin belgelerindeki öznel bilgilerin belirlenmesi ve çıkarılması yöntemidir. Duygu analizinin ana görevi belirli bir metnin kutupsallığını belirlemektir. Bu araştırmanın amacı, MapReduce ve KNIME aracı ile makine öğrenimi kullanarak büyük sosyal veriler için duygu analizini incelemektir. Bu araştırmada kullanılan veri setleri internetten gerçek zamanlı olarak alınmıştır. Araştırmada kullanılan model, güncel bir şemayı takiben, veri alma tarihini ve zamanını kullanmaktadır. Metin alanlarını, Twitter'da en sık yazılan cümlelerden girdi vektörü olarak toplamaktadır. Bu veri kümesini, çapraz doğrulama yöntemi kullanarak aynı türde bir test seti takip etmektedir. Rasgele bir yüzdelik eğitim seti, hata oranına ve performansa karşı test edilmiştir ve değerlendirilmiştir. Twitter API'sı, Twitter'dan veri toplamak için kullanılan bir diğer araçtır. Toplanan veriler Hadoop'tan KNIME aracına girilmiştir. Ayrıca, veri sınıflandırması için, bu projede makine öğrenme algoritmaları kullanılmıştır.

Anahtar Kelimeler: KNIME aracı, duyarlılık analizi, büyük veri, MapReduce, Hadoop, Twitter

ABSTRACT

Sentiment analysis is the use of natural language processing, text analysis, computational linguistics identifying and subjective information extracting from text documents. The main task of sentiment analysis is to determine the polarity of a given text. The aim of this research is to investigate sentiment analysis for big social data using MapReduce and machine learning with KNIME tool. The datasets used in this research are fetched from the Internet in a real-time basis. The model used in the research is employing the current date and time for receiving data and following an up-to-date schema. It gathers text fields as input vectors from the most frequently written phrases on Twitter. The collected datasets are passed to an evaluating stage that employs both confusion matrix and cross validation methods. Twitter API is another tool that was used to gather data from Twitter in this work. The collected data was fed to KNIME tool from Hadoop platform. Also, for data classification purpose, machine learning algorithms were used.

Keywords: Hadoop, big data, KNIME tool, MapReduce, sentiment analysis, Twitter

LIST OF FIGURES

	<u>Page No</u>
Figure 4.1 System design.....	20
Figure 4.2 Twitter API.....	21
Figure 4.3 Create new application in Twitter	22
Figure 4.4 Twitter application management console.....	22
Figure 4.5 Twitter API Connector (API)	23
Figure 4.6 Searching data from Twitter.....	23
Figure 4.7 Preprocessing step in KNIME tool.....	24
Figure 4.8 Tweets before preprocessing	25
Figure 4.9 Results after preprocessing.....	25
Figure 4.10 MapReduce workflow	27
Figure 4.11 Get data from Hadoop to KNIME.....	29
Figure 4.12 Prior polarity scoring.....	30
Figure 4.13 Labeling tweets.....	31
Figure 4.14 SVM classification	32
Figure 4.15 K-NN classification.....	33
Figure 4.16 SVM classifier based on 5k tweets by using confusion matrix method.....	36
Figure 4.17 SVM classifier based on 1k tweets by using confusion matrix method.....	38
Figure 4.18 K-NN classifier based on 5k tweets by using confusion matrix method	39
Figure 4.19 K-NN classifier based on 1k tweets by using confusion matrix method	41
Figure 4.20 Classification accuracy, precision, recall and specificity for 1k,5k tweets data set	42
Figure 4.21 SVM classification results for 1k tweets by using cross validation method	43
Figure 4.22 K-NN classification results for 1k tweets by using cross validation method.....	44

Figure 4.23 SVM classification results for 5k tweets by using cross validation method	45
Figure 4.24 K-NN classification results for 5k tweets by using cross validation method	46

LIST OF TABLES

	<u>Page No</u>
Table 4.1. Confusion matrix	34
Table 4.2 confusion matrix for three classes	34
Table 4.3. Confusion matrix for SVM classifier based on 5k tweets	35
Table 4.4. Confusion matrix for SVM classifier based on 1k tweets	37
Table 4.5. Confusion matrix for k-NN classifier based on 5k tweets.....	38
Table 4.6. Confusion matrix for k-NN classifier based on 1k tweets.....	40
Table 4.7. Classification accuracy for 1k, 5k tweets dataset.....	41
Table 4.8. Classification accuracy for 1k tweets data set	44
Table 4.9. Classification accuracy for 5k tweets data set	46

ABBREVIATIONS

AI	: Artificial Intelligence
API	: Application Programming Interface
AWS	: Amazon Web Services
BI	: Business Intelligence
CPU	: Central Process Unit
EMR	: Electronic Medical Records
HDFS	: Hadoop Distributed File System
IBM	: International Business Machines
k-NN	: k-Nearest Neighbors
KNIME:	: Konstanz Information Miner
ME	: Maximum Entropy
ML	: Machine Learning
NB	: Naïve Bayes
NLP	: Natural Language Processing
NLTK	: Natural Language Toolkit
NoSQL	: Non-unique Structured Query Language
OAA	: Oracle Advanced Analytics
PMML	: Predictive Model Markup Language
PNL	: Power Networking Lunch
RDBMS	: Relational Database Management System
SA	: Sentiment Analysis
SNS	: Social Networking Sites
SO	: Semantic Orientation
SSH	: Secure Shell

SVM : Support Vector Machine
UGC : User-Generated Content
VIP : Variable Information Processing

1. INTRODUCTION

Sentimental Analysis (SA) of data allows analyzing the feedback of a product. People publish information about a certain product on the social media such as Twitter. Thus, the online analysis of data is a useful practice. The social media platform used in this project is Twitter because most of the people get the feeling that when there are tweets, the value of the product increases. Deploying the Relational Database Management System (RDBMS) using the traditional methods would require powerful hardware resources [1].

The main contributors to the rapid growth of big data are the reduction of computing power and storage cost. In the past, for a company to make decisions based on the transaction data stored in a relational database, rather than less structured traditional data, some potentially important resources were ignored. The use of large data lines involves the development of the current architecture of enterprise data to integrate big data facilities and offer a better business value to the user. Big data allows companies to make specific decisions instantly to increase their profitability and shares in the market. However, establishing a confidential relationship, and defining new opportunities to organize and protect the means necessary to capture and protect the value of large data is a big challenge to overcome. In large quantities, the distillation and analysis of data can contribute to a better understanding of the company's increased productivity, improved competitive position and innovation. Potentially, the research has increasingly focused on analyzing new and different digital data streams, as well as analyzing new economic values of technology to show new opportunities in customer behavior to create and define proactive market trends [2, 3, 4].

SA is one of the major big data agendas which is focusing on large data analysis to identify trends and relationships, making smart predictions, providing information to an action and getting a business overview of the constant inflow of various information methods [5]. SA is typically used to analyze emotions of people with respect to things such as products, services, personal, themes, issues, events, and other entities and their emotional attributes, such as opinions, evaluations, attitudes, and emotions through texts, videos, and other means of online communication presented. The emotions here can be divided into three main categories, namely

positive, neutral, and negative. These categories include many somewhat different names and tasks, such as advice mining, opinion mining, emotions mining, subjective analysis, customer complaints, feel analysis, sentiment analysis, comments, mines and critical analysis. Many techniques have been introduced for SA in recent years. These techniques can be divided into the following categories: application-oriented, from joint price forecasting to speech analysis, population monitoring and client care sentimental analysis based, basic concepts, including word level, emotional level, sentence-level of SA, aspect-level of SA, the concept-level of SA class and analysis of linguistic characteristics, and social intelligence. These techniques are used for analyzing the production of online content such as the spread of the pandemic, emotions, and reactions to local events and other inputs [5].

In conjunction with Social Networking Sites (SNS) use, opinion mining has been recently initiated to fetch useful SNS texts and multimedia data in order to evaluate the true intentions of the user. To be more specific, mining has been used to understand the views and opinions of users on the meaning and real intent of the social network. The need to swiftly extract meaningful information from large amounts of data generated by SNS technology is on the rise, and a new method is needed to extract thoughts and ideas in real time from these large amounts of data. The extracted information can be useful for many areas, such as economics, politics, media, and culture [6].

Social media is a tool for social interaction between people and their communities. Social media creates a virtual network, share or/and exchange information, ideas, images, videos and texts. Social media contains various forms of multimedia information [6]. Specifically, Twitter contains various forms of information, such as video links, images links, and text data [6]. In this study, the focus is guided towards text data where the method of extracting information from the emotion of large text data is investigated. The data extracted are unstructured because the format is easy and free. The shape and structure of the unstructured data are so complex that they cannot be defined as standardized data unlike videos, images, and documents [6]. To extract useful information from large volumes of unstructured data available on the SNS, the unstructured data need to be structured. Until now, different techniques have been studied for the treatment of unstructured data, with an emphasis on morphological analysis. Therefore, there

has been a rise in the development of text extraction research [7-8-9]. The new development helps the extraction of information from semi-structured or atypical text-based data via Natural Language Processing (NLP) technology. The methods used here are based on Machine Learning (ML) algorithms to extract statistical meaningful information as well as extracting information from large amounts of text data. In addition, on the basis of text mining, a number of studies have also been conducted on opinion mining to determine emotional trends and preferences of social networking users such as positive, negative, and neutral [10, 11]. Recently, SA has been provided with a large data processing associated with a variety of free sources. Hadoop ecosystem [12] is the most commonly known data processing system. Hadoop will be discussed further in the following chapters. In this study, based on Hadoop Distributed File System (HDFS) [12] and MapReduce function [13], which can collect and store a variety of data, users have been intended to be analyzed using unstructured data generated by social networks.

1.1. Aim of Study

The purpose of this research is to investigate SA for big social data using MapReduce and ML with KNIME tool.

1.2. Overview of Thesis

Chapter 1 discusses the general review of the topic, the aim, and thesis organizations.

Chapter 2 presents the related literature on sentiment analysis.

Chapter 3 gives the theoretical framework of the thesis.

Chapter 4 presents the experimental results and discussion of the thesis.

Chapter 5 presents the conclusions and final recommendations of the thesis.

2. RELATED WORKS

2.1. Sentiment Analysis

The analysis of sentiments is a method used for discovering and extracting information of subjective nature from the main data, text analysis, computational linguistics and NLP [11]. So far, a large number of studies has been proposed and developed by researchers for big data analysis of the user's mood [14]. There are two ways to determine the sentence polarity: The first is using dictionaries and method dictionaries, the second method dictionaries are that the use and correctness of the power is very good, but there are still some problems at the expense of objectivity and classification [15].

Pak and Paroubek [16] proposed a method that can automatically collect corpuses for the purpose of sentiment analysis and opinion mining. By using corpuses, they have established an emotional classifier that can determine positive, negative and neutral emotions. Although they obtained a valid classification in their analysis of feelings, they did not consider the collection of material and the preservation of the original data.

Go *et al.* [17] disclosed a method that relies on remote testing for sensitivity analysis. They used a tweet label and a multi-class classifier to generate training data to determine the sentence polarity.

Barbosa and Feng [18] suggested a method for analyzing the feelings of Twitter. This method marks POS gram hashtags and user-friendly functions.

2.2. Social Big Data Processing Using Open Systems

The Hadoop ecosystem is providing a variety of large data processing options for open source projects [12]. The database administration framework, which is broadly utilized for substantial scale information handled by Hadoop frameworks, is a Non-Unique Structured Query Language (NoSQL) [19]. From a vast stored data, little useful information can be recovered utilizing a conformance model, which is less prohibitive than traditional relational databases. It is an exceptionally flexible database since it gives horizontal and vertical versatility. However, it does not utilize join operations amongst table and table patterns, and it gives a speedy response to

peruse and compose [20]. At present, some NoSQL-based database models are being presented in the scholarly world and industry, such as Amazon DynamoDB from <http://amazon.com/> [21], HBase from the Apache Software Foundation [22], MongoDB [23], Cassandra [12] and BigTable from Google [12], which are some of the commonly used ones.

MonDB, specifically, does not have a pattern and gives a consistent expression to discover and look whether the sentence adaptably incorporates a particular esteem. Since NoSQL can be expanded utilizing the horizontal extension technique (horizontal adaptability). Huge information preparing frameworks can be stretched out by including existing frameworks requiring little to no effort without overhauling the framework. The framework has a costly Central Process Unit (CPU). As a result, large amounts of data can be processed in parallel using MapReduce technology, filtering, data aggregation operations, statistics, and data retrieval compared to conventional relational database management systems.

2.3. Analysis of Sentiment Using Hadoop

So far, there have been many studies on Hadoop's feelings analysis. Khuc *et al.* [24] described big distribution systems, for instant Hadoop analysis of feeling. Here, the system contains two elements: dictionary builder and emotion classifiers. They can use the MapReduce framework and distribution database systems running on large-scale distributed systems. Bautin *et al.* [25] introduced TextMap access system Lydia, which allows the access to the system using Hadoop batches of interesting statistical data on a large number of people, places and things. Lydia consisted of five main elements: spidering, brand PNL, SA, physical analysis, aggregation and visualization.

Apache Hadoop is the application of the MapReduce. Hadoop has been applied successfully to datasets based on files. Apache Hadoop distributed computer project to develop a reliable system for open source software. Hadoop software library is a framework that allows the use of thousands of separate computers and calculates PB data, using distributed cluster processing of computers across large datasets. Existing tools are not designed to handle such large amounts of data. Hadoop averts the setbacks by providing efficient storage and computing power to avoid the disadvantages of processing a large amount of data.

2.4. Analysis of Sentiments Using MapReduce

MapReduce is a programming model that implements an associated algorithm for the use of distributed processing and generating large datasets on clusters. MapReduce programs run by filtering and sorting the Map () process and implementing the synthesis operation Reduce () process components. MapReduce is considered a new framework for the special processing of large sets of data distributed over the used source. The MapReduce is programming model for the generation and processing of a large dataset [26].

