International Journal of Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-5, Issue-1)

Research Article | **January 2016**

A Novel Lexicon Based Approach to Enhance the Accuracy of Sentiment Analysis on Big Data

G. Vaitheeswaran*
Research Scholar
Department of Computer Science
St. Joseph's College (Autonomous)
Tiruchirappalli, Tamilnadu, India

L. Arockiam
Associate Professor
Department of Computer Science
St. Joseph's College (Autonomous)
Tiruchirappalli, Tamilnadu, India

Abstract—

The widespread and best use of World Wide Web generate an enormous amount of unstructured data called, 'Big Data'. Such knowledge-rich data plays a vital role in the decision-making process. In day-today life, social media websites are serving as a powerful platform for sharing opinions which contains rich source of information on different aspects of zillion users'. Information can also be extracted in the form of sentiment polarity from the massive amount of unstructured/structured data by the analytical process, known as Sentiment Analysis. Dealing with the big data in the process of sentiment analysis demands big data technologies such as Hadoop Distributed File System and Flume, under the umbrella of Hadoop ecosystem. Twitter is a microblogging social media website contains rich source of information to carry-out sentiment analysis. An usual tweet contains word variations, emoticons, hash tags, contextual texts, etc. After preprocessing, the lexicon-based algorithm Senti_lexi classifies the tweets as positive or negative based on the emoticon and the contextual sentiment orientation of the words. A new algorithm Senti_Lexi has been proposed to provide better accuracy. This paper has also presented a big data platform for processing sentiment analysis.

Keywords— Big Data, Sentiment Analysis, Lexicon Based Approach, Hadoop Ecosystem.

I. INTRODUCTION

The rapid growth of the Internet and online activities (like clickstreams, blogging and micro-blogging, social media communications, e-commerce, online transactions, ticket booking, surveillances, conferencing, chatting, etc.,) have enabled the research and industrial community to extract, transform, load, and analyze very huge amount of structured and unstructured data, referred to Big Data. Such data can be analyzed using a combination of Data Mining, Text Mining, Web Mining, and Natural Language Processing techniques. The enormous amount of data related to customer opinions/reviews is quite difficult to analyze and needs extant approaches to get a generalized opinion summary. Various web resources, news reports, e-commerce web sites, social networks (YouTube, Facebook, Twitter, Pinterest, etc.,), blogs and forums are serving as a platform to express opinions, which can be utilized for understanding the opinions of the general public and consumers on social events, political movements, company strategies, marketing campaigns, product preferences, and monitoring reputations [1]. To achieve these tasks, research communities, academicians and industrialists have been working thoroughly on sentiment analysis for the past three and half decades.

Sentiment analysis (SA) is a computational study of opinions, sentiments, emotions, and attitude expressed in texts towards an entity [2]. Online media and Social Networking Sites (SNS) are used to express and share public experiences in the form of product reviews, blogs, and discussion forums. Collectively, these media contain vastly unstructured data, the combination of data formats such as text, images, audios, videos and animations that are useful in making public awareness for various issues. Online media affords the platform for broader sharing of ideas and boosting public for group discussions with open views. It delivers a better means to get a quick response and feedback on different Global issues and entities in the form of textual posts, news, images, and videos. Thus, it can be utilized to analyze peoples' opinions about learning the behaviors of consumer, patterns market, and trends of society [3]. Twitter has 320 million monthly active users and it posts 500 million tweets every day[1] and Facebook has 936 million daily and 1,440 million monthly active users[2] as of December, 2015. Thus, it helps as a good resource to extract heterogeneous opinions posted by people from diverse societies for different purposes such as improvement of quality of products and services, prediction of consumers' demand and taste, etc. The sentiment found within critiques, feedback and comments, provide fruitful information for many different purposes.

Twitter is a microblogging website which allows users to tweet not more than 140 characters. This short message contains rich source for processing sentiment analysis. An usual tweet holds images, audios, videos, url, word variations,

[1] https://about.twitter.com/company
[2] http://www.socialbakers.com/statistics/facebook/

emoticons, hash tags, contextual texts, etc., This creates a problem to analyze the polarity of words. After preprocessing the tweets, the general sentiment analysis tasks will be done using lexicon or machine learning based approaches. In this work we have use the lexicon based approach to classify the tweets as positive or negative using the emoticon and the contextual sentiment orientation of the words. We have proposed a new algorithm Senti_Lexi to provide better accuracy. This paper has also presented a big data platform for processing sentiment analysis. Measuring the depth of the sentiment, which mostly relies on the contextual word, is one of the major issues and challenges involved in the process of sentiment analysis. The proposed algorithm has mainly focused on Emoticon and contextual words.

