Journal of Pharmacognosy and Phytochemistry

Available online at www.phytojournal.com

E-ISSN: 2278-4136
P-ISSN: 2349-8234
JPP 2018; SP1: 695-700

Hein Htet
University of Technology (Yatanarpon Cyber City), Myanmar

Yi Yi Myint
University of Technology (Yatanarpon Cyber City), Myanmar

Correspondence
Hein Htet
University of Technology (Yatanarpon Cyber City), Myanmar

Social media (Twitter) Data analysis using maximum entropy classifier on big data processing framework (Case study: Analysis of health condition, education status, states of business)

Hein Htet and Yi Yi Myint

Abstract
Most of the people aren't aware about their health situation, and they don't interest which level has been stand by their nation in case of business and health states. These factors are considered as important things to be improved each nation. Therefore, it is needed to focus these things not only citizens but also the authorities of each country. But, it can be difficult to focus these stated factors without using the modern computer technology. Nowadays, most of the people friendly used social media and people have started expressing their feelings and activities on it. And so, social media is a valuable source to analyze these things by using data mining techniques. Therefore, social media (Twitter) data analysis system is developed to know about health condition, education status, and states of business which are good, fair, or bad based on the data that they post on the Twitter. Maximum Entropy classifier is used to perform sentiment analysis on their tweets to suggest these stated conditions. It is interacting with Twitter data (big data environment), and so, big data processing framework is built to efficiently handle large amount of Twitter data.

Keywords: sentiment analysis, big data, framework, maxent, twitter

1. Related Work
Sentiment analysis research on the Twitter data are doing by most of the researchers in the world. Syed Akib Anwar [1] proposed that Public sentiments are the main things to be noticed for collecting the feedback of the product. The twitter is the social media used in this paper for collecting the reviews about any product. The reviews collected are analyzed based on the locations, features and gender. Then, data extraction, data processing is performed and the product analysis using sentiment score is performed.
The nextresearcher**,** Aarathi Patil [2] proposes that the sentiment analysis can be done on any product or event by using the social media based on the location. There are four major steps involved in this paper. First step is to create Twitter application which is used to mine the twitter4j and analyze the data. Next, the tweets are collected by using the secret tokens from the twitter. These collected tweets are saved in an excel file. Preprocessing of data has been carried out and then, the filtered out tweets are classified by using the Naive Bayes classifier. The sentiment scores are provided as 1 for negative sentiment, 2 for neutral sentiment and 3 for positive sentiments.
Researchers, Gargi Mishra and Shivani Varshney [3] propose an efficient methodology to determine the people opinions from real time twitter data. Two phases are mainly involved in this paper: In the first phase, the web crawler is used for extracting the real time data by ASP.net. The creped knowledge from the twitter is stored in the database. The second phase involves the process of analysis of collected tweets that are stored at the database. After preprocessing process, the classifier (Support Vector Machine) is used in order to classify positive, negative and neutral sentiments.

2. Introduction
There are 195 countries in the world today and each particular country is existing at different positions according with the health condition, Education status, Business state, and Crime rate. These stated factors may also be influenced by their geographic location. According to their regional, rice is a staple food for most of the Asian countries and most Western countries eat wheat as their main food from the view of health aspect on food. Moreover, their affirmation and belief on health may also be different regionally. Furthermore,

The other two factors, Education and Business levels, are also important things to be focused mainly for improving a country. People needed to be aware about their countries education and business situation and always needed to be checked which level has been positioned in accordance with the above stated things. This system is intended as a monitoring system of the conditions about health, education, and business for each individual social media (Twitter) users. Besides, it can also be surveillance this stated conditions relate with the continents such as Asia, Europe, Africa, etc. To commerce with, this intended system is based on the emotional data of each user that they describe on the social media. On the contrary, it is not focused on the physical data such as population, life expectancy, currency, educational habitat, etc. Not many people enjoy talking about health and fitness, especially when it concerns with their health problems. In addition, it may be difficult to mine about each country business and educational status in general based on insufficient irrelevant data. Nowadays, people use Twitter and have stated sharing in the public domain about their feelings, and activities. This system is focused only on Twitter users' created tweets which are composed news, politics, and life conversation for predicting each Twitter user and alternatively, each continental area' health, education, and business condition with the aid of data mining technologies. Additionally, this system is also intended for the Authorities of each country government to check about their nation' business, education, and health states. This surveillance system is developed to check about the health condition, states of business, educational status which is good, fair, or bad based on the data that they post on the Twitter by doing text analysis. This paper purposes are fetching tweets by applying the Twitter API, preprocessed these data on the cloud server and crawled to the big data processing framework to store for further usages. Then, it is analyzed by the MaxEnt classifier and finally produced about the above mentioned condition (positive, negative, and neutral) for the specific users and continental areas.