Map: The master node that accepts the input and splits it into smaller sub-problems, then distributes them to the work node. A working node can repeat doing so, resulting in a multi-level tree structure. The work node processes a small problem and sends the results to its main node [27].

Reduce: The main node that collects all the results of the sub-problems, and somehow combines them to form the final output - for its first attempt to solve the problem which was found [27].

3. THEORETICAL FRAMEWORKS

In recent years, not only there has been a big growth of data, but also for practical reasons, its in-depth analysis. The computer system of gigabytes or even terabytes of measurement of data to operate on, considering the user and computer system with fast speed with a continuous data generation. Researchers and computer engineers have coined the term “big data” to name this pattern. The main qualities of the substantial volume of information are assortment and speed. For the benefit of a vast volume, it cannot be effortlessly taken care of by traditional database frameworks and individual machines. Speed implies that information need to make a constant speed, which corresponds to an assortment of various shapes, for example, content, pictures and recordings. The ascent of enormous information has a few reasons.

One of them is the increase in the number of cell phones, especially smartphones, tablets and portable Personal Computer (PCs) which are usually connected to the internet. Another reason is that the PC framework has started to be utilized as a part of numerous segments of the economy, from government and nearby specialists, and from pharmaceutical services to the finance division. As a side effect of the organization's information investigation of various exercises with a specific end goal to better comprehension of clients' requirements and the anticipation of future patterns. Past research attempts showed that precise investigation can be utilized to foresee drifts in money related markets [28].

In addition to the expansive size, the fundamental concern is to dissect and translate extensive information and applications that the partner sees. Information examination, otherwise called information mining, can be utilized, for example, ML, computerized reasoning and insights of various systems. Additionally, it is imperative to consider the span of the information to be handled, which thus decides the calculation or a current technique given if important.

3.1. Big Data

“Big Data” has several definitions, one of which is stating that “Big data refers to more than the size of a typical database software tools for capturing, storing, managing and analyzing datasets” [29]. This definition focuses on the key aspects of a large volume of data, speed and diversity [30], which is according to the IBM report [31], creating a daily “2.5 million bytes of

data”. These figures are increasing every year. This is due to the universal access to the Internet as well as the increasing number of equipment and data transmission devices. Specifically, the ones involved in processing data from a variety of real-time operating systems, social media for instance, which continue to accumulate information about client activity and interaction.

This aspect needs to store data in a variety of ways to maximize the speed. Sometimes using the database system or no column-oriented schema (NoSQL), one can do the job because the data is large. However, a well-structured large data is not only difficult but also creates major opportunities such as: creating transparency, optimizing and improving the performance, generating additional profits, new discovery ideas, services and products.

3.2. Social Media

One of the greatest trend of data is the rise of Web 2.0. It is the transfer from a static website to having a major change as User-Generated Content (UGC) interactive website. Web 2.0 has prompted to the notoriety of many administrations; for example, online journals, podcasts, interpersonal organizations, and bookmarks. Clients can make and share information in an open or closed group, contributing to a lot of data transmission. Web 2.0 has prompted to the creation of web-based social networking, which is presently in the group to make. This permits the contribution and trade of information with others through electronic media.

Online networking can likewise be abridged as “in view of three main components: content, group and Web 2.0” [32-33]. Every component is a key consideration in the online networking. One of the most imperative considerations is advancing online networking from cell phones connected to the Internet, for example, PDAs and tablets. Twitter is a microblogging stage which allows the enlisted users to post and read messages called “tweets”. Each tweet is limited to 140 Unicode characters long – as the restrictions on SMS transport. Users not exclusively can see the tweets. Users cannot set up relationships or relationships of concern.

Subscribe to different clients who are called "devotees", and get refreshes continuously from that person. Be that as it may, users ought not to include individuals as their supporters. Twitter can be gotten from an assortment of administrations, for example, the official Twitter site, portable applications, and SMS administrations from outsiders. Since Twitter is an extensive

variety of administrations, particularly in the United States, in light of the fact that the data structure is minimized, so it constrains users to make a short remark. In addition, Twitter is tool used by politicians and other VIP members of public relations for communication purpose. This has cultural and social impact on large communities. Thus, Twitter was selected as experimental data sources.

3.3. SA for Big Data Problems

In spite of the fact that SA is one of the primary plan of substantial data, there is not many working strategies about the SA framework for extensive data. This section concentrates on this view.

The big data is related to some issues, for example, speed, volume, esteem, instability, and veracity of the data. The amount of data is the primary attribute of expansive data, truth be told, this is the principle reason why the term “big data” was coined. A cozy relationship with the volume variable is the speed, which creates and handles continuous streams of data on the pertinent sensor, and in this way should be broken down. At the point when there are tons of data produced continuously, there should be a way for determining the genuineness instability, legitimacy, confusion, and unwavering quality. Given these elements and issues identified with the species, and in this manner consider the quality and exactness of the data, in light of the fact that the data is created in various arrangements and styles. Next, comes the issue of data qualities, which ought to be utilized as a part of an opportune way.

The decision of where the data is considered legitimate and, in this way, must be put away in unpredictable or duration irrelevant. The actualities above demonstrate that not just vast data brings new sorts of data and capacity components; additionally, it brings another kind of analysis. Big data analysis is a continuum, not a free aggregation action, including various diverse data handling. Here, the absence of this data in its unique frame data model to characterize the importance of every component in different contexts is a big challenge.

Toward the start of this new analysis, a few new questions ought to be considered. These questions incorporate revelation, iterative, adaptable mining and prescient limit and decision administration [34]. The issue is founded because of the reality that the data esteem is typically covered up underneath the surface of the arrangement of data gathered, and cannot be dictated

by the disclosure procedure. In addition, the genuine relationships inside the enormous measure of data are not generally known ahead of time. In this way, uncovering a diagram is frequently an iterative procedure, until the appropriate response is discovered. An inescapable issue related with vast data limit is the adaptability. While distributed computing is utilized for vast data, the iterative way of expansive scale data analysis needs to be utilized additional time and assets to tackle the issue by hand. Recognition, mining, and anticipating how distinctive data components identify each other is a similar issue.

3.4. Analysis of Sentiments Using Big Data Framework

SA concentrates mainly on the identification of the composer's emotions. The strategy for accomplishing this goal can be separated into two main classes, content and no content. SA is firmly identified with the mining of opinion, which is characterized as the object, characteristic of the object, the emotional value of opinions, that expressed [35].

Although mining opinion was introduced earlier, but due to its commercial value, it should be pointed out that SA in the field of large data has got more and more attention [36]. Indeed, social media is increasingly dependent on Feedback and its related activities. Subsequently, organizations must be online to tune in the voice of their clients (and in this manner are the fundamental points of interest of SA). For instance, a showcasing effort to advance the item is a great way. Also, responding to grievances and considering the general concept of people in their advertising and vital arrangements. In such manner, the objective is to comprehend online messages of emotional direction (otherwise called polar), to monitor the sender, and to comprehend the subjects and topics as well as the fame of the message [37].

In spite of the fact that SA ponders have been conducted for over 13 years, yet not to highlight the substantial data. Rather they have been used for the closeness analysis of online networking where a huge data stage gives a few clients SA services [38]. Given the vast number of online networking movements, web-based social networking is the initial phase in the analysis to comprehend the extent of the data to be gathered for analysis. By and large, data can be constrained to certain hash labels and linked to catchphrases.

Hadoop pre-prepares data to distinguish full-scale drifts or to discover which pieces of information are helpful, for example, the estimation of the range. It allows companies to use inexpensive product servers from the new data release to potential high value. Many big organizations use Hadoop as a leading form of predecessor analysis. Hadoop is filtering, sorting, or a large number of new data pre-processing and distillation mechanisms. It is used to produce a popular choice theory contains more useful “information” from denser data. Pre-processing involves filtering a new data source, which is reasonable for further analysis in the data distribution center.

MapReduce permits us to acquire unstructured data, data conversion (mapping) that is significant data, and afterward cured data for the final reporting. Every one of these ways occurs in parallel in the Hadoop group on every one of the hubs. A straightforward case of MapReduce is the web-based social networking messages that can be mapped into a rundown of words and tally their events. Such a rundown is then diminished to one word that seems day by day tally of the number [39].

Once noteworthy data is put away in Hadoop, it can be stacked into existing Business Intelligence (BI) stage or utilize effective self-benefit apparatuses (for example, PowerPivot and PowerView) to be examined specifically. Utilizing SQL Server as a corporate BI stage, clients have a few options to get to their Hadoop data. These options incorporate Sqoop, SQL Server Integration Services, and Polybase.

An Oracle database and through a combination of inward calculations and R open source calculation, available by the means of SQL and R language, as well as the introduction of Oracle Advanced Analytics (OAA) can be deployed to find shrouded relationships in the data. OAA is superior data extraction and the open combination of the source language R. Thus, it is essential for performing a prescient analysis inside the database, data mining, content mining, measurable analysis, intelligent design and progressed numerical registering.

Amazon Web Services (AWS) utilizes cloud AWS training stack, which offers a scripting method for gathering online networking messages, for instance, tweets from the famous Twitter platform. These tweets are stored in Amazon S3. Some framework applications, such as

Amazon EMR, can be then used for analyzing the archived data. This offers the ability to create an EMR bunch. This bunch utilizes Python NLTK SA program to record the data. At that point, it assesses the yield recorded to control tweets aggregation emotions.

Big data analysis tools' main feature is the real-time analysis support to help users stay ahead of the competition. For example, the analysis of data derived from a variety of disjoint dashboard systems. This dashboard goes beyond the data repository with many formats (overview) and has the ability to track trends in the data flow based on constructed decisions (actions).

SA method in the analysis tools is mainly achieved by the application of brand-oriented management, where the phase begins with how to conduct a study on public and consumer content analysis, and companies' trend and absorption of strategic decisions. SA's various instruments can be utilized to track social advertising. These instruments can be divided into reference analysis or content analysis. The reviewed systematic applications, for instance, Tweepchup and Social Sprout, do not give a top to bottom analysis of the message content, watchwords and significant patterns in web-based social networking organizations. However, these applications are generally free. Content analysis involves high expenses, basically because of its intuitive dashboards and multi-language abilities. The instruments used here incorporate Radian6, soften water, Simplify 360, mark watches and Hootsuite, where their elements capabilities incorporate following, analysis and summary of demographics. Free applications, on the other hand, perform content analysis, yet with low precision.

While a considerable number of the mentioned applications use online networking content improvement, yet their designs do not utilize substantial limit of data ingestion apparatuses. Applications, for the most part, are centered around the collection of moderate online messages, emotion classification within messages, subjectivity removal and personalized visualization. Late applications incorporate SA Horton Works, which concentrates on huge data SA, coordinated Flume and Power View to gather and view data. This tool restricts the SA functionality, as it is constructed exclusively with respect to the standard Python NLTK emotion motor. Application of the SA content analysis innovation based permits a more profound and thorough, and bringing about more prominent exactness.

3.5. Sentiment Analysis General Approach

SA, or opinion mining, extricates positive or negative opinions from unstructured contents. At first, Hearst and Wiebe proposed content mining in the view of direction (i.e., contents containing opinions, emotions, impacts and partialities). In content analysis, traditional structures, for example, the nearby analysis may not matter to the discussions. Accordingly, emotional analysis has lately been utilized as a part of many types of system based tasks [40]. The emotional classification has a few imperative attributes, including different undertakings, elements, and systems. Three critical emotional respectful errands are as per the following:

- Identifying whether the content is objective or subjective, or whether the subjective content has a positive or negative direction
- Determining the classification level (record/sentence level)
- Identifying the source/focus of the mindset

Two common questions include classifying the direction as positive or negative [41]. In addition, a few specialists are focused on grouping information into opinion/subjective or truthful/objective. Furthermore, a few scientists attempt to characterize emotions, for example, bliss, bitterness, outrage and dread, instead of emotions. Emotional extremity classification is a delegated archive level, sentence and expression class classification. Record level classification characterizes archives as neutral, negative and positive. The sentence class classification only considers and orders a sentence to figure out if the sentence is subjective or objective. So as to catch different emotions that may exist in a solitary sentence, the expression classification is performed. In addition, in order to classify the levels of horizons and emotions, different assumptions have been made for the source of emotions and goals. The features used for the classification of emotional polarity and the techniques based on ML are described in detail in the next section.

3.6. Features of Sentiment Analysis

3.6.1. Explicit Characteristics

In SA examination, four sorts of express elements are utilized, namely language structure, semantics, connect based, and style-based components. The language structure quality is the

most common arrangement of SA components. In addition, these characteristics contain the expression design, utilizing the Part-Of-Speech (POS) label n-gram designs, and POS labels [42]. They show that an expression example, for instance, “*n+aj*” (as a rule took after by a descriptor) as a rule demonstrates a positive emotional orientation, and “*n+dj*” (a thing took after by negative modifiers) generally express negative emotions.

In 2004, Wiebe connected a set in which some parts of the settled n-gram were traded with the general words mark. Whitelaw *et al.* [43] connected an arrangement of altered components where the nearness of these elements changes the evaluation qualities of the dictionary extend. Connect/reference analysis is connected to interface based functions to distinguish emotions from the Web and records. Efron *et al.* [44] have demonstrated that the perspectives website pages are connected to each other. Interface based elements have been utilized as a part of restricted research, and in this manner the legitimacy of these components is still misty. The complex elements contain auxiliary and lexical properties for some past stylometric/creator works. Vocabulary and auxiliary style markers have been utilized for constrained emotional analysis.