The rest of this paper has been organized as follows: In section background, a brief review on big data, sentiment analysis tasks and approaches have been provided. In the literature review section, a brief review on some related work on sentiment analysis using big data analytical tools has been given. The methodological framework for sentiment analysis on big data has been presented in the section 'research design and methodology'. The proposed algorithm has been explained in the section 'proposed work'. In section software package, the detailed overview about the tools available for this problem has been provided. Section conclusion has concluded the paper.

II. BACKGROUND

A. Big Data

In the world there are 7,388,180,600 billion people as of statistics on Dec 2015[3]. From that 3,266,493,250 users are using the Internet as of statistics on Dec 2015[4]. The data is being generated simultaneously from various sources including social media websites such as, YouTube, Twitter, Facebook, etc., and other applications such as weather prediction, e-governance, health care management leads to big data [4]. Big data is a buzz word. There is no standard definition of big data. The definitions given by some organization are given below.

The Big Data Commission at the Tech America Foundation offers the following definition:

"Big Data is a term that describes large volumes of high- velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information " [5]

Researchers at McKinsey have proposed an intentionally subjective definition:

"Big data refer to datasets whose size is beyond the ability of the typical database software tools to capture, store, manage, and analyze" [6].

B. Sentiment Analysis

Sentiment analysis [7], also called *opinion mining*, is the field of study that analyzes people opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. There are also many other names, e.g., *sentiment analysis*, *opinion mining*, *opinion extraction*, *sentiment mining*, *subjectivity analysis*, *affect analysis*, *emotion analysis*, *review mining*, etc. However, they all are now under the umbrella of sentiment analysis or opinion mining [7]. While in industry, the term *sentiment analysis* is more commonly used, but in academia, both *sentiment analysis* and *opinion mining* are frequently employed. They basically represent the same field of study.

It contains a huge problem space. For example, a sentence "This car is expensive, but nice", is apparently positive, as it expresses a favorable sentiment of the speaker, who is probably going to purchase the product. However, the sentence "This car is nice, but expensive" is negative, as it expresses the lack of enthusiasm of the user for purchasing the product.

Sentiment analysis approaches

Three approaches for performing sentiment extraction are as follows and described in the table 1:

(i) machine learning approach,
(ii) lexicon-based approach and
(iii) hybrid approach.

Table 1. Sentiment Analysis Approaches

Sentiment classification	Approaches	Advantages and Limitations
Lexicon based	Dictionary based. Corpus based. Ensemble approaches.	**Advantages** Best method for domain dependent. High accuracy for domain dependent. Wider term coverage. **Limitations** Finite number of words in the lexicons

[3] http://www.worldometers.info/world-population/
[4] http://www.internetlivestats.com/internet-users/

Machine learning based	Support Vector Machines, Bayesian Networks, Naïve Bayes, Random Forest, Maximum Entropy.	**Advantages** Ability to adapt and create trained models for specific purposes and contexts. **Limitations** Low applicability on new data, due to the demand of labeled data.
Hybrid based	Lexicon and Machine Learning based.	**Advantages** High accuracy on new data. Sentiment lexicon constructed using public resources for sentiment detection. Sentiment words as features in machine learning method. **Limitations** Noisy data

III. LITERATURE REVIEW

Geeta et al. [8] sentiment analysis is the process of detecting the contextual polarity of text. This work has focused on one of the famous micro blogging platform, Twitter, for performing sentiment analysis. A simple and completely automatic approach has been proposed for analyzing the sentiment of users on Twitter. Tweets express positive and negative polarity through a completely automatic procedure by using only emoticons in tweets. The sentiment classifier has been built to process an actual creek of tweets and its content have been classified as positive, negative or neutral. The classification has been made without the use of any pre-defined dictionary or polarity thesaurus. The thesaurus has been automatically inferred from the scraping of tweets. The author has observed that the simple system captured the polarity perceptions matching reasonably well with the classification done by human judges.