This paper is composed with eighth sections. Section two and threeare about the introduction and overall system architecture. Data collection, Data store and processing, data analysis are in the section four, five, and six consequently. Experimental results explanation is in the section seven and the last eighth section is about the conclusion and future work.

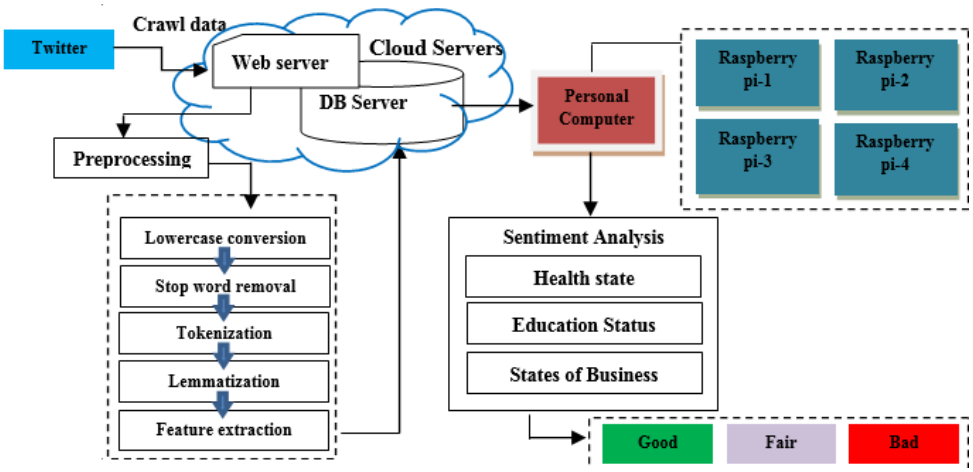

Fig 1: Overall System Architecture

3. Overall system architecture

Overall system architecture described in figure-1, it is firstly crawled data from the Twitter via Twitter API to the cloud web server. These raw tweets-data need to be cleaned and so,it is going to the preprocessing stage. Preprocessing stage is very important stage which transforms the raw data into the valuable data for doing analytical process. Firstly, it is needed to convert the raw tweets text into the same format (lowercase). After that, it is needed to remove stop words such as a, an, the, is, etc. Then, tokenization which breaks tweets set intoeach words is performed and it is needed to pass to the lemmatization stage that is grouping words based on the several grouping format in this system which are nouns, verbs, adjectives and adverbs.

After that, feature sets correspondence with the health, education, and business data are added to the feature vector and this resultant feature vector will be stored back to the cloud database server. In the local side, big data distributed processing framework is built using one PC for master node and four Raspberry pi boards for data nodes. Big dataprocessing ecosystem such as Hadoop, MapReduce, and Hive data warehouse is configured within this framework.

Sentiment analysis for stating the behavior of Twitter users or continental health, business and education, is performed at local side framework and the predicted outcome (good, fair, or bad) percentage results are shown to the tested Twitter' users.

4. Data collection

This section is also the first stage of doing data analytics and involves collection of data from several types of data sources, data marts and data warehouses. In this system, data are gathering from the social media- Twitter. Open API is used to fetch data from the Twitter. To get data from the Twitter, it is needed to create Twitter developer account. After getting, it is also needed to develop a new own application on the Twitter and getting API key and API secret key. Then, it creates the access token and will get access token and access token secret. After fulfill these requirements, it can be obtained Twitter data with the help REST API and Twitter4J. It is returned the data file to the developers by the csv or json format. These are some sample training data concerned with health, business, and education.