3.6.2. Implicit Characteristics

The investigation of understood components in SA concentrates on semantic and etymological guidelines to distinguish embedded messages. These messages are normally not spoken to by predefined watchwords. Rather, the significance is passed by utilizing a comparable concept-based expression. Semantic elements attempt to distinguish polarities or give force related scores to words and expressions. Hatzivassiloglou and McKeown [45] developed a Semantic Orientation (SO) technique which is later delivered by Asur and Huberman [46]. SO technique was broadened utilizing idle semantic analysis [47].

Manual or semi-naturally created emotional dictionary [48] often utilizes an arrangement of significant auto-produced terms. These terms are physically separated and encoded with extremity and force information. Client characterized labels are utilized to demonstrate regardless of whether certain expressions have positive or negative emotions. Self-loader dictionary generate instruments to construct an arrangement of strong subjectivity, frail

subjectivity, and target things. They additionally utilize different components, for example, pack of-words to order English archives as a subjective or goal. Another strategy for explaining semantics as a word/expression is the evaluation group. The initial term list is created using WordNet, and then manually it filters these lists to build the dictionary.

WordNet

WordNet was created at Princeton University in 1986. It is an English vast electronic vocabulary database, which continues to create and keep up. WordNet incorporates synonyms from the primary sentence structure classes, for example, things, verbs, descriptors and intensifiers. The current version of WordNet (3.1) contains more than 117,000 synsets that include 81,000 things synsets, 3600 verb synsets, 19,000 descriptors synsets and 3600 intensifiers synsets [49]. Most present reviews utilize WordNet and SentiWordNet. WordNet has been utilized for synonym collection whereas SentiWordNet has been utilized to distinguish the semantic orientation of each sentence or separated element.

SentiWordNet

SentiWordNet is a vocabulary asset for popular opinion mining. It is a dictionary foundation, like WordNet, yet it reaches out with vocabulary information about every synset contained in WordNet's emotions. Three distinct polarities, to be specific excitement, antagonism and objectivity, are relegated to every synonym in WordNet. The two most common versions of SentiWordNet utilized as a part of many reviews are SentiWordNet 1.0 and SentiWordNet 3.0. In addition to being utilized for monolingual learning, SentiWordNet can be utilized as a part of multilingual SA [50].

SenticNet

SenticNet works by utilizing sentic calculations. It is the most recent semantic asset created for concept level SA. It utilizes Artificial Intelligence (AI) and semantic web innovation to better recognize, translate and control characteristic language seen on the web. SenticNet is a learning base that can be utilized as a part of numerous regions of improvement, for example, expansive social data analysis, e-health, human-PC interaction, etc. [51].

3.7. Konstanz Information Miner (KNIME)

Content analysis and emotional analysis [55] are incorporated into the perspective, the calculation of emotions and the content of the subject. This is done by identifying the analysis of characteristic language content data. The emotional analysis is the utilization of common language handling, content analysis and computational phonetics to recognize and separate the subjective information of the content record. The essential undertaking of emotional analysis is to decide the extremity of a given content. The dictionary action and the emotional analysis process are performed through KNIME [56], which is an easy way to understand how graphical functions of the whole analysis are handled. KNIME utilizes six stages to process a content: perusing and reporting analysis, distinguishing named substances, separating and controlling, word checking and catchphrase extraction, handling and visualization. The work process, improvement and utilization of the accompanying errands KNIME Implementation are as follows:

- Data retrieval from the database
- Develop and execute dictionaries
- Scores reviewer

3.8. Classification of Sentiments via Machine Learning

The ML approach deploys the ML calculation and utilizes language elements to improve the execution of the framework using test data. Huge data structures, for example, Mahout and Pentaho contain libraries and modules for ML strategies that can be executed to perform emotional classification. In the context of vast data analysis, the client ought to decide the kind of calculation to be connected to the hand data, and this calculation is executed by an expansive data analysis apparatus for a particular critical thinking reason (e.g., prescient analysis).

As a rule, two arrangements of archives are required for ML-based classification. These are the preparation and test sets. The classifier utilizes the preparation set to take in the report properties, and uses the test set to confirm the classifier execution. Content classification utilizing the ML strategy can be divided into supervised learning techniques and unsupervised learning strategies. The supervisory technique utilizes countless of learning archives. At the

point when these labeled preparing records are elusive, the maneuver should be switched to the unsupervised strategies. The supervisory approach accomplishes reasonable legitimacy. However, typical area particular and language-related require labeled data, which is generally work escalated. In the meantime, unsupervised techniques have their popularity in light of the fact that the openly accessible data is typically unmarked, and along these lines requires a strong solution. In this manner, semi-regulated learning has been presented and has brought on a considerable attention in the emotional classification. Unsupervised learning utilizes countless data and labeled data to produce a superior learning model.

Various ML methods have been utilized to perform classification errands in SA. The most prevalent ML methods that have made awesome progress in content categorization are Support Vector Machines (SVM), Naive Bayesian (NB), and Maximum Entropy (ME). Other well-understood ML techniques in characteristic language handling are k-Nearest Neighbor (k-NN), ID3, C5, Centroid Classifier, Winnow Classifier, and N-Element Syntax Model.

Support Vector Machine

SVM is a measurable classification strategy for auxiliary hazard administration standard utilizing computational learning hypothesis. Contrasted with other ML innovations, for example, NB and ME, SVM has turned out to be a productive approach to ordering traditional content [52].

Naive Bayesian

The NB classifier is a straightforward likelihood classifier in light of the Bayesian hypothesis. NB is especially appropriate when the information has a high dimension. NB is a basic and compelling calculation that has been broadly utilized as a part of writing classification [53]. At the point when the number of elements is little, NB is better than SVM. The calculation can likewise be enhanced when joined with different strategies, for example, traditional dictionaries. A straightforward NB classifier can be improved in order to see more intricate models. This is achieved for discovering a more fitting component selection and undesirable element (commotion) evacuation.

Maximum Entropy

ME is another ML classifier that has ended up being powerful in numerous regular language applications. Alternatively, to NB, ME does not make assumptions about the relationship between elements. This improves its conceivable performance when the conditional autonomy assumption is not fulfilled. Now and again, the ME entropy classification model is better than the dictionary-based approach in distinguishing sentences in emotional sentences, for instance, where the words in the dictionary cannot express emotional propensities [54].

k-Nearest Neighbor Machine Learning Algorithm

In pattern recognition, the nearest k -neighbor calculation is a nonparametric technique for regression and classification [57]. In both cases, the information consists of the k nearest learning tests in the function space. The yield relies upon whether k -NN is utilized for classification or regression:

In the k -NN classification, the yield is a participation to the class. The question is sorted by most of the votes of its neighbors, where articles are doled out to the most common class in their closest k neighbors (k is a positive whole number, generally small). In the event that $k = 1$, the question is basically appointed to the closest neighbor class.

In the k -NN regression, the yield is the characteristic estimation of the question. This esteem is the normal of the estimations and its k closest neighbors.

k -NN is a case-based learning or inactive realizing, where the function is only privately approximated, and all calculations are continued to the classification. k -NN calculation is one of the least difficult calculations of all programmed learning calculations.

Independent of classification and regression, it might be helpful to assign the weights of neighboring contributions, so that later neighbors contribute more than the normal. For instance, the common weighting plan is to give each neighbor $1/d$ of the weight, where d is the separation to the neighbor [58].

The area is acquired from an arrangement of known objects of an assortment of items (for k -NN classification) or protest characteristic qualities (for k -NN regression). This can be

considered as a learning set for the calculation; however, no express preparing venture is required.

The detriment of the k-NN calculation is that it is delicate to the neighborhood structure of the data. The calculation ought not to be confused with k-implies, another prominent learning method.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Hadoop version 2.6.0, CentOS Linux 7(Final) OS, VMware Workstation 10.0.2, Eclipse IDE JAVA version jdk 1.7.0_79, mac OS, KNIME 3.1 have been used for the experiments.

4.1. System Design

Data analysis in this research is conducted based on the steps shown in Figure 4.1. Generally, the analysis is divided into five consecutive steps as follows:

- Social media platform selection: The choice of this research work is based on Twitter for data collection. The collected data is fed to the KNIME tool for preprocessing.
- Saving data to Hadoop: KNIME tool sends the preprocessed collected data to a local server (Hadoop) for word counting, which is done using MapReduce function.
- Processed data transfer: The result of the words count is transferred back to KNIME for further analysis.
- Classification: The received processed data from Hadoop at KNIME side is classified using k-NN and SVM.
- Visualization: The final stage in data analysis is evaluate the data using confusion matrix and cross validation methods in order to produce the visualized output.

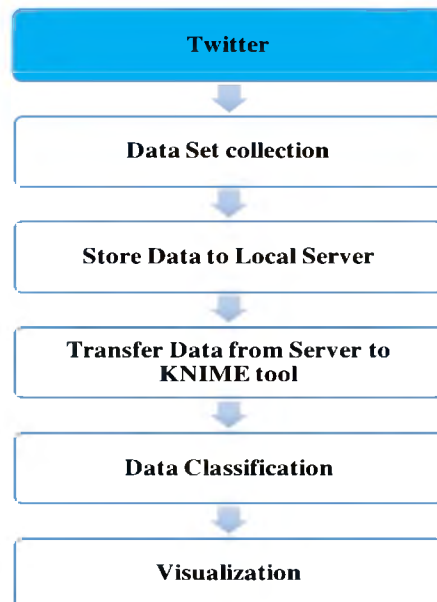


Figure 4.1 System design

4.2. Dataset

The first step is collecting the datasets from twitter. The datasets which have been used in this research are fetched from the Internet in real-time. The model is based on the current date and time of data receiving and it follows it in an up-to-date schema. It gathers text fields as input vectors from the most frequently written phrases on Twitter. This data set is followed by a test set in the same type using cross-validation method. A random percentage of the training set is tested and scored against the error rate and performance.

Twitter has been used as the main source of data input in this research. There has been a collection of about 1k to 5k tweets from the USA, where the data was obtained from (tender or eHarmony) set. Although the tweets expire in one or two days, the main focus was on testing and evaluating the proposed work. As it is the main case, there is data about global warming and major worldwide topics throughout Twitter. The timing is captured but it is less important due to the short lifetime of tweets. The tweets data was loaded from Twitter by an API connector. Here, Twitter APIs have two types for user access (REST APIs and streaming APIs). Figure 4.2 illustrates how the users request data from the Twitter website through HTTP server by using REST API [61].

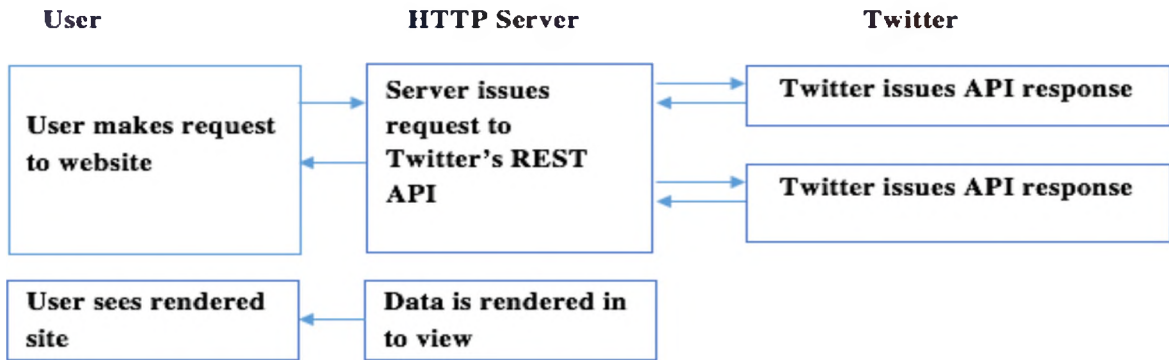


Figure 4.2 Twitter API

4.3. Data Preparation

In this project, KNIME tool is used for analyzing tweets data obtained from Twitter website. The choice was made for using KNIME tool as it is a free open source, and it is more user friendly and newer than all of tools that work with data mining. KNIME tool consists of many

nodes each of which performs a particular task. The node in the KNIME tool is considered as the basic processing unit. In order to obtain the datasets from Twitter, a Twitter application form has to be created. Creating the required Twitter application is shown in Figure 4.3.

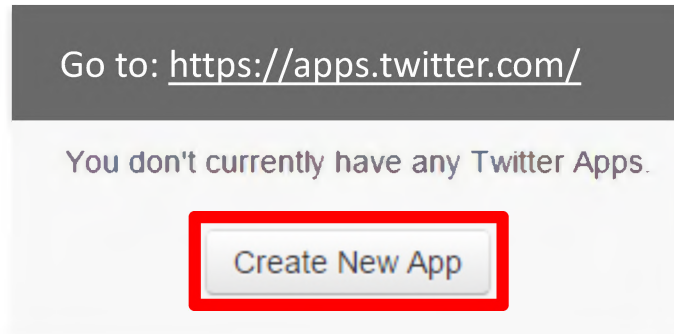


Figure 4.3 Create new application in Twitter

After the application has been created, the application should be opened and managed. This is due to the fact that once an application is managed, the followings can be activated and used: API key, API key secret, access token, and the access token secret which are necessary for accessing data in KNIME tool. Figure 4.4 shows how to get API key, API key secret, access token, and the access token secret.

