Big Data Analysis Using Naive Bays and Neural Network

Sozan A. Mahmood[1], MzhdaHewaHama[2]

Computer Department, Sulamani University,
KRG-IRAQ.

[1] susanara80@yahoo.com, [2] mzhda.hiwa@gmail.com

ABSTRACT

In the current years, volume of data increased dramatically. The structured, semi-structured and unstructured data come from different sources like blogs, e-mails, social networks, wikis and tweets. Big data analytics refers to capability of extracting useful information from these types of data that could not be managed by current methodologies and relational database management systems or data mining software tools. In this work sentiment analysis as one of the user cases of the big data analytics has been investigated. Sentiment analysis is a task of determining polarity in opinions, feelings, and attitudes expressed in text sources. Two different Bags of words have been made from training set by applying Multinomial Naïve Bayes algorithm and Forward Neural Network. These two bags and an existing Bag are applied to the test set, Naive Bayes has the highest F-Score.

Keywords: Big Data, Data Analysis, naïve bays, Neural Network

INTRODUCTION

We are living in the era of information explosion which large scale amount of data is getting increasingly larger because of virtual worlds, wikis, blogs, e-mails, online games, VoIP telephone, digital photos, tweets, traffic systems, bridges, airplanes and engines, satellites, RFID and weather sensors. [1]

Río et al., 2014 [2] analyzed the performance of several techniques used to deal with imbalanced datasets in the big data scenario using the Random Forest classifier. Miller et al., 2014 [3] propose treating spam detection on Twitter as an anomaly detection problem. Md. Zahidul Islam, 2013[4], was used big data within the cloud, in this study an application has been built for collecting, organizing, analyzing and visualizing of the data from retail industry, which has been gathered from indoor navigation systems and social networks like Twitter and Facebook. Cheng et al.2013 [5] suggest the potential of swarm intelligence in big data analytics, they concluded that with the application of the swarm intelligence and combining with data mining techniques more rapid and effective methods can be designed to solve big data analytics problems. T. Chardonnens, 2013 [6], is provided solution for high velocity stream, Twitter and Bitly are used as a datasets. Storm is mainly used for stream processing system. Use cases has been done in order to analyze the scalability of storm .The goal was to analyze the performance based on the number of worker nodes, using three nodes shown better performance than others. J. Mehine, 2011 [7]. This thesis has introduced how to process a large-scale data and analysis data using PigPig is one of subproject of Hadoop that is located on top of MapReduce ,and how it is related to the MapReduce programming model. Feng Fengand et al., 2015 [8], this research focuses on the visualization and quantitative study in bibliographic databases by taking the university–industry collaboration studies as an example.

Leena and Luna International, Oyama, Japan.
(株) リナアンドルナインターナショナル, 小山市、日本.
ISSN: 2186-8476, ISSN: 2186-8468 Print
www.ajsc. leena-luna.co.jp

The main aim of this work is to analyze Sentiments of twitter to determine people feeling about any kind of subjects, to get familiar with methodologies, tools, techniques in Big Data like Apache Hadoop and its characteristics and components like MapReduce paradigm, Apache Hive and HDFS. In this work MongoDB as NoSQL databases are used to store unstructured data like JSON files, and Using python NLTK library for preprocessing datasets. Also,applying machine learning algorithms like Naïve Bayes and Neural Network in sentiment analyses and making Bag of Words and Implementation of algorithms using Hadoop.

THEORETICAL BACKGROUND

Hadoop

Hadoop is a Top level Apache project, open source software framework that's written in java programming language which allows the distributed processing of massive data sets across different sets of servers. Hadoop is designed to scale up from single server to a thousands of machines, each offering local computation and storage [9].

The most important software for big data analytic is Apache Hadoop which is based on MapReduce programming paradigm and a distributed file system which called HDFS. It allows writing programs using Apache Hive that quickly process large amount of data in parallel on large clusters of computer nodes. Apache Hive codes compile and translate to MapReduce jobs. They divide the input dataset into independent subsets that are processed by map tasks in parallel. After that steps of reducing tasks to obtain the results are followed. Hadoop consists of two main components:

HDFS

Is a Java-based file system that provides scalable and reliable data storage that is designed to span large clusters of commodity servers. HDFS was designed to be a scalable, fault-tolerant, distributed storage system that works closely with MapReduce [7]. HDFS is highly fault-tolerant and have designed to be deployed on low-cost hardware [10].