Positive tweets

1. "Health risks of light drinking in pregnancy confirms that abstention is the safest approach\u2026 https://t.co/YcfHfoZfcB UK"

2. "Experts urge action on poisons in Asia https://t.co/LyhA4Msv5j"

Neutral tweets

1. "RT@Wadsamnews:https://t.co/g7A42bKQeG\nWomen's Contributions to Agriculture Economy in Afghanistan Need Recognition #AfghanWomen #Agriculture"
2. "Discover #corruption risks for companies investing in #Kosovo https://t.co/sLjEQg8gQeby @gan_integrity"

Negative Tweets

1. "Example of how bad the American education system is: I used to think Egypt was near Iranand only recently learned it was in northern Africa"
2. "@Nervana_1 What do U mean, Egypt Well Developed, Cultured & Education in Egypt Started Long time, Even they Early Teachers to Gulf Nations..?"

5. Data store and processing

In Data Store and Processing stage, data is storing into apache HBase which is an open source, distributed, consistent, non-relational database, which provides low-latency, random read/write operations on top of Hadoop Distributed File System (HDFS).At shown in figure-2, the resultant feature words are imported to the HDFS with the aid of FLUME which is used for the purpose of import and export unstructured data. The imported data from the HDFS must be send to the map reduce paradigm which is also a primary usage tools in solving big data problems.

Twitter API can response up to 3600 tweets per user and maximum 200 tweets once at a time. A tweet will have maximum 140 characters and so, there will be the data size 3.6 Mb for each Twitter user to predict their health state. The more the tested users, the more storage space is required to persistence their data. Therefore, big data processing framework, Hadoop ecosystem is applied.

5.1 Hadoop

Hadoop is an open source platform based on the Java, is used for handling large amount of users' testing data that are used to predict their health state and two primary components of Hadoop mainly used in this proposed system are:

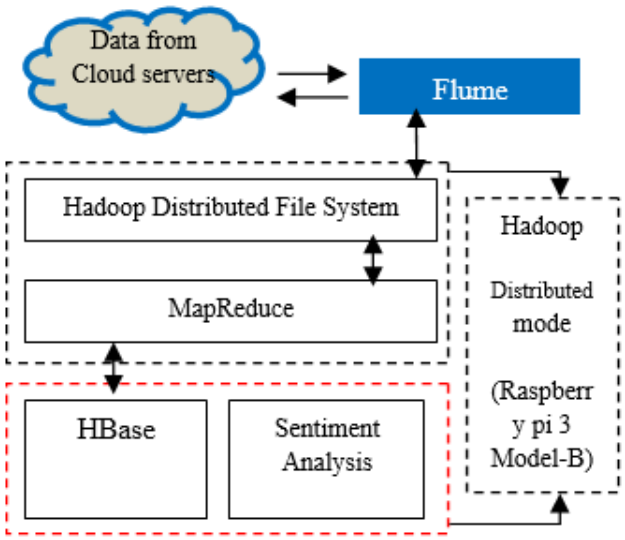

Fig 2: Workflow between HDFS, MR, Hive

5.1.1 Hadoop Distributed File System

The Hadoop Distributed File System is a versatile, resilient, clustered approach to managing files in a big data environment. It is a data service that offers a unique set of capabilities needed when data volumes and velocity are high. There are two types of service in HDFS- NameNode and Data nodes.

5.1.2 Hadoop MapReduce

MapReduce (MR) is a data processing paradigm that takes a specification of how the data will be input and output from its two stages- map and reduce. This process starts with a user request to run a MapReduce program and continues until the results are written back to the HDFS. MR tasks will be performed on the testing data to store efficiently for further processing.

The first step of Map phase is to locate and read the input file containing the raw data. This is the function of Input Format and Record Reader. Input Format decides how the file is going to be broken into smaller pieces for processing using a function called Input Split. The, it is assigned to a Record Reader (RR) to transform the raw data for processing by the map. RR is the class which actually loads the data from the source [4]. It is the class which converts the data into <Key, Value> pairs as sample examples, [<health, 2>, <medical, 1>], [<risk, 3>, <medical, 2>], [<health, 1>, <defense, 1>]. The Mapper will receive one <Key, Value> pair at a time until out split is consumed.

The Reduce phrase is performed by gathering intermediate results from the output of the Map phrase and then, by shuffling, sorting, and combining to get the desired result. The output of reduce task is also a key and a value such as [<health, 3>], [<medical, 3>], [<risk, 3>], and [defense, 1]. Then, Hadoop provides an Output Format feature which takes key-value pair and organizes the output for writing to HDFS. Finally, Record Writer is used to write the data to the HDFS.