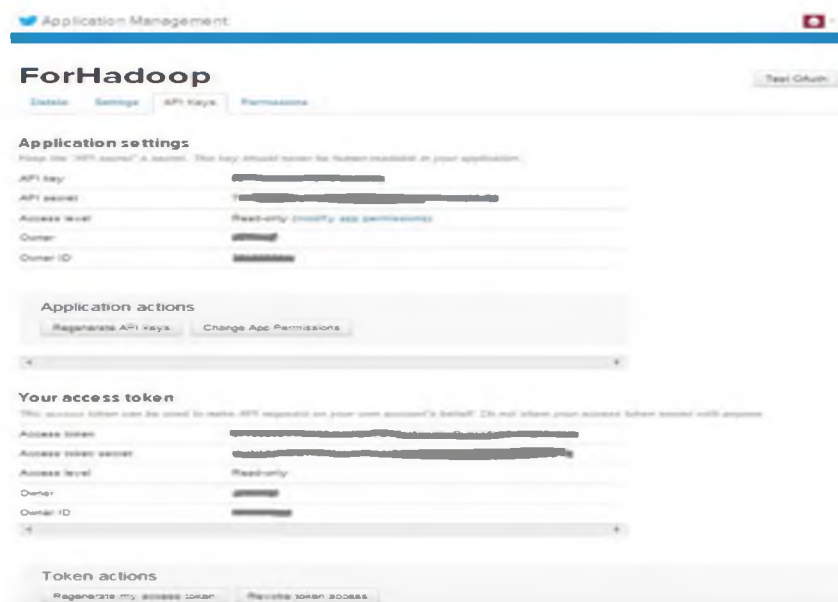


Figure 4.4 Twitter application management console

Twitter API Connector (API) is the essential node employed in this project for connecting KNIME tool and Twitter platform. The applied the configuration of Twitter API Connector (API) node is shown in Figure 4.5.

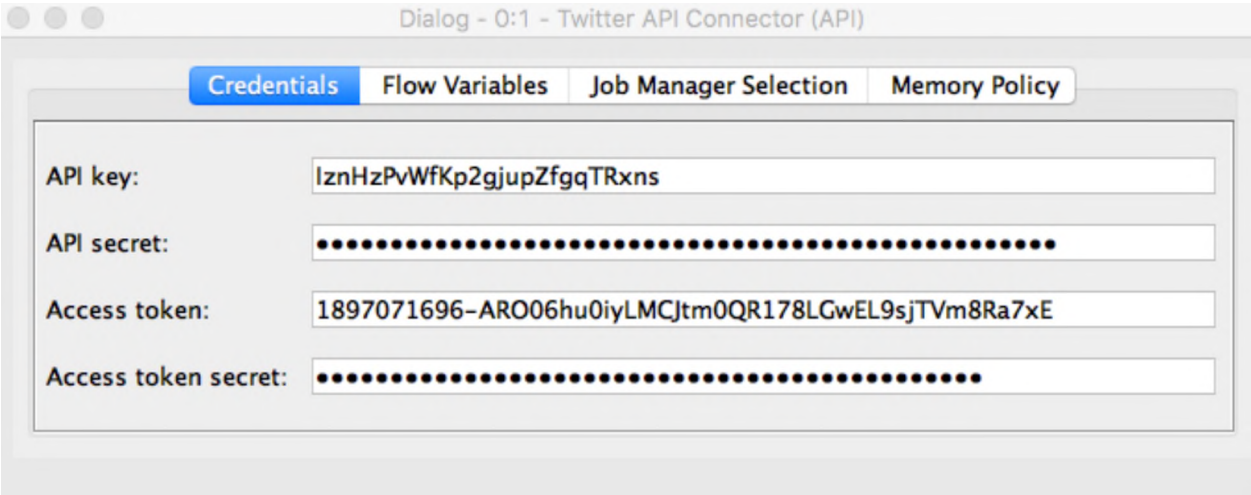


Figure 4.5 Twitter API Connector (API)

Figure 4.5 shows the required four necessary fields that needed to be configured. These are API key, API key secret, access token, and access token secret, which are generally appearing in the first step of creating a Twitter application according to the explained methodology.

The Twitter Search node allows the search for tweets on Twitter and download the specified searched data to KNIME tool.

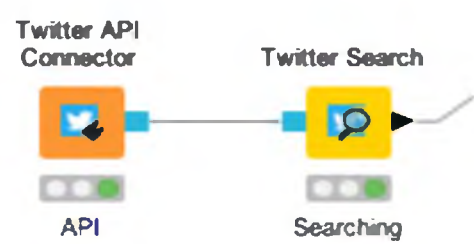


Figure 4.6 Searching data from Twitter

Figure 4.6, shows two nodes, the “Twitter API Connector”, and the “Twitter Search” which allows the user to search for tweets in KNIME tool.

4.4. Preprocessing

Preprocessing is an important step in this project because of fully filtered raw data acquired from Twitter. Here, the acquired data has a lot of undesired parts which should be eliminated by the preprocessing steps. Some of the preprocessing steps are well-known such as separating the tokens, parts of speech, and stop of words which include tagging, extraction, and representation.

Stop words are not added to the analysis and subsequently are removed amid pre-processing steps. While the procedure is caused by different administrators to bring parts of speech and words to the root structure. Parts of speech in tagging content are crucial to use in the dialect discourse to detect different components. Due to their poor condition, printing data are a big challenge. They are often added to the pre-processing and it is one of the steps required to remove the highlights and this requires a compelling level. Besides, removing the highlights, highlight identification is fundamental in achieving any analysis.

The workflow removes punctuation marks such as (., ; - .) and other commands (@,\$,£....). Then (httpcomlrtlvlialht) were removed using “rexes filter” node. And then the document was tagged with part of speech to understand the language of your topic (noun, verb, adv.....). And then they were added to a word bag with separation (noun, verb, adv., and adjective...), and then the document was tagged with stop of word node to remove stop of word words like (a, the, of, from...) [59]. That is shown in Figure 4.7.

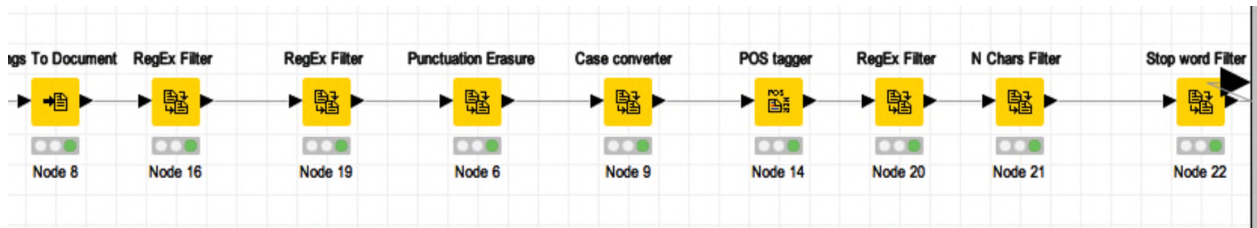


Figure 4.7 Preprocessing step in KNIME tool

Figure 4.8 shows the tweets before preprocessing that contain many number, punctuation marks, and uppercases.

Row ID	S Tweet
Row0	#ThankYouUndertaker – my life has been so much better for having known you. #Wrestlemania https://t.co/earZXZ9uYF
Row1	Like I always say, chase your dreams because dreams do come true, and this ...
Row2	SHE HAD TO WALK 24 HOURS TO FIND WATER & FOOD. YET SHE IS GRATEFUL...
Row3	RT @ELHAE: ELHAE • Aura 2 • 4.14.17 • RT to save a life... https://t.co/48X...
Row4	@thatkidtelles2 greatest moment of my life
Row5	Handmade #TreeofLife clock in wood #Zen #GiftsIdeas #Buddha
Row6	RT @phanlands: hello @doddleoddle i hope u dont mind me turning ur life int...
Row7	RT @changes: life gets a whole lot more beautiful once you start living for your...
Row8	RT @loneblockbuster: Films are a great way to temporarily escape the un-rele...
Row9	RT @RealMickFoley: #ThankYouUndertaker – my life has been so much better ... #Wrestlemania https://t.co/earZXZ9uYF
Row10	RT @forgivejacob: JUSTIN AND ALEX ARE DATING IN REAL LIFE OH MY GOD #...
Row11	You can have everything in life you want, if you will just help enough other peo...
Row12	RT @BCSxPSS_1976: "As an ultra I identify myself with a particular way of life. ...

Figure 4.8 Tweets before preprocessing

Finally, this preprocessed step was completed and shown as result in Figure 4.9.

Row ID	Document
Row0	"thankyouundertaker my life has been so much better for having known you wrestlemania htt...
Row1	"like i always say chase your dreams because dreams do come true and this was one of thos...
Row3	"rt elhae aura rt to save a life https"
Row4	"greatest moment of my life"
Row7	"rt life gets a whole lot more beautiful once you start living for yourself and accept the fact th...
Row10	"rt justin and alex are dating in real life oh my god why https"
Row11	"you can have everything in life you want if you will just help enough other people get what th...
Row13	"trouble dont last always bitches better live life while you can pops gum"
Row14	"rt why do we close our eyes when we pray cry kiss or dream because the most beautiful thi...

Figure 4.9 Results after preprocessing

Figure 4.9, the illustration of how the preprocessing steps operate on the data received from Twitter through the KNIME nodes. Also, Figure 4.9 creates a new document that labels searched tweets data by row numbers, where the second column is presenting the preprocessed data.

4.5. Saving Files

Hadoop is a framework of open source software. It is not a unique software that can be downloaded and installed on a computer; rather, it is a framework used for running processing applications on big data.

In this project, the open source set of Hadoop tools has been used for distributed processing of large datasets. Here, it is designed to scale up from a single server to a hindered machine as each of them has its own storage and computation. Rather than relying on a hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer. Therefore, this delivers a highly-available service on top of a cluster of computers, each of which may prone to failure.

The project includes the following modules:

Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data. It is designed for storing a large file that could be in the range of megabyte or gigabyte in size.

Hadoop MapReduce: A YARN-based system for parallel processing of large datasets.

4.5.1. MapReduce

The data which has been loaded from Twitter should be saved in a file. The saved data should be then split into independent chunks. These chunks are processed by the map function in a completely parallel manner. The framework sorts the outputs of the maps, which are then fed to the reduce function. Typically, both the input and the output of the job are stored in a file-system. The framework takes care of the scheduling tasks where it monitors them and re-executes the failed tasks. Figure 4.10 shows MapReduce workflow.

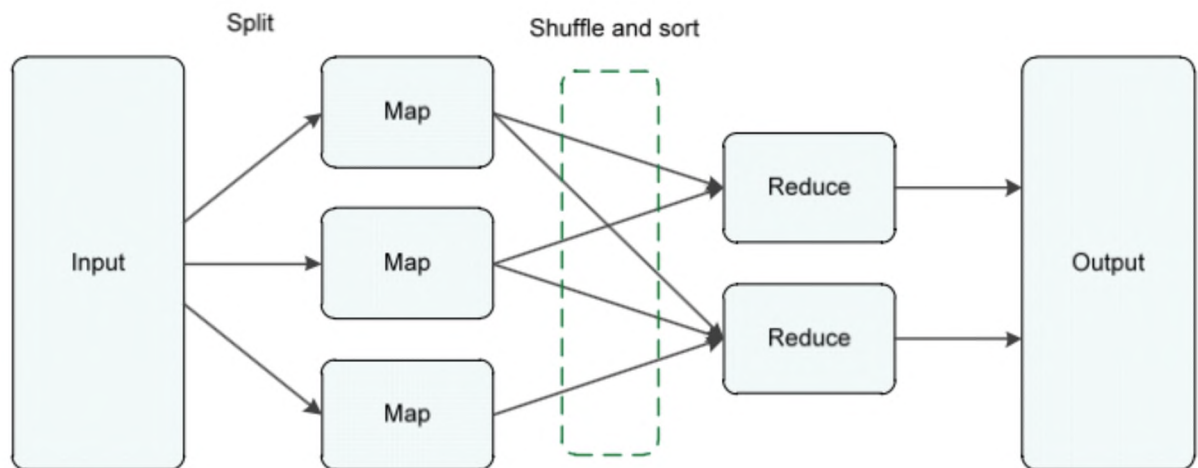


Figure 4.10 MapReduce workflow

- **Operating the MapReduce**

Input: each line of the text

Output: (Label, word _ count)

Map (docID, line):

Label=line= [0]

Text=Pre_Processing ([line [1:]])

Text Array=text. Split ('')

Emit (label,textArray.length)

The mapper receives every single tweet as an input. After processing the tweets and splitting them by spaces, an array of the words is obtained from the saved file of the processed tweets. In this mapper, the word number is simply emitted along with the label. For example, if the input tweet is:

“I like the movie”

After preprocessing, the following is obtained:

“Like movie”

Then the output of mapper is

“Word 2”

This means this tweet it has 2 valid words.

The output will be written into intermediate file and passed into the Label Reducer.

Label Reducer

Pseudo for Reduce:

Input” (label, word_ count)

Output” (label, tweet number: word number)

Reduce (label, int [])

UniqueLableNumber++

For count in int []:

Reduce (label, int [])

UniqueLabelNumber++

For count in int []:

wordNumWithLabelY++

wordNumWithLabelY+=count

emit (label,docNumWithLabelY: wordNumWithLabelY)

The reducer takes all the word counts corresponding to the same label as input. The input is also restructured by shuffle and sort process before passed into the reducer. In the reducer, there is another global variable named UniqueLabelNumber, which is the number of all unique labels, because each reducer only accepts one kind of label. For the word count array, it will iterate

over each element and add the value to a variable wordNumWithLabelY, which is the number of all words in the tweets with label Y. At the same time, another variable docNumWithLabelY will be incremented by one to calculate the number of the tweet with label Y. At last, the variable docNumWithLabelY and wordNumWithLabelY will be combined together and emitted out along with the label [60].

After the MapReduce has processed the data, it creates the output file in the Hadoop server. This file can be downloaded to the KNIME tool by using the Secure Shell (SSH) Connection node, as shown in Figure 4.11

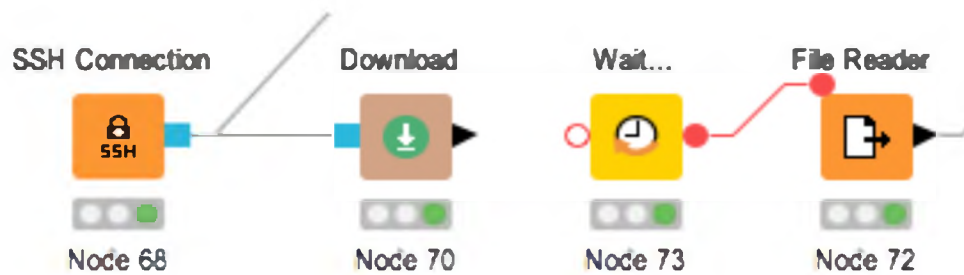


Figure 4.11 Get data from Hadoop to KNIME

In Figure 4.11, the SSH Connection node is used for connecting the Hadoop server and KNIME tool. It should be configured by inserting the host name and the password of the server. Then, the desired file can be downloaded to the KNIME using the Download node. The KNIME here processes the output file that has been retrieved from the server.