MapReduce

MapReduce is a Heart of Hadoop, a programming model for parallel processing of tasks on a distributed computing system and an associated implementation for processing and generating large data sets. This algorithm allows splitting of a single computation task to multiple nodes or computers for distributed processing. As a single task can be broken down into multiple subparts, each handled by a separate node, the number of nodes determines the processing power of the system. As MapReduce is an algorithm, it can be written in any programming language.[11]

A MapReduce program consists of two user-specified functions, the Map phase and the Reduce phase as shown in figure (1). In the Map phase, each mapper reads raw input, record by record, and converts it into Key/Value pair [(k,v)], and feeds it to the map function then map function performs a computation on a key value pair. The Map function operates on each of the pairs in the input and produces intermediate output in the form of new key/value pairs depending upon how the user has defined the Map function. Output of the map function is then passed to the reduce function as input [12]. The reduce function then, applies an aggregate function on its input merges all intermediate values associated with intermediate keys (e.g. count or sum values), and stores its output to the disk.

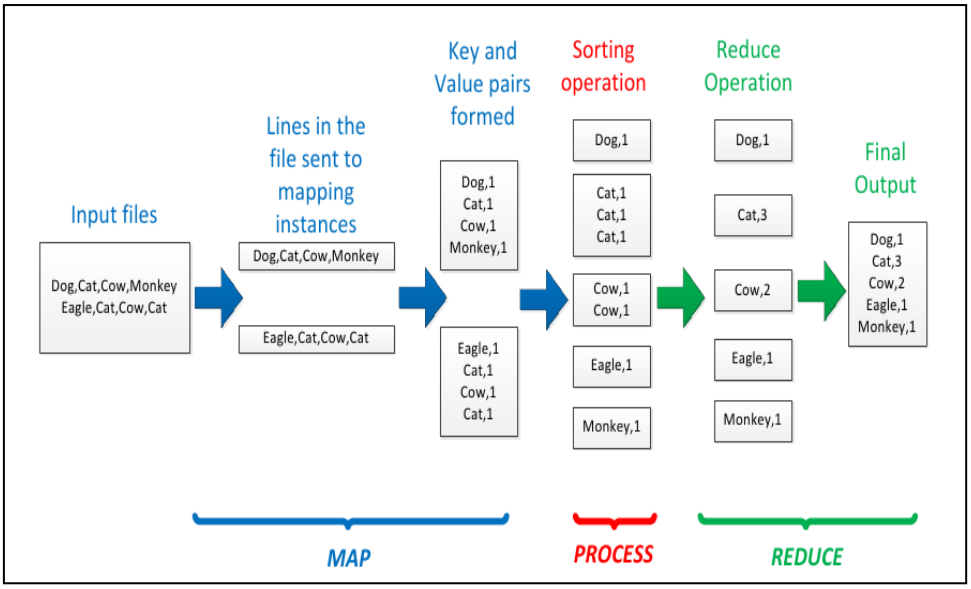

Figure 1. Illustration of MapReduce Algorithm [9]

The mapping function takes the input file, separates the records, and sends them to different nodes or mapping instances for processing. The mapping function then splits the document into words and each record is given to one cluster to perform mapping tasks in parallel.

Then assigns a digit "1" to inform a key-value pair for further computation. The intermediate output is in the form of (word, 1) and is sorted and grouped into individual nodes to calculate the frequency. The resultant output from the sort operation is then fed to the reduce function, which sums up the outputs from different nodes and generates the final output containing the frequency.

Hive

The Apache Hive is an open source data warehouse software built on top of Hadoop for providing data summarization, query, and analysis [13]. Queries are compiled into MapReduce jobs it means that queries are automatically translated in to MapReduce and then executed on Hadoop [14]. HCatalog Is a component of Hive built on top of hiveis a table and storage management layer for Hadoop that enables users with different data processing tools — Pig, MapReduce to more easily read and write data on the grid. HCatalog supports reading and writing files in any format for which a SerDe can be written. By default, HCatalog supports RCFile, CSV, JSON, and SequenceFile, and ORC file formats [15].