5.2 Raspberry pi Hadoop Cluster

Hadoop has developed into a key enabling technology for all kinds of Big Data analytics scenarios. To inspire the prototype of the Hadoop, it can be installed open source Apache Hadoop from scratch on Raspberry pi 3 Model B. Hadoop is designed for operation on commodity hardware so it will do just fine on Raspberry Pi as a Master and Slaves type. At shown in figure-3, IoT based big data processing framework is built by using the four Raspberries pi, four cat-5 network cables, and five-port switch. Other additional devices which are on/ off switch adding for each Raspberry and fixing a cooling fan by upside down with the help of motor driver.

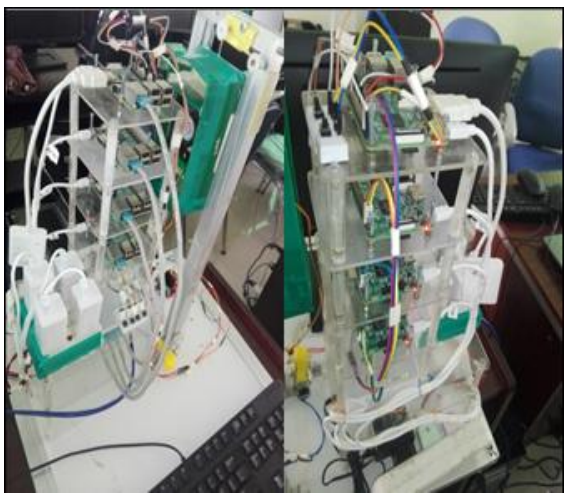

Fig 3: Hadoop Cluster with Raspberry pi

5.3 HBase

HBase is highly configurable, providing a great deal of flexibility to address huge amounts of data efficiently. HBase is a columnar database and it is modeled after Google Big Table and capable of hosting very large tables (billions of columns/rows) because it is layered on Hadoop clusters of commodity hardware. Some benefits of columnar database are that

- Data can be highly compressed and the compression permits columnar operations to be performed very rapidly.
- Column based DBMS is self-indexing, it uses less disk space than a relational database management system (RDBMS) containing the same data.
- Column architecture doesn't read unnecessary columns.
- Avoid decompression costs and perform operations faster.

6. Data analysis

Data analytics refers to methods and tools for analyzing large sets of data from diverse sources aiming to support and improve decision making. A number of organizations used sentiment analysis in order to collect feedback from the user. In this proposed system, it analyzes the data to focus the health, education, and business state of social media users. It can also be analyzed these factors according to each continent. As shown in figure-4, the sentiment analysis steps consist of two phases: Training phase and testing phase. In both phases, raw input texts need to be preprocessed and extract features which are described details in the next sections.

6.1 Preprocessing text data

The pre-processing is necessary because there are some words or expressions in the review that don't return any meaning and by the presence of those words that cannot get the correct sentiment analysis. So by doing pre-processing it can also get higher accurate results.

In pre-processing, the following steps contained: Lowercase conversion, stop word Removal, Text Segmentation and Normalization. The following are the resultant data after doing these steps.

Positive Tweets

1. "health risks of light drinking in pregnancy confirms that abstention is the safest approach\u2026 URL"
2. "experts urge action on poisons in Asia URL"

Neutral Tweets

1. "AT_USER URL women's contributions to agriculture economy in afghanistan need recognition rfghan women agriculture"
2. "discover corruption risks for companies investing in kosovo URL by AT_USER"

Negative Tweets

1. "example of how bad the american education system is I used to think egypt was near Iranand only recently learned it was in northern africa"
2. "AT_USER what do u mean, egypt Well developed, cultured & education in egypt Started Long time, even they early Teachers to gulf nations"