4.6. Prior Polarity Scoring

A number of the used features is based on prior polarity of words that depend on the dictionary which was created for this study. The dictionary has about 4500 English words. It contains negative and positive English words. The scores should be first normalized by diving each score by the scale.

In this research, the words with a polarity less than or equal to -1 are considered as negative, higher than or equal to 1 are considered as positive and the rest of words are considered as neutral. Figure 4.12 provides the workflow that presents how the data are being labeled.

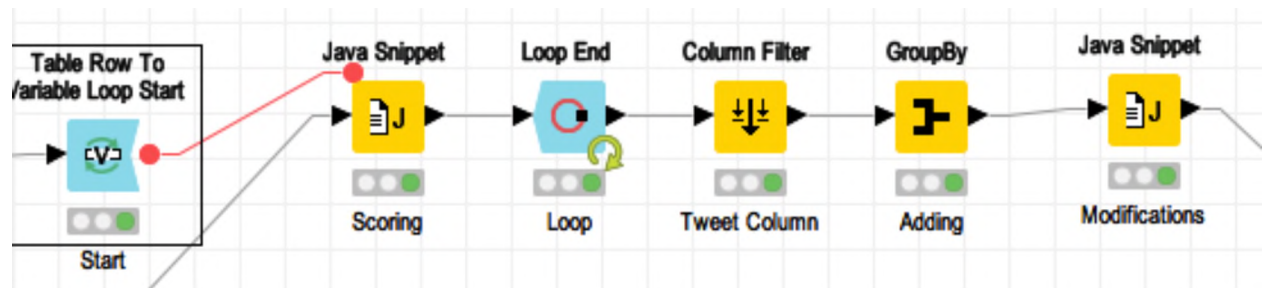


Figure 4.12 Prior polarity scoring

Figure 4.12 shows how a group of nodes is used to label the data. This workflow is used in this project for distinguishing the positive, negative, and neutral data and then modifying the data using the JavaScript node.

S TweetText	Sum(V...	S String...
"your voice is so tender can you say my name at least it aint that"	1	Positive
"rt high court dismisses voter register audit case against the k...	0	Natural
"youre going to have no problem in succeedinghe groaned agai...	2	Positive
"rt khai photos of daughter stephanie showing off her tender fl...	1	Positive
"rt sekarang gini kenapa pengadaan kertas dan alat tulis di ka...	1	Positive
"rt father i need you your soft tender touch in my life i know yo...	2	Positive
"the bar tender wants to get us drunk"	1	Positive
"opposition loses case to block kpmg voter audit tender httpsel...	-2	Negative
"how toward blast away away from suppliers legal tender onyx ...	-2	Negative
"model toys eztec scientific toy g gauge g scale coal tender nor...	0	Natural
"rt muhammed was tender kind towards cats he appreciated ...	7	Positive
"rt opposition loses case to block kpmg voter audit tender http...	-2	Negative
"love me tender"	2	Positive
"rt eager blowjob and tender swallow httpsblowjob blowjob s...	1	Positive
"i am extremely sorry that money to be broken in and put an i...	0	Natural
"rt father i need you your soft tender touch in my life i know yo...	2	Positive
"rt video namba ni tender ya kee akimshirikisha tekno ngazk...	1	Positive
"whoever got tender to frame those pictures of the president a...	-2	Negative
"baby baby baabeim coming home girl to your tender sweet lo...	3	Positive
"rt khai photos of daughter stephanie showing off her tender fl...	1	Positive
"rt breaking energy minister gone tina who mistakenly reveale...	-3	Negative
"rt father i need you your soft tender touch in my life i know yo...	2	Positive
"but for real how tender is this shit couples on tv are so over dr...	1	Positive
"bodohnya ente jemb semanggi pakai csr ngapain tender masi...	-1	Negative
"rt video namba ni tender ya kee akimshirikisha tekno ngazk...	1	Positive
"rt baby baby baabeim coming home girl to your tender sweet...	3	Positive
"rt whoever got tender to frame those pictures of the presiden...	-2	Negative
"tender carnation by georgiana romanovna httpstcojkdvjrrbpg"	0	Natural
"i may have put in an early morning tender for a short shift with...	1	Positive
"rt sekarang gini kenapa pengadaan kertas dan alat tulis di ka...	1	Positive
"rt kpc paid per cent of the value of the tender just for paper...	0	Natural
"rt father i need you your soft tender touch in my life i know yo...	2	Positive
"rt doodle	0	Natural
"rt but for real how tender is this shit couples on tv are so over...	1	Positive
"rt sfiso buthelezi is a former prasa chairman who presided ov...	3	Positive
"rt"	0	Natural

Figure 4.13 Labeling tweets

Figure 4.13 shows original tweets with labeled sentiments. The tracked sentiments are in positive, negative, and neutral classes.

4.7. Text Classification

There are various classification algorithms that can be found in the literature. However, this project applies some of the available classification methods in KNIME, such as SVM, k-NN, etc. In this project, the target is the sentiment data labeling, as it is the case in all supervised machine learning. This data is stored as a category in the used documents. The target is extracted from these documents and appended as a string column, using the "Category to class" option. Then the documents are labeled positive, negative or neutral sentiment based on these labels.

Text Classifier node collection provides a dictionary-based classifier for text documents. Using a set of labeled sample documents, one can build a dictionary and use it to classify uncategorized documents.

- SVM machine learning algorithm

In this study, this algorithm has also used two nodes for classification, namely SVM Learner and SVM Predictor with HyperTangent kernel function, this function has two parameters: delta with value 0.1 and kappa with value 0.5. The SVM Learner node builds an SVM PMML (Predictive Model Markup Language) from an input training data model. The SVM Predictor node is responsible for classifying the labeled dataset based on SVM PMML into three classes (positive, negative, and neutral). Here, the evaluation result should be one of the following three categories: +1 (positive sentiment), zero (neutral sentiment), and -1 (negative sentiment).

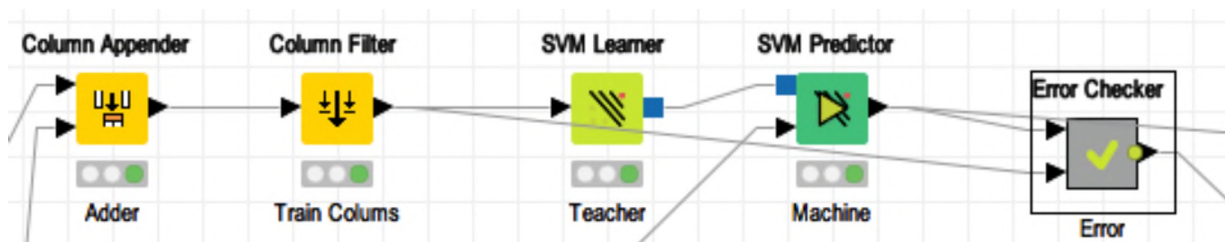


Figure 4.14 SVM classification

Figure 4.14 presents the SVM deployment using two nodes, namely the SVM Learner and SVM Predictor to text classification.

- k-NN machine learning algorithm

It is the simplest classifier that mainly depends on the category labels. k-NN uses the parameter “ k ” in classifying the datasets, in this application the number of neighbor is 3 . The main concept behind the k-NN is first training the system using the existing data. Then, finding a predefined number of training sample nearest in distance to the new point and estimate the labels from these. The amount of samples can be user-defined constant and the common alternative for similarity measure is the Euclidean distance equation. In Figure 4.15, the presentation of the used k-NN classification in this work has been demonstrated.

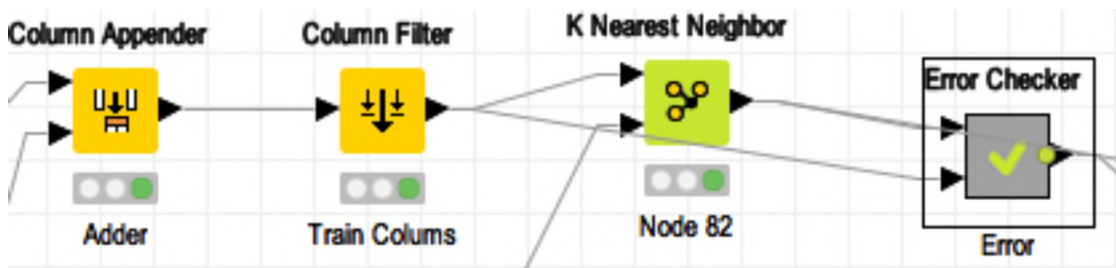


Figure 4.15 K-NN classification

Figure 4.15 shows the k-NN classification method that uses the k Nearest Neighbor node for classification and the Error node for error checking.

4.8. Experimentation and Results

This part explains the produced results of the research work on sentiment classification for Twitter. The dataset is dependent on positive, negative and neutral words for each of the sentiment classification. This classification is done using the machine learning algorithms SVM and k-NN.

Twitter has been used as a main source for raw data in this project. There has been a collection of 1k to 5k tweets from USA Twitter data. The tweets expire in one or two days, yet they will be used for testing and evaluating the project. Generally, there is data about global warming and major worldwide topics throughout Twitter. The timing is captured but it is less important due to the short lifetime of the tweets. After preprocessing and scoring have been achieved, data is fed to the classification phase where SVM and k-NN are used. In this project, for evaluating the

effectiveness of machine learning methods, the following measure are used: precision, recall, specificity, and accuracy, based on confusion matrix and cross validation methods.

- **Confusion Matrix:**

A confusion matrix is a table that is often used to describe the performance of a classification model or a classifier. Confusion matrix comprises three cases: Sensitivity, Specificity and Accuracy. Confusion matrix includes four cases True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). These values are utilized to calculate the Sensitivity, Specificity and Accuracy. TP is a properly reported true positive data test result. FP is any error reported in the data indicating improper test results, which are negatives as positives. TN is a properly reported true negative data test result. FN, false negative is any error reported in the data indicating improper test results, which are positives as negatives.

Table 4.1. Confusion matrix

	Predicted condition	
	Predicted condition positive	Predicted condition negative
	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

The confusion matrix has dimension of NxN (3x3) table containing TP, TN, FP, FN, TE and FE as elements.

Table 4.2 confusion matrix for three classes

	Predicted condition		
	Predicted condition positive	Predicted condition negative	Predicted condition Neutral
	True Positive (TP)	False Neutral(FE)	False Negative (FN)
	False Neutral(FE)	True Negative (TN)	False Positive (FP)
	False Negative (FN)	False Positive (FP)	True Neutral (TE)

$$Precision = \frac{TP}{TP+FE+FN} \quad (4.1)$$

$$Recall = \frac{TN}{TN+FE+FP} \quad (4.2)$$

$$Specificity = \frac{TP+FE+FE+TN}{TP+FE+FE+TN+FN+FP} \quad (4.3)$$

$$Accuracy = \frac{TP+TN+TE}{Total} \quad (4.4)$$

- **Confusion Matrix for SVM:**

This confusion matrix is obtained after the SVM implementation. Table 4.3 shows the confusion matrix for SVM classification based on 5k tweets. The predicted numbers of TP, TN, FP, FN, FE and TE are presented in Table 4.3.

Table 4.3. Confusion matrix for SVM classifier based on 5k tweets

		Classified		
		Positive	Negative	Neutral
Actual	Positive	448	350	228
	Negative	350	2522	52
	Neutral	228	52	1400

Also, the precision, recall, specificity and accuracy are found as follows:

$$Precision = \frac{448}{448 + 350 + 228} = 43.6 \%$$

$$Recall = \frac{2522}{2522 + 350 + 52} = 86.2 \%$$

$$\text{Specificity} = \frac{448 + 350 + 350 + 2522}{448 + 350 + 350 + 2522 + 228 + 52} = 92.9\%$$

$$\text{Accuracy} = \frac{448+2522+1400}{5000} = 87.4 \%$$

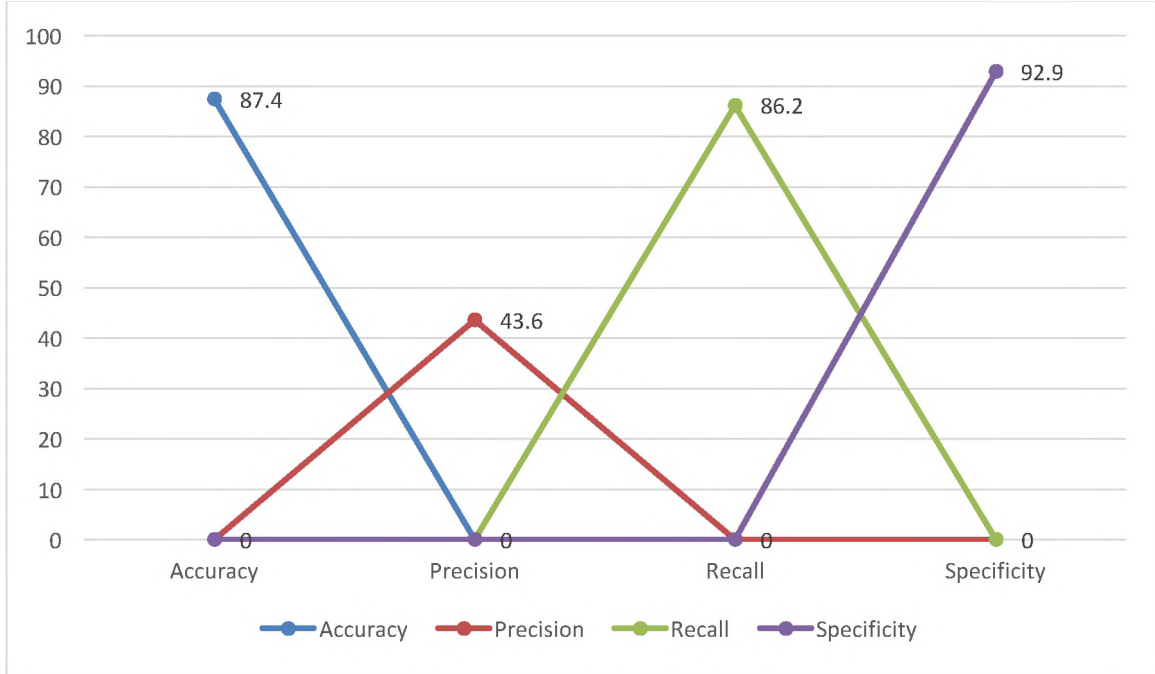


Figure 4.16 SVM classifier based on 5k tweets by using confusion matrix method

Figure 4.16 shows the confusion matrix for SVM classifier based on 5k tweets as input data. The accuracy, precision, recall and specificity of SVM method are obtained as 87.4%, 43.6%, 86.2%, and 92.9% respectively.