NOSQL Databases

Evolution of the SQL databases begins in the late 1990s. After a few years it became a serious competitor to RDBMS because RDBMS has its own set of problems when applied to massive amounts of data. The problems relate to efficient processing, effective parallelization, scalability, and costs. So in the year 2009 and 2010 there were organized NoSQL conferences.[16].

PROPOSED SYSTEM

To implement the proposed system, twitter sentimental data has been investigated, refined analyzed and visualized as one of the user cases of big data analytics. The algorithms have been implemented in Horton works Sandbox, which is a personal and portable Hadoop environment that has been installed on a VMware virtual machine environment. Apache Hive

Leena and Luna International, Oyama, Japan.
(株) リナアンドルナインターナショナル, 小山市、日本.
ISSN: 2186-8476, ISSN: 2186-8468 Print
www.ajsc. leena-luna.co.jp

has been used for implementing the algorithms and processing all the data inside the HDFS. Hive provides HiveQL is a Query language used for implementation of all algorithms then automatically converted into MapReduce and then executed on Hadoop. Figure (2) shows the steps of proposed system implementation.

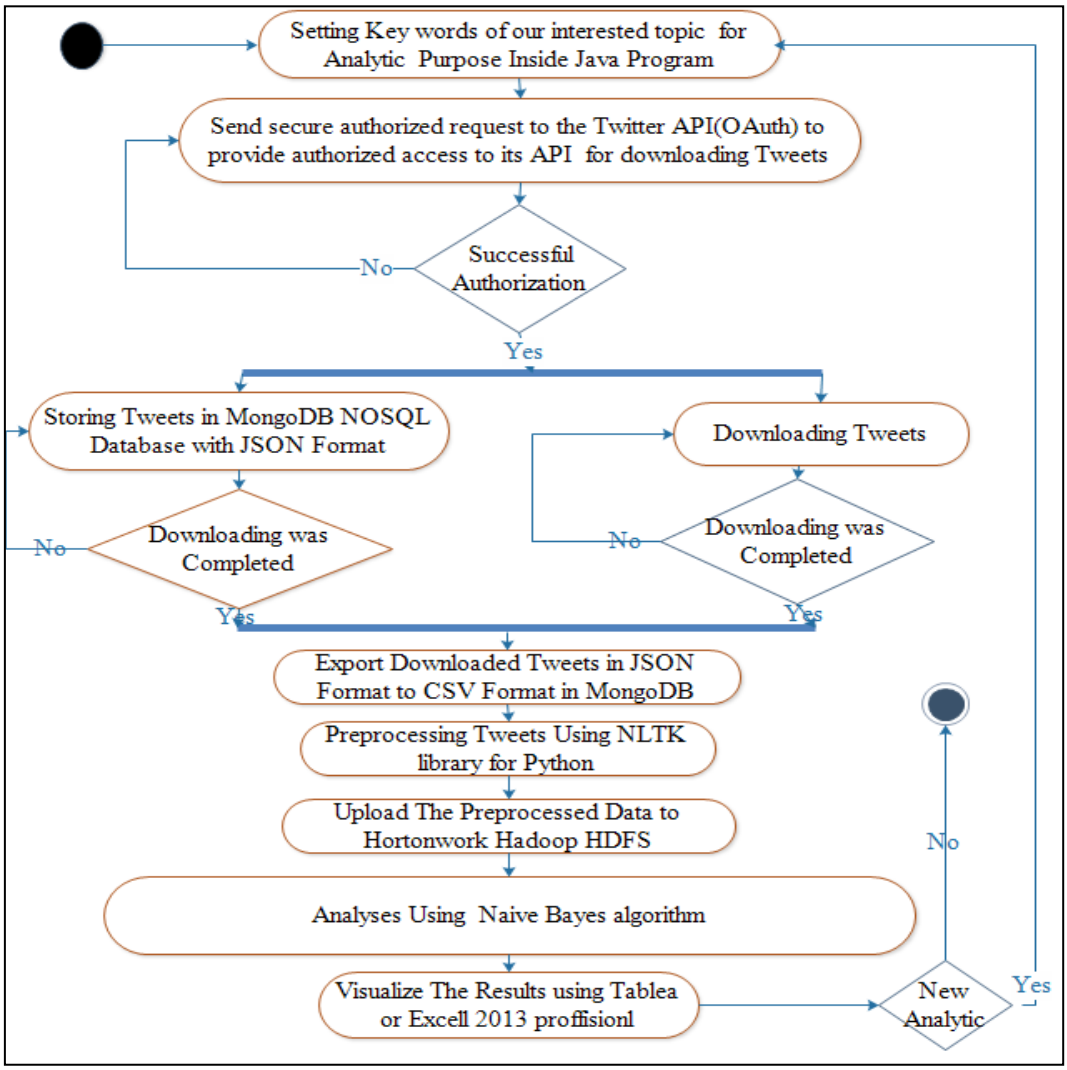

Figure 2. Overview of System Implementation

Twitter Sentiment Analysis Dataset contains 1,578,627 classified tweets, which have been sorted by tweets content alphabetically. Each row is marked as "1" for positive sentiment and "0" for negative sentiment.

The public tweets directly have been downloaded using a java program and Twitter API has been stored in a mongoDB. For preprocessing Stemming stage is used to find out the root/stem of a word. Lemmatization is very similar to stemming, but is more akin to synonym replacement. A lemma is a root word, as opposed to the root stem. As the first part of the work after preprocessing about 30% of the corpuses has been used for testing the algorithms, while the rest (70%) has been dedicated towards training set to train the neural network and Naïve Bayes algorithms to classify sentiments.