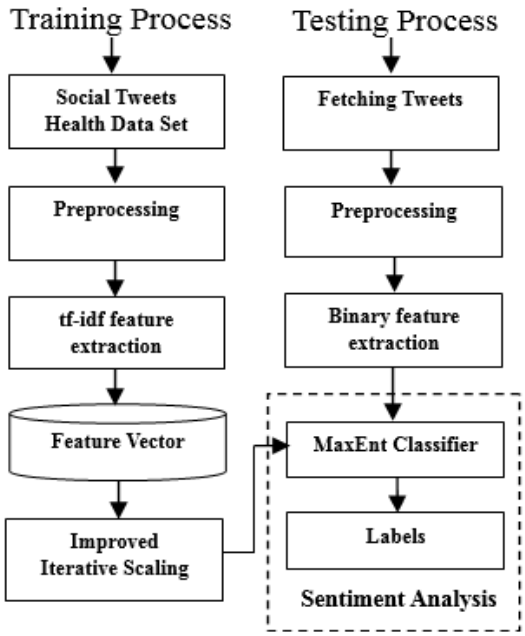

Fig 4: Schematic block for Sentiment Analyzers

6.2 Feature Extraction: tf-idf

The system is trained by the Maximum Entropy Classifier on TF-IDF [5] weighted word frequency features.It is also perform like that frequently occurring words present in all files of corpus irrespective of the sentiment, like in this case, 'health', 'fever', 'wellness', 'happy', 'sick', etc. will be selected as features words.

Term Frequency increases the weight of the terms (words) that occur more frequently in the document.

$$tf(t,d) = \log\big(F(t,d)\big) \quad (1)$$

where $F(t,d)$= number of occurrences of term 't' in document 'd'

Inverse Document Frequency diminishes the weight of the terms that occur in all the documents of corpus and similarly increases the weight of the terms that occur in rare documents across the corpus.

$$idf(t,D) = \log\left(\frac{N}{N_{t\in d}}\right) \quad (2)$$

6.3 Maximum entropy classifier

Supervised machine learning algorithm (MaxEnt) classifier is effectively used in a number of natural language processing applications. Sometimes, it outperforms Naive Bayes at standard text classification. Its estimate Of P(c | d) takes the exponential form as in Eq. (3),

$$P_{ME}(c/d) = \frac{1}{z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right) \quad (3)$$

Where, Z (d) is a normalization function. Fi, c is a feature/class function for feature fi and class c, as in Eq. (4),

$$F_{i,c}(d,c') = 1, if\ (n_i(d) > 0\ and\ c' = c)$$
$$0, \text{otherwise} \quad (4)$$

This classifier works by finding a probability distribution that maximizes the likelihood of testable data. The followings are the basic steps on doing Maximum Entropy Classifier:

- Collect a large number of training data which consists of samples represented on the following format: (x_i,y_i).
- Summarize the training sample in terms of its empirical probability distribution.
- For each word w and class c, define a joint feature f(w,c) = N where N is the number of times that 'w' occurs in a document in class c.
- Iterated to get optimization result, assign a weight to each joint feature so as to maximize the log-likelihood of the training data [6].

6.4 Classifier trainer algorithms

This paper is introduced and compared with two classifier trainer algorithms.

1. Improved iterated scaling (IIS)

Firstly, IIS is used for calculating the parameters of a maximum entropy classifier given a set of constraints. IIS [7] performs by collecting labeled documents 'D' which is also training data and a set of features function fi. For every feature fi, estimate its expected value on the training documents. Then, initialize all the λ's to be zero and iterate the convergence. At each step, IIS must find an incrementally more likely set of parameters. If it is guaranteed that IIS succeeds in improving the likelihood, then it is known that it will converge to the globally optimal set of parameters. That is both the maximum likelihood solution for the parametric form and the solution with the maximal entropy. The log likelihood of an exponential model can be calculated by using equation 5.

$$l(\Lambda/D) = log \prod_{d\epsilon D} P_\Lambda(c(d)/d) = \sum_{d\epsilon D}\sum_i \lambda_i f_i(d, c(d)) - \sum_{d\epsilon D} log \sum_c exp \sum_i \lambda_i f_i(d, c) \quad (5)$$

2. Generalized iterative scaling (GIS)

GIS is a method that searches the exponential family of a Maximum Entropy solution of the form:

$$P^{(0)}(x) = \prod_i \mu^{(0)f_i(x)} \quad (6)$$

The next iteration of each is intended to create an estimate that will match the constraints better than the last one. Each iteration 'j' follows the steps:

1. Compute the expectations of all the f_i's under the current estimate function, i.e.,

$$\sum_x P^{(j)}(x) f_i(x) \quad (7)$$

2. Compare the present values with the desired ones, updating the

$$\mu_i^{(j+1)} = \mu_i^{(j)} \cdot \frac{K_i}{E_{p^{(j)}} f_i} \quad (8)$$

3. Set the new estimate function:

$$P^{(j+1)} = \prod_i \mu_i^{(j+1)f_i(x)} \quad (9)$$

4. If convergence or near-convergence is reached stop; otherwise go back to step1.

7. Experimental results

All experimental results of proposed research are showing that it is on the 8768 training data sets and testing on the 1645 data. By applying the Maximum Entropy Classifier with the use of Improved Iterated Scaling (IIS) algorithm, this system is iterated with the number of 100 iterations.