Table 4.4 shows the confusion matrix for SVM classification based on 1k tweets. The predicted numbers of TP, TN, FP, FN, FE and TE are presented in Table 4.4.

Table 4.4. Confusion matrix for SVM classifier based on 1k tweets

		Classified		
		Positive	Negative	Neutral
Actual	Positive	180	30	45
	Negative	30	535	40
	Neutral	45	40	170

Also, the precision, recall, specificity and accuracy are found as follows:

$$\text{Precision} = \frac{180}{180 + 30 + 45} = 70.5 \%$$

$$\text{Recall} = \frac{535}{535 + 30 + 40} = 88.4 \%$$

$$\text{Specificity} = \frac{180+30+30+535}{180+30+30+535+45+40} = 91.1\%$$

$$\text{Accuracy} = \frac{180+535+170}{1000} = 88.5\%$$

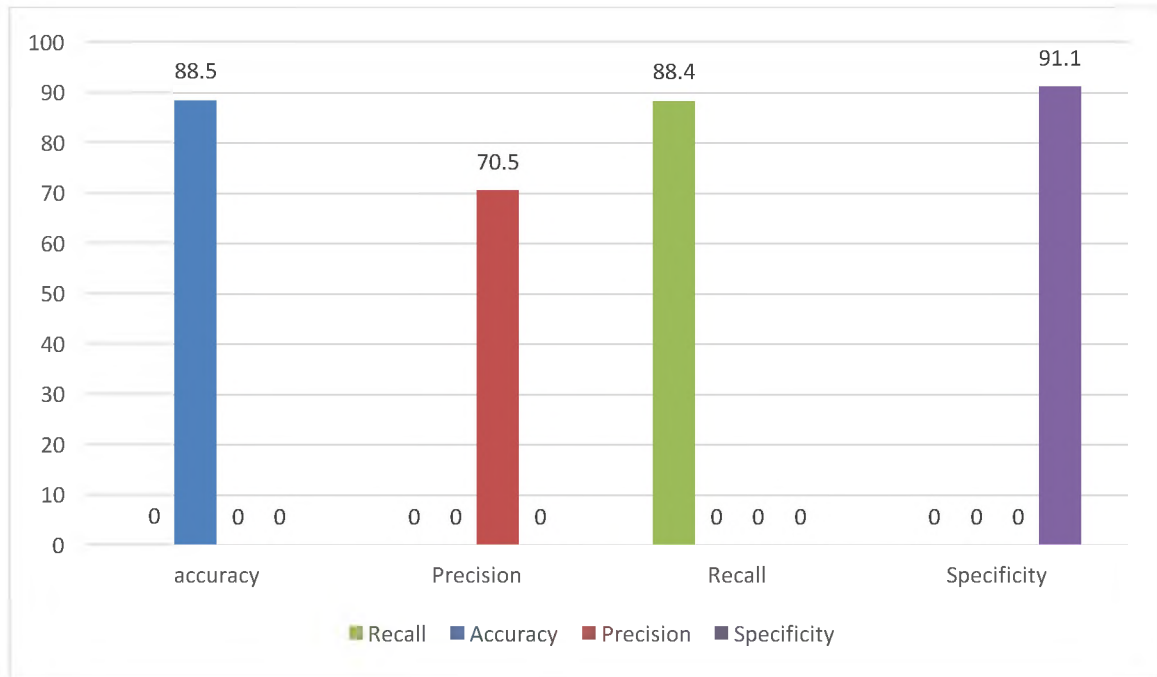


Figure 4.17 SVM classifier based on 1k tweets by using confusion matrix method

Figure 4.17 shows the confusion matrix for SVM classifier based on 1k tweets as input data. The accuracy, precision, recall, and specificity of SVM method are obtained as 88.5%, 70.5%, 88.4%, and 91.1 respectively.

- **Confusion Matrix for k-NN**

This confusion matrix is obtained after k-NN implementation. Table 4.5 shows the confusion matrix for k-NN classification based on 5k tweets. The predicted numbers of TP, TN, FP, FN, FE and TE are presented in Table 4.5.

Table 4.5. Confusion matrix for k-NN classifier based on 5k tweets

		Classified		
		Positive	Negative	Neutral
Actual	Positive	543	74	241
	Negative	74	3259	57
	Neutral	241	57	826

Also, the precision, recall, specificity and accuracy are found as follows:

$$\text{Precision} = \frac{543}{243 + 74 + 241} = 97.3\%$$

$$\text{Recall} = \frac{3259}{3259 + 74 + 57} = 96.1 \%$$

$$\text{Specificity} = \frac{543 + 74 + 74 + 3259}{543 + 74 + 74 + 3259 + 241 + 57} = 92.9\%$$

$$\text{Accuracy} = \frac{543 + 3259 + 826}{5000} = 92.5 \%$$

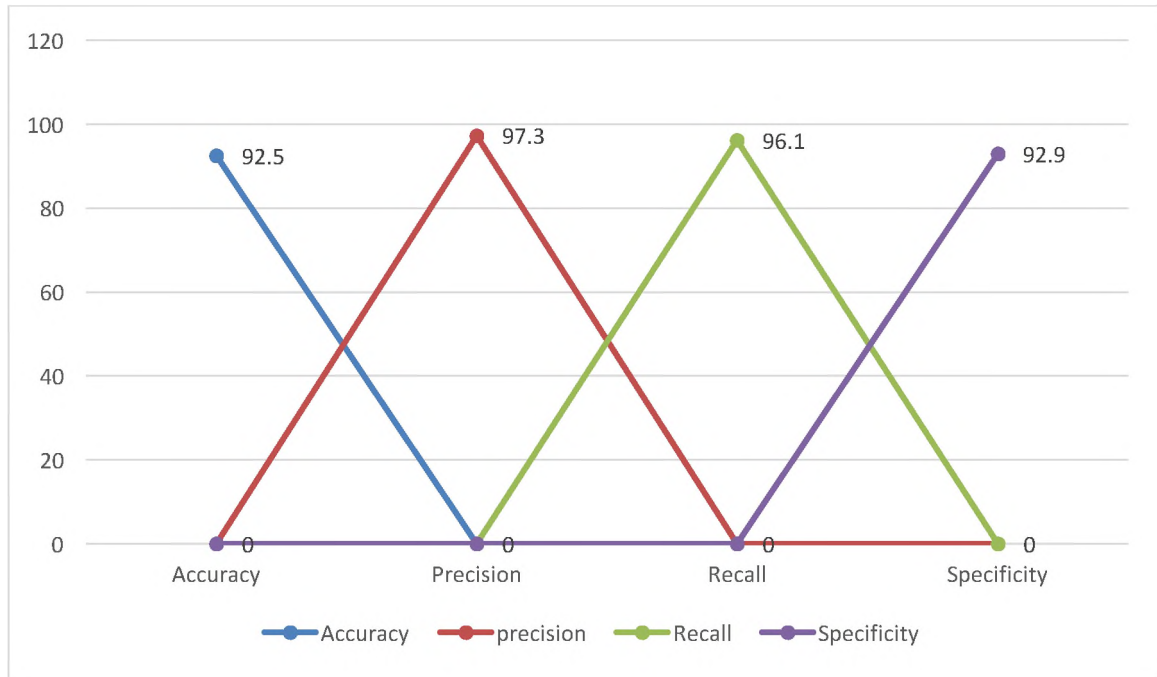


Figure 4.18 K-NN classifier based on 5k tweets by using confusion matrix method

Figure 4.18 shows the confusion matrix for k-NN classifier based on 5k tweets as input data. The accuracy, precision, recall, and specificity of k-NN method are obtained as 92.5%, 97.3%, 96.1%, and 92.9% respectively.

Table 4.6 shows the confusion matrix k-NN classification based on 1k tweets The predicted numbers of TP, TN, FP, FN, FE and TE are presented in Table 4.6.

Table 4.6. Confusion matrix for k-NN classifier based on 1k tweets

		Classified		
		Positive	Negative	Neutral
Actual	Positive	81	8	20
	Negative	8	750	19
	Neutral	20	19	122

Also, the precision, recall, specificity and accuracy are found as follows:

$$\text{Precision} = \frac{81}{81 + 8 + 20} = 74.3 \%$$

$$\text{Recall} = \frac{750}{750 + 8 + 19} = 96.5 \%$$

$$\text{Specificity} = \frac{81 + 8 + 8 + 750}{81 + 8 + 8 + 750 + 20 + 19} = 95.5\%$$

$$\text{Accuracy} = \frac{81+750+122}{1000} = 95.3\%$$

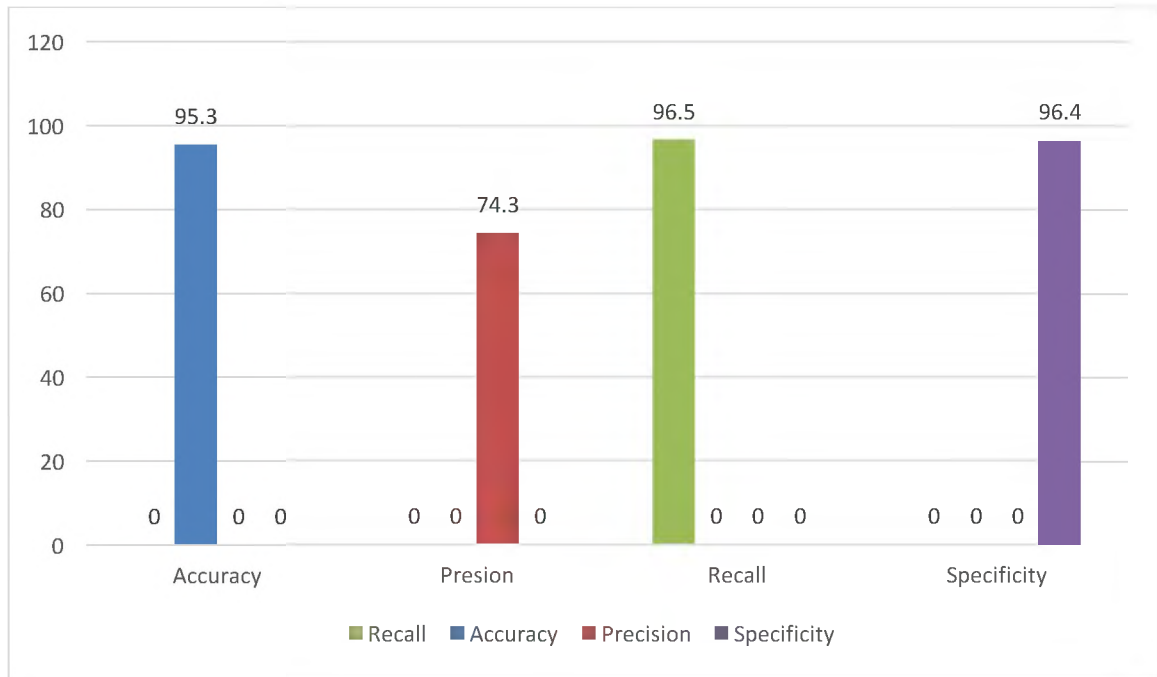


Figure 4.19 K-NN classifier based on 1k tweets by using confusion matrix method

Figure 4.19 shows the confusion matrix for k-NN classifier based on 1k tweets as input data. The accuracy, precision, recall and specificity of k-NN method are obtained as 95.3%, 74.3%, 96.5% and 95.5% respectively.

The accuracy, precision, recall and specificity of each of the used classification methods have been obtained. This procedure is done using the confusion matrix evaluation for k-NN and SVM for 1k and 5k tweets, as provided in Table 4.7.