In both algorithms Bag of Words from training set has been made. Later Neural Network algorithm has been used to prepare dictionary of positive and negative words.

The simplest way to compute frequent item sets is to consider all possible item sets, compute their support, and check whether they are higher than the minimum support threshold.[17] The following steps for naïve Bayes's rule applied to documents and classes, for each document d and class c:

1. From training corpus, extract Vocabulary */*list of words*/*

   Calculate P(C_j) terms/ likelihood */*

   For each C_j in C do

   $docs_j$=all docs with class C_j

   $$P(C_j)=\frac{|docs_j|}{|total\ number\ of\ decuments|}$$

2. Calculate $P(W_k|C_j)$ terms

   $Text_j$=single doc containing all $docs_j$

   For each word W_k in bag of words

   n_k=number of occurences of W_k in $Text_j$

   $$P(W_k|C_j)=\frac{n_k+\alpha}{n+\alpha|size\ of\ bag\ of\ words|}$$

3. For each given docs in test set

   $$\Gamma(POSITIVES) = P(C_j)\prod_{I=1}^{K}\left(P(W_k|C_j)\right)$$

   $$\Gamma(NEGATIVES) = P(C_j)\prod_{I=1}^{K}\left(P(W_k|C_j)\right)$$

   IF $\Gamma(POSITIVES) >= \Gamma(NEGATIVE)$ *THEN THE SENTENCE IS POSITIVE*

   ELSE NEGATIVE

Activity diagram of making bag of words using Naïve Base has been illustrated in the figure (3), the operations that needs to be done are clarified using the diagram. There are three different bags of words in order of analyzing the sentiment data. The first bag of words has been made by Naïve Bayes algorithm and the second one has been made by Neural Network and the last one is available it is ready.

Leena and Luna International, Oyama, Japan.
(株) リナアンドルナインターナショナル, 小山市、日本.
ISSN: 2186-8476, ISSN: 2186-8468 Print
www.ajsc. leena-luna.co.jp