The bar chart shown in figure 5 is the practical analysis result of the health state for Asia. The system is predicted that positive (34.34%), negative (22.18%), and neutral (43.46%) based on the testing datasets 1645 tweets. And so, it is generally concluded that Asia health state is fair situation. Furthermore, figure 6 is about the analysis result of Education rates on Asia and the system return 38% (positive), 32%(neutral), 30%(negative) respectively. Besides, figure 7 is about the analysis result of Business rates on Africa and the system return 23% (positive), 20% (neutral), and 30% (negative). Similarly, figure 8 is about the health analysis result of president Mr. Barack Obama based on the tweets that he posted on Twitter and the system return 78.6% (positive), 25.3% (neutral), and 6.1% (negative). So, he can be suggested as a healthy man by the system.

In table1, the evaluation results are discussed based on the experimental results between two classifier trainer algorithms, Improved Iterative Scaling (IIS) and Generalized Iterated Scaling (GIS).

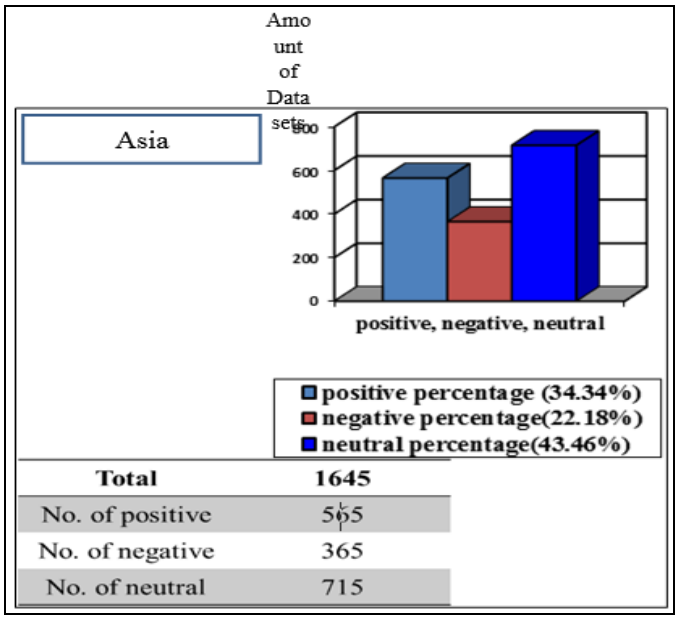

Fig 5: Asia' Health Analysis Result

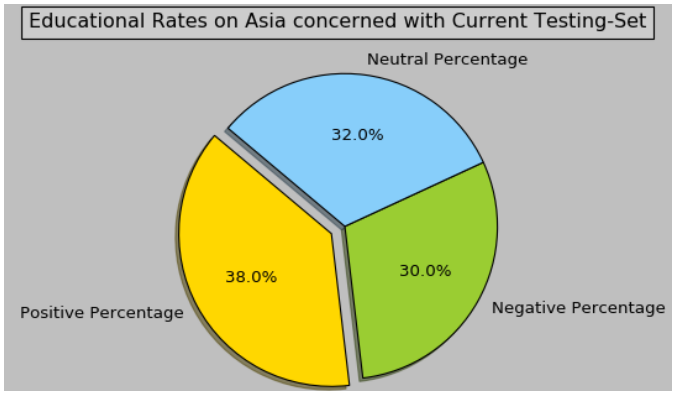

Fig 6: Asia' education analysis result

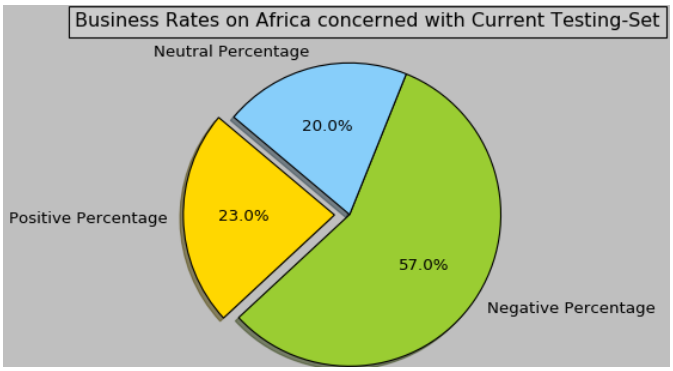

Fig 7: Africa' business analysis result

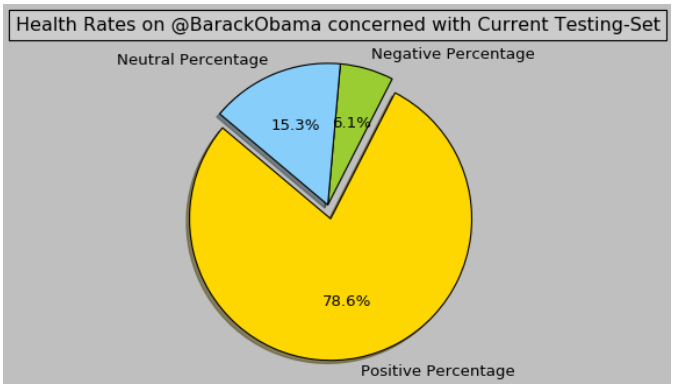

Fig 8: Obama' health analysis result

For performance comparison, the following two factors are considered:

- **Precision** = fraction of the returned result that are relevant to the information need
- **Recall**= fraction of the relevant documents in the collection that were returned by the system

Table 1: Evaluation Results Comparison

Process Status	Health Data Sets (8768)	
	IIS	GIS
Training Time	3-hr & 32 min	3hr&14min
Testing Time	1.55-min	1.05-min
Positive Precision	0.944250	0.899910
Positive Recall	0.957763	0.905531
Neutral Precision	0.901954	0.912234
Neutral Recall	0.912263	0.779985
Negative Precision	0.881472	0.801172
Negative Recall	0.857440	0.784542
Average Precision	0.909225	0.871105
Average Recall	0.909155	0.823352
Average Accuracy	90.91%	84.72%

8. Conclusion and future work