Table 4.7. Classification accuracy for 1k, 5k tweets dataset

Machine learning algorithm	Data size	Accuracy	Precision	Recall	Specificity
SVM	1000	88.5%	70.5%	88.4%	91.1%
SVM	5000	87.4%	43.6%	86.2%	92.9%
k-NN	1000	95.3%	74.3%	96.5%	95.5%
k-NN	5000	92.5%	97.3%	96.1%	92.9%

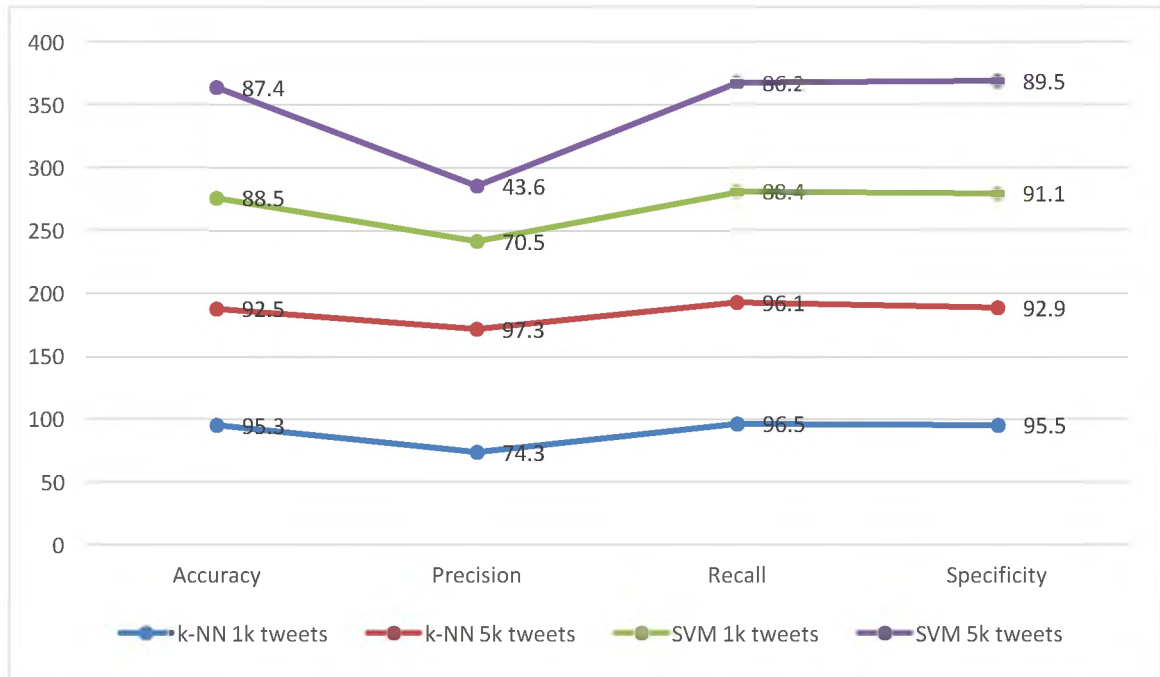


Figure 4.20 Classification accuracy, precision, recall and specificity for 1k,5k tweets data set

Figure 4.20 shows the classification accuracy, precision, recall and specificity for two sets of samples for 1k and 5k tweets datasets based on SVM and k-NN methods. By comparing these two results, the k-NN classifier is better than the SVM classifier. The accuracies for 1k and 5k using SVM are 88.5% and 87.4% respectively, and for k-NN are 95.3% and 92.5% respectively.

- **Cross-Validation Method**

Cross-validation is used for performance evaluation of machine learning methods. In this project, the performance evaluation of both algorithms is obtained using N cross validation process, where N is a number of folds that is equal to 5.

Figure 4.23 is representing the SVM classification results. The results here are based on 1k tweets processed and classified by the SVM classification method.

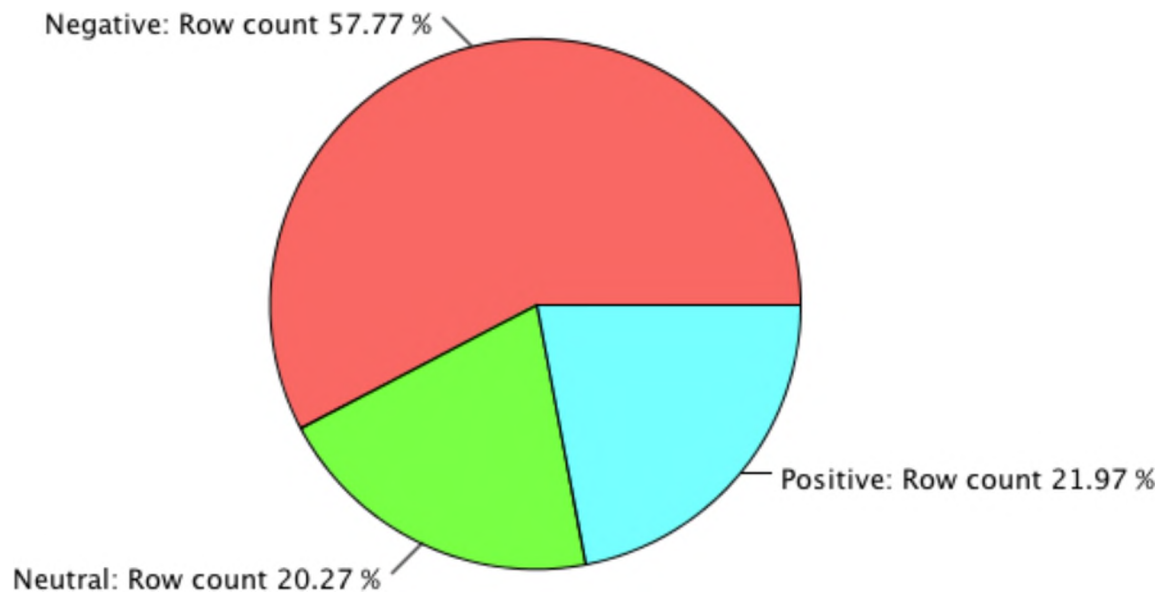


Figure 4.21 SVM classification results for 1k tweets by using cross validation method

In Figure 4.21, the output is distributed in such a way that (20.27%) is classified neutral, (21.97%) is classified positive and (57.77%) is classified negative. The accuracy of the used SVM is 83.73% based on the corresponding 5-fold cross-validation method.

k-NN classification result is shown in Figure 4.24. This result is obtained from 1k tweets where these tweets are processed and classified using the k-NN classification method. The value of k here was set to 3 for all experiments.

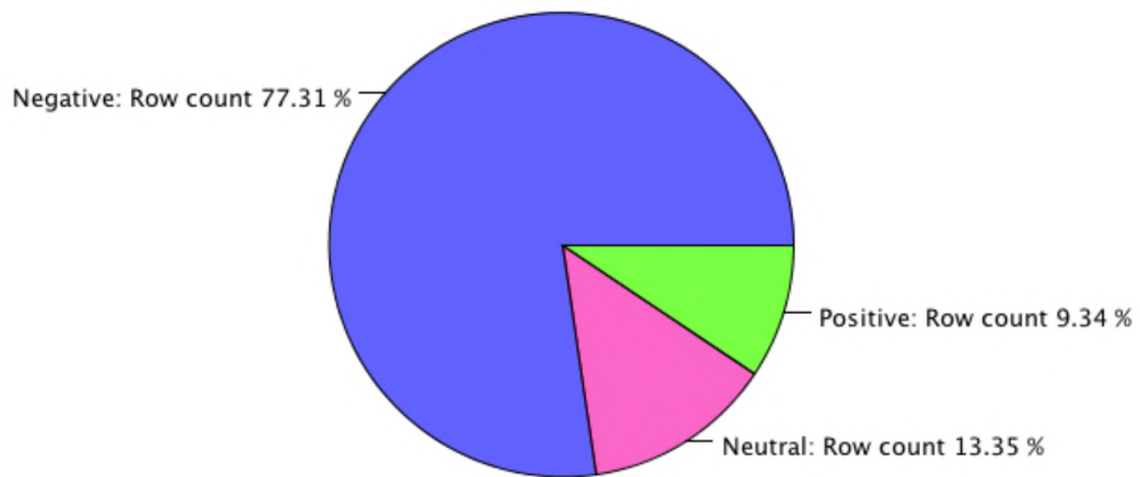


Figure 4.22 K-NN classification results for 1k tweets by using cross validation method

As shown in Figure 4.22, the output for 1k tweets data classification is providing that 21.97% of the opinions are positive, 57.77% of the opinions are negative and 20.27% of the opinions are neutral. The accuracy of the SVM is 83.73% based on the corresponding cross validation. When the k-NN machine learning classification algorithm was used for tweets, the analysis provided 9.34% of positive opinion, 77.31% of negative opinion and 13.35% of neutral. Here, the accuracy of the k-NN is 87.34% based on the 5-fold corresponding cross-validation.

The accuracy of each of the used classification methods has been obtained. This procedure is done using the cross-validation, and the results for k-NN and SVM accuracies are provided in Table 4.8.

Table 4.8. Classification accuracy for 1k tweets data set

Methods	The accuracy percentage of the tested data
	The percentage of each method
k-NN	87.34%
SVM	83.73%

Figure 4.23 shows the SVM classification results. The results here are obtained when 5k tweets are classified and tested using the SVM classification method. Basically, the produced output is classified to (38.79%) of natural, (6.41%) of positive and (54.8%) of negative. The accuracy of the SVM classification for 5k tweets is 80.13% based on the corresponding 5-fold cross-validation.

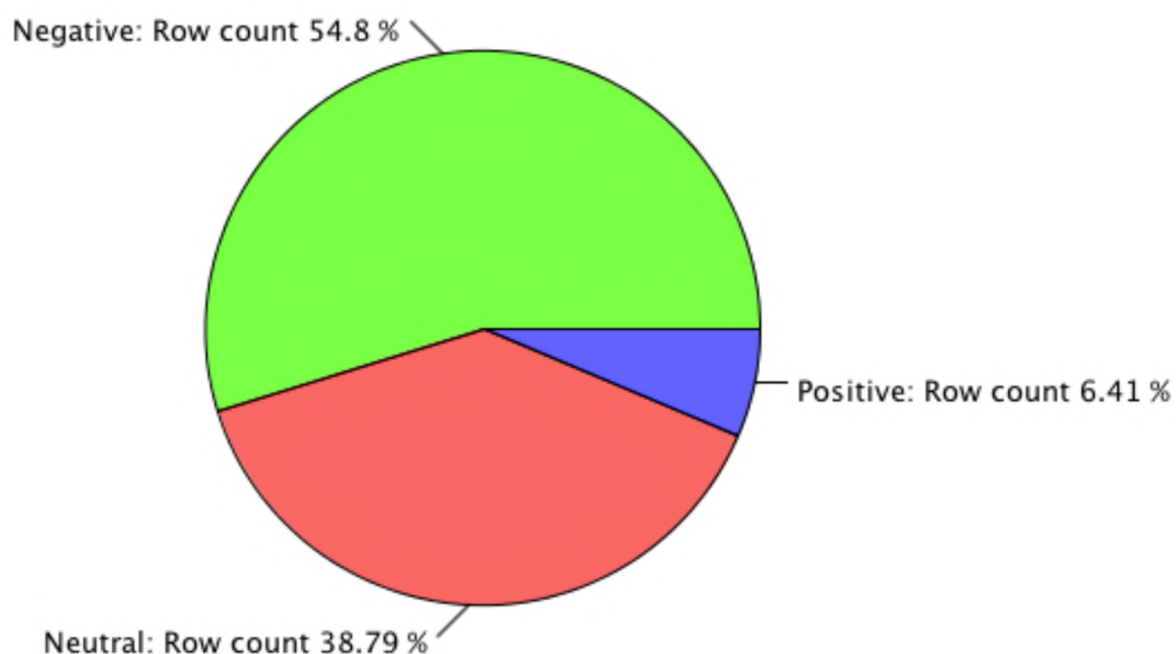


Figure 4.23 SVM classification results for 5k tweets by using cross validation method

Figure 4.24 shows the k-NN classification results. The results here are obtained when 5k tweets are classified and tested using the k-NN classification method.

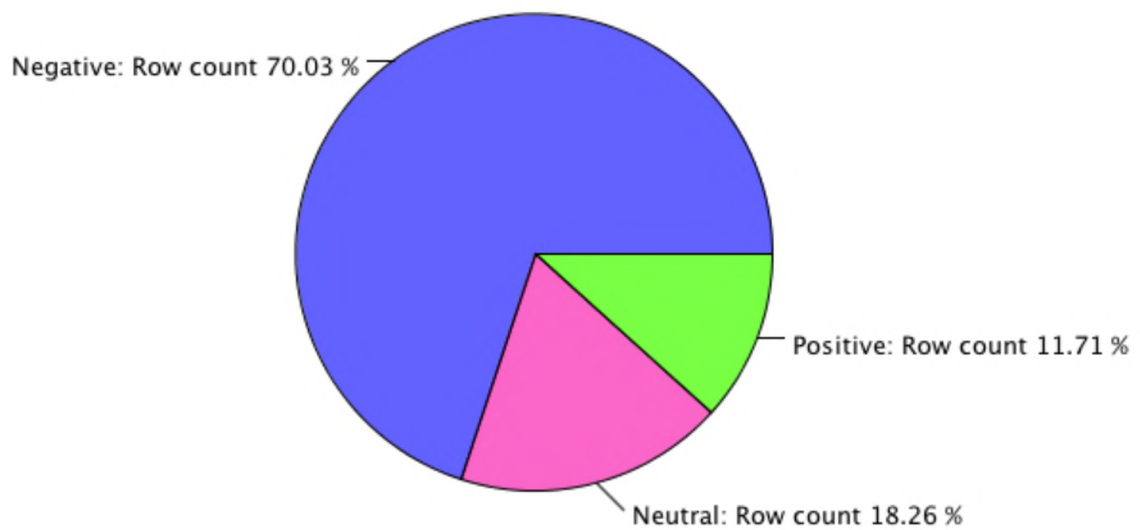


Figure 4.24 K-NN classification results for 5k tweets by using cross validation method

Figure 4.24 show the results for 5k tweets data. The results indicate that 6.41% of opinions is positive, 54.8% is negative and 38.79% is neutral. The accuracy of the SVM is 73.13% for the corresponding cross-validation. When the k-NN machine learning algorithm was used for classifying tweets data, the results indicate that 11.71% is classified as positive opinion, 70.13% is classified as negative opinion and 18.26% is classified as neutral. The accuracy of the k-NN is 85.01% based on the corresponding 5-fold cross-validation.