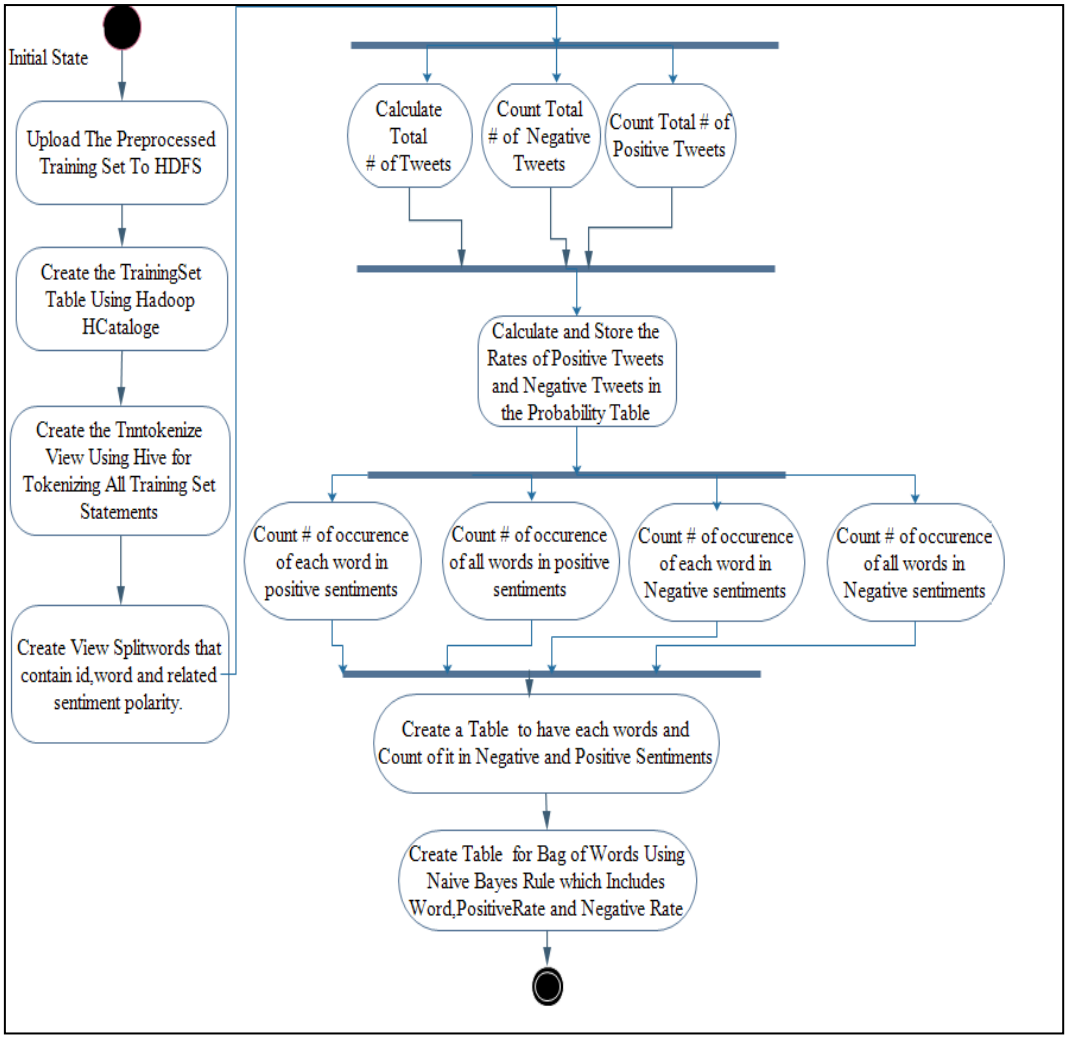

Figure 3. Activity Diagram to Make Naïve Bayes Bag of Words

After creation of bag of words the naïve Bayes algorithm need to be implemented on Hadoop Using Apache Hive as shown in figure (4).