There are some kinds of work for further extending to get more accurately results relate with health. The first thing is that Hadoop MR used in this proposed system has some drawbacks in the case of input/ output transaction. And so, it can be replaced by Hadoop YARN which is more convenient in I/O transaction. Then, Raspberry pi boards are used for building big data framework. This is feasible in it because it can only be used in resizable case of testing Twitter data and the resultant resizable data is stored back to the cloud database server. Therefore, Raspberries should be replaced more powerful desktop computers in the case of building distributed cluster for real-time big data framework. Moreover, it is needed to be added meta-data analysis in the case of sentiment health state analysis.

This monitoring and surveillance system is developed by doing social media sentiment analysis to be useful for the people who are very poor to understand about their realistic health state and to be aware for each government about their regions business and education situation. Therefore, it can be expected to build the sentiment analyzer for solving this. So, this proposed system consists of two main parts:

- Building social media mining sentiment analyzer
- Examining big data solutions on this analytics case

This system will be developed by using Python programming language, Open API, open source big data platforms and cloud computing technology.

References

1. Syed Akib Anwar. Localized Twitter Opinion Mining Using Sentiment Analysis, India, 2015.
2. Aarathi Patil. Location Based Sentiment Analysis of Products or Events over Social Media, India, 2014.
3. Gargi Mishra, Shivani Varshney. Location Based Opinion Mining of Real Time Twitter Data, New Delhi, India, 2016.
4. Park Sun Pyo. Introduction of Big Data and Hadoop Ecosystem, Head of Korea-Myanmar E-learning Center, Korea, 2016.
5. Ganesh Shinde, Sachin Deshmukh N. Sentiment TFIDF Feature Selection Approach for Sentiment Analysis, International Journal of Innovative Research in Computer and Communication Engineering, 2016.
6. Nipun Mehra, Shashikant Khandelwal, Priyank Patel. Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews. International Journal of Computer Applications (0975-8887), 2016.
7. Kamal Nigam, John Lafferty. Using Maximum Entropy for Text Classification, Andrew McCallum, Carnegie Mellon University Pittsburgh, 2013.