After testing the specified methods using 5k tweets as input data, the accuracy of each method based on the cross validation is calculated. This procedure is done using a special node that calculates the accuracy for each method. Different accuracy results are found for each of k-NN and SVM classifiers as provided in Table 4.9.

Table 4.9. Classification accuracy for 5k tweets data set

Methods	The Accuracy Percentage of the Tested Data
	The percentage of each method
k-NN	85.01 %
SVM	80.13 %

According to the results, k-NN provides a better classification results than SVM using the same datasets, as evaluated by both cross-validation and confusion matrix methods. Where the accuracies for 1k and 5k using SVM are 87.9% and 88.1% respectively, and for k-NN are 93.3% and 89.7% respectively as evaluated by confusion matrix method, and where The accuracy for 1k and 5k for SVM is 83.73%, 80.13% and for k-NN is 87.34%, 85.01%, as evaluated by cross-validation.

5. CONCLUSIONS AND RECOMMENDATION

In this proposed system, we have developed a system that analyze the tweets data using the Big Data and machine learning methods. The data sets which have been used in this research are fetched from the internet in real-time. The model is based on the current date and time of receiving data, following an up-to-date schema. It gathers text fields as input vectors from the most frequently written phrases from Twitter. The datasets used in this research are fetched from the Internet in a real-time basis. The model used in the research is employing the current date and time for receiving data and following an up-to-date schema. Twitter API was used to collect the data from Twitter. The collected data were inputted from Hadoop to KNIME tool. To manage the big data, we have used Hadoop technology which is developed in Java. Hadoop will mine the big data and provide us the sentiment based on the tweet we have. Machine learning algorithms used in this project for classification are SVM and k-NN. According to the results, k-NN provides a better classification results than SVM using the same datasets, as evaluated by both cross-validation and confusion matrix methods. The accuracies for 1k and 5k using SVM are 88.5% and 87.4% respectively, and for k-NN are 95.3% and 92.5% respectively as evaluated by confusion matrix method, and the accuracy for 1k and 5k for SVM is 83.7%, 80.13% and for k-NN is 87.34%, 85.01%, as evaluated by cross-validation.

6. REFERENCES

- [1] **Chetna Bachhav, Monika Gite, Komal Jadhav, Kamini Malode.,** 2015. Sentimental Analysis On Big Data, International Journal for Research in Engineering Application & Management (IJREAM), Vol-01, Issue 01, pp. 1-3
- [2] **Bernabé-Moreno, J., Tejada-Lorente, A., Porcel, C., Fujita, H. and Herrera-Viedma, E.,** 2015. CARESOME: a system to enrich marketing customers acquisition and retention campaigns using social media information. Knowledge-Based Systems, 80, pp.163-179.
- [3] **Harrigan, P., Soutar, G., Choudhury, M.M. and Lowe, M.,** 2015. Modelling CRM in a social media age. Australasian Marketing Journal (AMJ), 23(1), pp.27-37.
- [4] **Malthouse, E.C., Haenlein, M., Skiera, B., Wege, E. and Zhang, M.,** 2013. Managing customer relationships in the social media era: introducing the social CRM house. Journal of Interactive Marketing, 27(4), pp.270-280.
- [5] **Mohd Sharef, N., Mat Zin, H. and Nadali, S.,** 2016. Overview and future opportunities of sentiment analysis approaches for big data. Journal of Computer Science, 12(3), pp.153-168.
- [6] **Ha, I., Back, B. and Ahn, B.,** 2015. MapReduce functions to analyze sentiment information from social big data. International Journal of Distributed Sensor Networks, 11(6), p.417502.
- [7] **Tan, A.H.,** 1999, April. Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases (Vol. 8, pp. 65-70).
- [8] **Mei, Q. and Zhai, C.,** 2005, August. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 198-207). ACM.
- [9] **Park, K.M. and Hwang, K.B.,** 2011. A bio-text mining system based on natural language processing. Journal of KIISE: Computing Practices and Letters, 17(4), pp.205-213.

- [10] **Pang, B. and Lee, L.**, 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), pp.1-135.
- [11] **Kang, B., Song, M. and Jho, W.**, 2013. A study on opinion mining of newspaper texts based on topic modeling. *Journal of the Korean Society for Library and Information Science*, 47(4), pp.315-334.
- [12] **Borthakur, D.**, 2007. The hadoop distributed file system: Architecture and design. *Hadoop Project Website*, 11(2007), p.21.
- [13] **Chang, J.Y.**, 2013. Automatic retrieval of SNS opinion document using machine learning technique. *The Journal of The Institute of Internet, Broadcasting and Communication*, 13(5), pp.27-35.
- [14] **Kouloumpis, E., Wilson, T. and Moore, J.D.**, 2011. Twitter sentiment analysis: The good the bad and the omg!. *Icwsn*, 11(538-541), p.164.
- [15] **Tang, H., Tan, S. and Cheng, X.**, 2009. A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7), pp.10760-10773.
- [16] **Pak, A. and Paroubek, P.**, 2010, May. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREc* (Vol. 10, No. 2010).
- [17] **Go, A., Bhayani, R. and Huang, L.**, 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report*, Stanford, 1(12).
- [18] **Barbosa, L. and Feng, J.**, 2010, August. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 36-44). Association for Computational Linguistics.
- [19] **Han, J., Haihong, E., Le, G. and Du, J.**, 2011, October. Survey on NoSQL database. In *Pervasive computing and applications (ICPCA)*, 2011 6th international conference on (pp. 363-366). IEEE.
- [20] **Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M., Chandra, T., Fikes, A. and Gruber, R.E.**, 2008. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2), p.4.

- [21] **Sivasubramanian, S.**, 2012, May. Amazon dynamoDB: a seamlessly scalable non-relational database service. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 729-730). ACM.
- [22] **George, L.**, 2011. HBase: The Definitive Guide: Random Access to Your Planet-Size Data. " O'Reilly Media, Inc."
- [23] **Chodorow, K.**, 2013. MongoDB: the definitive guide. " O'Reilly Media, Inc."
- [24] **Khuc, V.N., Shivade, C., Ramnath, R. and Ramanathan, J.**, 2012, March. Towards building large-scale distributed systems for twitter sentiment analysis. In Proceedings of the 27th annual ACM symposium on applied computing (pp. 459-464). ACM.
- [25] **Bautin, M.**, Ward, C.B., Patil, A. and Skiena, S.S., 2010, April. Access: news and blog analysis for the social sciences. In Proceedings of the 19th international conference on World wide web (pp. 1229-1232). ACM.
- [26] **Lämmel, R.**, 2008. Google's MapReduce programming model—Revisited. Science of computer programming, 70(1), pp.1-30.
- [27] **Gillick, D., Faria, A. and DeNero, J.**, 2006. Mapreduce: Distributed computing for machine learning. Berkley, Dec, 18.
- [28] **Da, Z., Engelberg, J. and Gao, P.**, 2011. In search of attention. The Journal of Finance, 66(5), pp.1461-1499.
- [29] **Manyika, J.**, Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H., 2011. Big data: The next frontier for innovation, competition, and productivity.
- [30] **Dumbill, E.**, 2012. What is Big Data? An Introduction to the Big Data Landscape.[Blog post, January 2011.].
- [31] **Zikopoulos, P. and Eaton, C.**, 2011. Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.
- [32] **Zajicek, M.**, 2007, May. Web 2.0: hype or happiness?. In Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A) (pp. 35-39). acm.
- [33] **Ahlqvist, T., Bäck, A., Halonen, M. and Heinonen, S.**, 2008. Social media road maps exploring the futures triggered by social media: VTT Tiedotteita–Valtion Teknillinen Tutkimuskeskus.

- [34] **Asur, S. and Huberman, B.A.**, 2010, August. Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 492-499). IEEE.
- [35] **Sharef, N.M. and Haghanikhameneh, F.**, 2014, August. Content-Based Analysis Method for Sentiment Scoring in Microblogging Mining. In SoMeT (pp. 398-414).
- [36] **Agnihotri, R., Dingus, R., Hu, M.Y. and Krush, M.T.**, 2016. Social media: Influencing customer satisfaction in B2B sales. *Industrial Marketing Management*, 53, pp.172-180.
- [37] **Batrinca, B. and Treleaven, P.C.**, 2015. Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, 30(1), pp.89-116.
- [38] **Sharef, N.M.**, 2014. Review of sentiment analysis approaches in big data era. *Proceedings of the Malaysian National Conference of Databases, (NCD' 14)*, Serdang, pp: 7-12.
- [39] **Nirmal, V.J. and Amalarethinam, D.G.**, 2015. Parallel Implementation of Big Data Pre-Processing Algorithms for Sentiment Analysis of Social Networking Data. *International Journal of Fuzzy Mathematical Archive*, 6(2), pp.149-159.
- [40] **Aggarwal, C.C., Orlin, J.B. and Tai, R.P.**, 1997. Optimized crossover for the independent set problem. *Operations research*, 45(2), pp.226-234.
- [41] **Pang, B., Lee, L. and Vaithyanathan, S.**, 2002, July. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). Association for Computational Linguistics.
- [42] **Fei, X., Wang, H. and Zhu, J.**, 2010, August. Sentiment word identification using the maximum entropy model. In *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on* (pp. 1-4). IEEE.
- [43] **Whitelaw, C., Garg, N. and Argamon, S.**, 2005, October. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 625-631). ACM.

- [44] **Efron, M.**, 2004. Cultural orientation: Classifying subjective documents by cociation analysis. In AAAI Fall Symposium on Style and Meaning in Language, Art, and Music (pp. 41-48).
- [45] **Hatzivassiloglou, V. and McKeown, K.R.**, 1997, July. Predicting the semantic orientation of adjectives. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (pp. 174-181). Association for Computational Linguistics.
- [46] **Asur, S. and Huberman, B.A.**, 2010, August. Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 492-499). IEEE.
- [47] **Lee, H., Surdeanu, M., MacCartney, B. and Jurafsky, D.**, 2014, May. On the Importance of Text Analysis for Stock Price Prediction. In LREC (pp. 1170-1175).
- [48] **Riloff, E., Wiebe, J. and Wilson, T.**, 2003, May. Learning subjective nouns using extraction pattern bootstrapping. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 (pp. 25-32). Association for Computational Linguistics.
- [49] **Chaumartin, F.R.**, 2007, June. UPAR7: A knowledge-based system for headline sentiment tagging. In Proceedings of the 4th International Workshop on Semantic Evaluations (pp. 422-425). Association for Computational Linguistics.
- [50] **Yong, K.K., Mahmud, R. and Woo, C.S.**, 2011. Lexical Database for Multiple Languages: Multilingual Word Semantic Network. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 5(8), pp.819-824.
- [51] **Poria, S., Cambria, E., Ku, L.W., Gui, C. and Gelbukh, A.**, 2014, August. A rule-based approach to aspect extraction from product reviews. In Proceedings of the second workshop on natural language processing for social media (SocialNLP) (pp. 28-37).
- [52] **Xia, R., Zong, C. and Li, S.**, 2011. Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 181(6), pp.1138-1152.

- [53] **Khairnar, J. and Kinikar, M.,** 2013. Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications*, 3(6), pp.1-6.
- [54] **Ye, Q., Zhang, Z. and Law, R.,** 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3), pp.6527-6535.
- [55] **Fei, X., Wang, H. and Zhu, J.,** 2010, August. Sentiment word identification using the maximum entropy model. In *Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2010 International Conference on (pp. 1-4). IEEE.
- [56] **Pang, B. and Lee, L.,** 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1-2), pp.1-135.
- [57] **Altman, N.S.,** 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), pp.175-185.
- [58] **Jaskowiak, P.A. and Campello, R.J.G.B.,** 2011, August. Comparing correlation coefficients as dissimilarity measures for cancer classification in gene expression data. In *Proceedings of the Brazilian Symposium on Bioinformatics* (pp. 1-8). Brasília.
- [59] **Awrahman, B. and Alatas, B.,** 2017. Sentiment Analysis and Opinion Mining within Social Networks using Konstanz Information Miner. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 9(1), pp.15-22.
- [60] **Jaskowiak, P.A. and Campello, R.J.G.B.,** 2011, August. Comparing correlation coefficients as dissimilarity measures for cancer classification in gene expression data. In *Proceedings of the Brazilian Symposium on Bioinformatics* (pp. 1-8). Brasília.

BANAN JAMIL AWRAHMAN

MARZABOTO STREET, HALABJA, SULAIMANYAH, KURDISTAN, IRAQ

EMAIL: jameel.banan101@gmail.com

DOB: 1/JAN Y/1990.

EDUCATION

- **B.SC.**, (COMPUTER SCIENCE) 2007-2011: **SULAIMANIA UNIVERSITY**, KURDISTAN, IRAQ
- **HIGHER SCHOOL**, 2006-2007: 89% FROM **HALABJAH HIGH SCHOOL FOR GIRLS**, KURDISTAN, IRAQ

EMPLOYMENT AND EXPERIENCE

- DECEMBER 2012 – PRESENT, *ASSISTANT PROGRAMMER*, TECHNICAL COLLEGE OF APPLIED SCIENCES IN HALABJA / SULAIMANI POLYTECHNIC UNIVERSITY, KURDISTAN, IRAQ

COMPUTING SKILLS

- PROGRAMMING SKILLS: C++, C# .NET, JAVA.
- OPERATING SYSTEMS: WINDOWS, MAC, LINUX
- APPLICATIONS: PHOTOSHOP, MS OFFICE
- DATABASE: SQL SERVER 2008
- WEB TECHNOLOGIES: HTML, PHP

PROJECTS

B.SC., PROJECT TITLE – HAND RECOGNITION USING JMF.

PERSONAL DETAILS

NATIONALITY: IRAQI

KNOWN LANGUAGES: ENGLISH, ARABIC, KURDISH, TURKISH, AND PERSIAN.