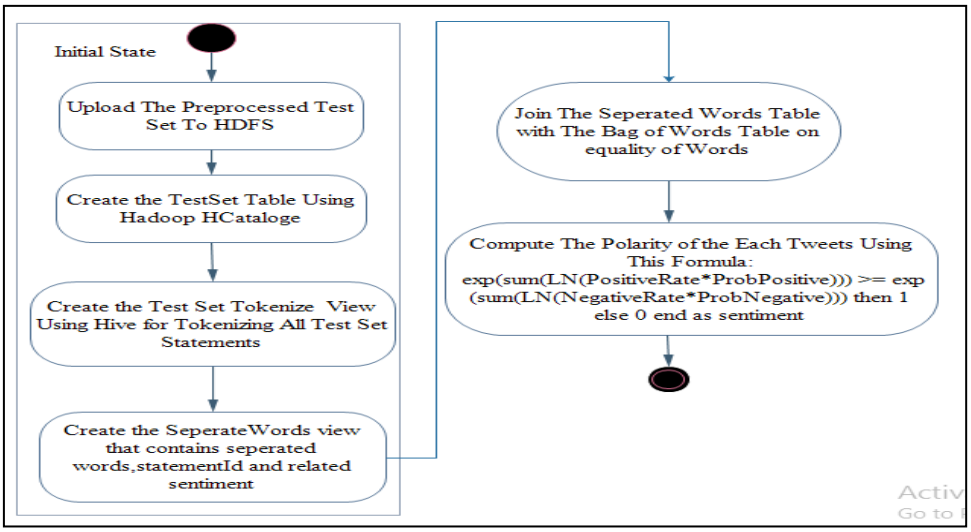

Figure 4. Naïve Base Algorithm Implementation on Hadoop Using Apache Hive

FEED FORWARD NEURAL NETWORK

Neural Network has emerged as an important tool for classification. The recent vast research activities in neural classification have established that Neural Networks are a promising alternative to various conventional. In this proposed system neural network text classifier contain 3 layers .The input layer contain n nodes which is the size of bag of words, hidden layer and one output node for showing 0 or 1 result. Figure (5) shows the activity diagram to make neural network bag of words.

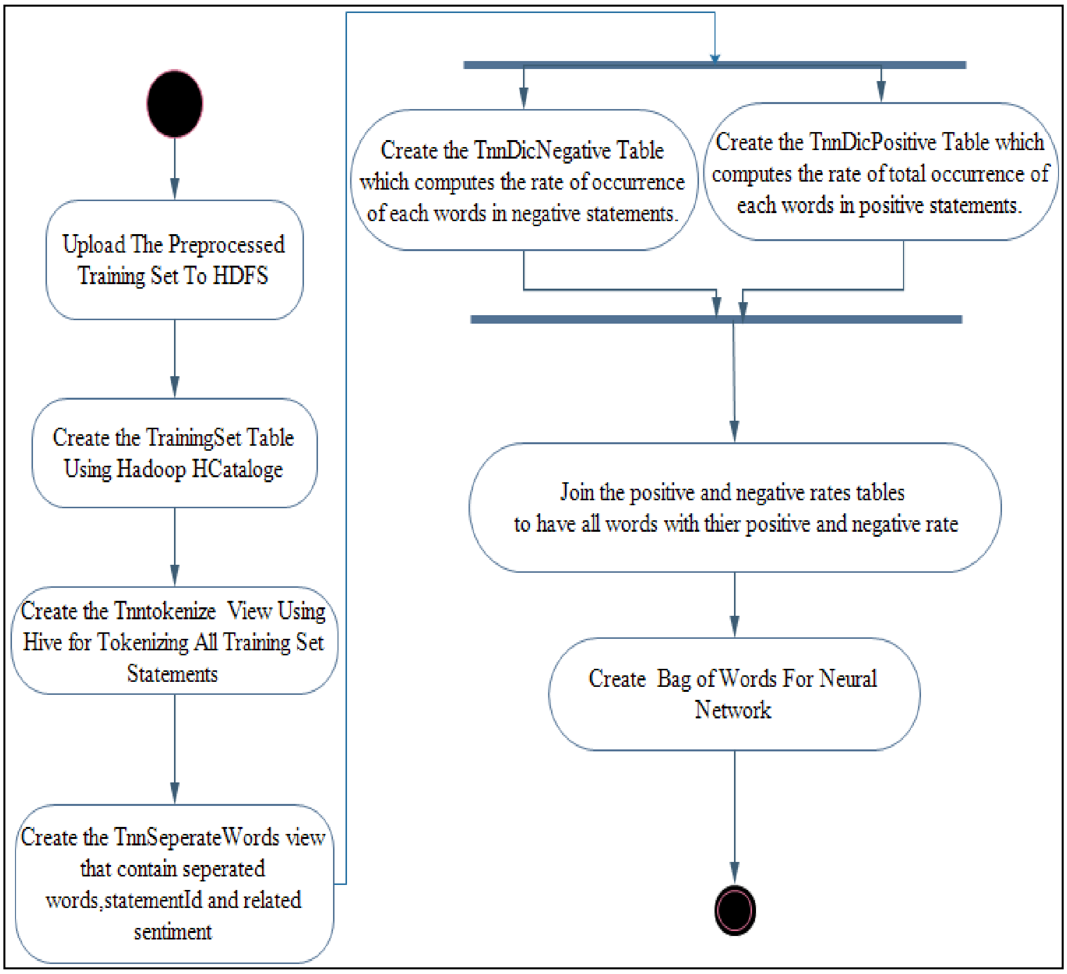

Figure 5. Activity Diagram to Make Neural Network Bag of Words

ACCURACY COMPARISON

Three different algorithms have been applied on the test set data. For each algorithm Precision, Recall and F1 has been computed and used for the accuracy comparison of those three algorithms. In order to calculate the Precision, Recall and F-Score the following equations need to be used:

$$\text{Precision} = \frac{True\ Positive}{Number\ of\ Predicted\ Positives} = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \ldots\ldots (1)$$

$$\text{Recall} = \frac{True\ Positive}{Number\ of\ Actual\ Positives} = \frac{True\ Positive}{True\ Positive + False\ Negative} \ldots\ldots\ldots\ldots (2)$$

$$\text{F-Score} = \frac{2 * Presision * Recall}{Presision + Recall} \ldots\ldots\ldots\ldots (3)$$

ISSN: 2186-8476, ISSN: 2186-8468 Print
www.ajsc. leena-luna.co.jp

Table 1. Accuracy comparison among Naïve Bayes, Neural Network and Existing BOW

Algorithm	*Precision*	*Recall*	*F Score*
Neural Network Algorithm	0.782398072	0.697214929	0.74
Naïve Bayes Algorithm	0.777090497	0.741529992	0.76
Existing Bag of Words Algorithm	0.549014525	0.904401909	0.68

As mentioned before the sentiment analyses can be used in many different areas. Organizations use sentiment analysis to understand how the public feels about something at a particular moment in time, and also to track how those opinions change over time. Table (2) shows the information about the categories that has been selected for the purpose of this work globally over 50,000 tweets has been downloaded, refined, processed and visualized. For each category about 10,000 tweets has been downloaded, preprocessed, refined, analyzed and visualized. All tweets has been selected in English languages and they are form different periods of time from 01/12/2014 to 15/01/2015.Figure (6) shows Results Chart of Sentiments of Iraq, and Figure (7) shows Visualization of Sentiment Data about Kurd around the World.

Table 2. Sentiment Analyses Result

Category	*Keywords*	*Total Tweets*	*Positive %*	*Negative%*
Competitors	{“IPhone 6”}	10341	19%	81%
Reputation	{“Obama”}	10628	72%	28%
Reputation	{“ISIS”}	10242	70%	30%
Social	{“Marriage”}	10389	69%	31%
Social	{"Kurd","Kurdistan","North of Iraq","Kurdish", "Peshmerga","Kobani","Shangal","Sinjar"};	10051	53%	47%

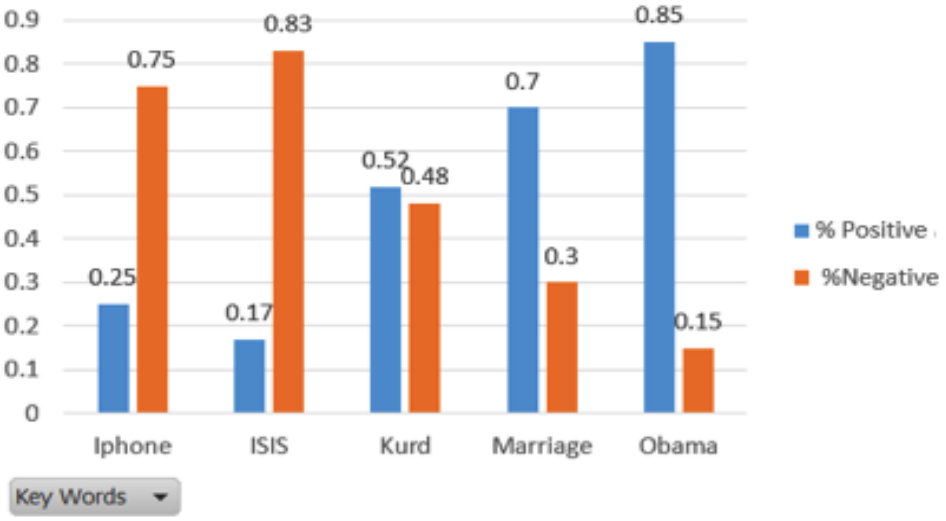

Figure 6. Results Chart of Sentiments of Iraq

Figure 7. Visualization of Sentiment Data about Kurd around the World

CONCLUSION

Naïve Bayes and Feed Forward Neural Network have been applied into an available data set to containing above 1,500,000 Tweets and for each algorithm a specific Bag of Words has been obtained. Then using the Bag of Words and applying both algorithms on test set data, interesting results has been achieved. Moreover a ready to use Bag of Words has been downloaded and in the same way with applying Neural Network algorithm some new results have been obtained that was not good enough as the previous Bag of Words that has been made by applying the algorithms on the data set. Actually, our made Bag of Words works much better than ready to use one. Using the applying the same algorithm on test set for proposed system Bag of Words showed 14% better accuracy. Multinomial Naïve Bayes and its based Bag of Words have also shown better accuracy performance than others. It showed almost 3% better accuracy than Neural Networks with its based Bag of Words and more than 17% than ready to use Bag of Words. Preprocessing using NLTK library for python on tweets, improved the accuracy for about 2%.Moreover it decreased the size of all files for about 25%. This reduction of the files size lead to perform and apply the algorithms in a much more efficient and convenient way.

Leena and Luna International, Oyama, Japan.
(株) リナアンドルナインターナショナル, 小山市、日本.
ISSN: 2186-8476, ISSN: 2186-8468 Print
www.ajsc. leena-luna.co.jp

ACKNOWLEDGMENT

The University of Michigan Sentiment Analysis competition on Kaggle and Twitter Sentiment Corpus by Niek Sanders are appreciated for providing data sets for the experiments.

REFERENCES

[1] Noseworthy, G. (2012). Infographic: Managing the Big Flood of Big Data in Digital Marketing. Available at: http://analyzingmedia.com/2012/infographic-big-flood-of-big-data-in-digital-marketing/

[2] Río, S.d., López, V., Benítez, J. M., & Francisco, H. (2014). "On the use of MapReduce for imbalanced big data using Random Forest". *Information Sciences, 285,* 112-137.

[3] Miller, Z., Dickinson, B., Deitrick, W., Hu, W., & Alex, H. W. (2014),"Twitter spammer detection using data stream clustering" *Information Sciences, 260,* 64-73.

[4] M. Islam (2013). "*A Cloud Based Platform for Big Data Science".* Master Thesis, Linköping University Department of Computer and Information Scienc.

[5] Cheng, S., Shi, Y., Qin, Q., & Ruibin, B. (2013). "*Swarm Intelligence in Big Data Analytics*". Berlin: Springer-Verlag.

[6] Chardonnens, T. (2013). "*Big Data analytics on high velocity streams".* Master Thesis, Software Engineering Group Department of Informatics University of Fribourg.

[7] Mehine, J. (2011). "*Large Scale Data Analysis Using Apache* Pig". Master's Thesis, University of Tartu.

[8] Fenga, F., Zhanga, L., Dub, Y., & Weiguang, W. (2015). "Visualization and quantitative study in bibliographic databases: A case in the field of university–industry cooperation". *Journal of Informetrics, 9*(1), 118–134.

[9] Chetan, S. (2014)." *Big Data Analytics using Neural Networks*". Master Thesis, San José State University SJSU ScholarWorks.

[10] Hadoop definition. (2014). Retrieved October 23, 2014, from http://hadoop.apache.org

[11] Benoy, B. (2014). "*Emergence and Taxonomy of Big Data as a Service*". Working Paper CISL.

[12] Raste, K. S. (2014). "*Big Data Analytics - Hadoop Performance Analysis"*. Master Thesis, San Diego State University.

[13] Borthakur, D. (2007). "The hadoop distributed file system: Architecture and design". *Hadoop Project Website, 11,* p.21.

[14] Wang, G. (2012). "*Evaluating MapReduc System Performance: A Simulation Approach*". Ph.D Thesis, Virginia Polytechnic Institute and State University ,Virginia, USA.

[15] Dionysios, L. (2011). "*Architectures for Stateful Data-intensive Analytics".* Ph.D. Thesis University of California, San Diego.

[16] Catalog, H. (2014). Available at: https://cwiki.apache.org/confluence/display/Hive/Home

[17] Cios, K. J., Pedrycz, W., Swiniarski, R. W, & Kurgan, L. (2007). *Data Mining, A Knowledge Discovery Approach*. Berlin: Springer-Verlag.

Leena and Luna International, Oyama, Japan.
(株) リナアンドルナインターナショナル, 小山市、日本.
ISSN: 2186-8476, ISSN: 2186-8468 Print
www.ajsc. leena-luna.co.jp