Vo Ngoc Phu et al. [9] have explored different methods of improving the accuracy of sentiment classification. The sentiment orientation of a document can be positive, negative, or neutral. A new combined dictionary with 21137 entries has been built. The new dictionary has many verbs, adverbs, phrases and idioms which are not present in the existing dictionary. The proposed method based on the combination of Term-Counting method and Enhanced Contextual Valence Shifters method has improved the accuracy of sentiment classification. The combined method had an accuracy of 68.93% on the testing dataset, and 69.224% on the training dataset. All of these methods were implemented to classify the reviews based on new dictionary and the Internet Movie data set.

Ramesh R, et al. [10] have proposed machine learning based approach for sentiment analysis for datasets from social media, which can be termed as big data. In order to process the big data, big data technology hadoop was used. The author has aimed at improved accuracy of results, and speed of processing.

Ilkyu Ha et al. [11] have proposed other processes of sentiment analysis on hadoop framework to enable parallel process of data. The proposed work used HDFS for storage and MapReduce function for sentiment analysis. The presented work has improved in less time through parallel processing.

Hasan Saif et al. [12] have performed sentiment analysis on Twitter which attracted much attention recently due to its wide applications in both, commercial and public sectors. In this paper SentiCircles, a lexicon-based approach for sentiment analysis on Twitter has been proposed. Different from typical lexicon based approaches, which offer fixed and static prior sentiment polarities of words regardless of their context, SentiCircles has taken into account the co-occurrence patterns of words in different contexts in tweets to capture their semantics and update their preassigned strength and polarity in sentiment lexicons accordingly. This approach has performed for the detection of sentiment at both entity-level and tweet-level. The proposed approach has been evaluated on three Twitter datasets using three different sentiment lexicons to derive word prior sentiments. The results have shown that the proposed approach significantly outperformed the baselines in accuracy and F-measure for entity-level subjectivity (neutral/polar) and polarity (positive/negative) detections. For tweet-level sentiment detection, the proposed approach has performed better than the existing SentiStrength methods by 4–5% in accuracy in two datasets, but falls marginally behind by 1% in F-measure in the third dataset.

Suresh et al. [13] have highlighted the web as an excellent source for assembling consumer opinions, such as customer reviews of products, forums, discussion groups, and blogs. This paper has focused on online customer reviews of productsand made two contributions. First, it has proposed a framework for analyzing and comparing consumer opinions of competing products in map and reduce environment for better analysis. Second, a new lexicon based technique has been proposed to extract neutral reviews and restrict them from being categorized under positive or negative. Experimental results have proved the effectiveness of the proposed technique and its ability to defeat the existing method significantly.

Chetan Kaushik et al. [14] have found a technique to efficiently perform sentiment analysis on big data. In the research, sentiment analysis has been performed on a large data set of tweets using Hadoop and the performance of the technique has been measured in the form of speed. The proposed technique has produced 73.5% accuracy. The experimental result

has proved the efficiency of the proposed technique in handling big sentiment data sets. Though the proposed technique has been implemented in the single node sandboxed configuration, it could also be implemented in the multimode configuration for better results.

IV. MOTIVATION AND OBJECTIVE

From the above literature review, most of the works have been carried out using lexicon based approach. The emoticon and contextual words has been used separately to find the polarity of tweets. This motivated to build a new model based on combining both emoticon and contextual word identification.

The objective of this work is to provide a novel lexicon based approach using emoticon and contextual word identification to classify the sentiment words of tweets and also to establish a big data platform for sentiment analysis.

V. RESEARCH DESIGN AND METHODOLOGY

In fig 1 the methodological diagram has been shown to represent an overall method for sentiment analysis on big data and to provide the outline of this paper. This research work mainly falls into Level 2.

Level 1: Collecting the tweets from twitter using API tools and then preprocessing and storing them in the desired format. The preprocessing step includes stopwords removal, url removal, audio and video removals, stemming, and lemmatization.

Level 2: An algorithm has been proposed for emoticon handling, contextual word; and briefly explained in the proposed work section. A mathematical approach has been used for sentiment score computation to classify the tweets as positive, negative and neutral.

Level 3: Using the commonly used polarity measures such as precision and recall, the accuracy and F-score will be calculated. Later, the positive, negative and neutral tweets will be classified with and without emoticon and contextual word; this will show the improved accuracy.

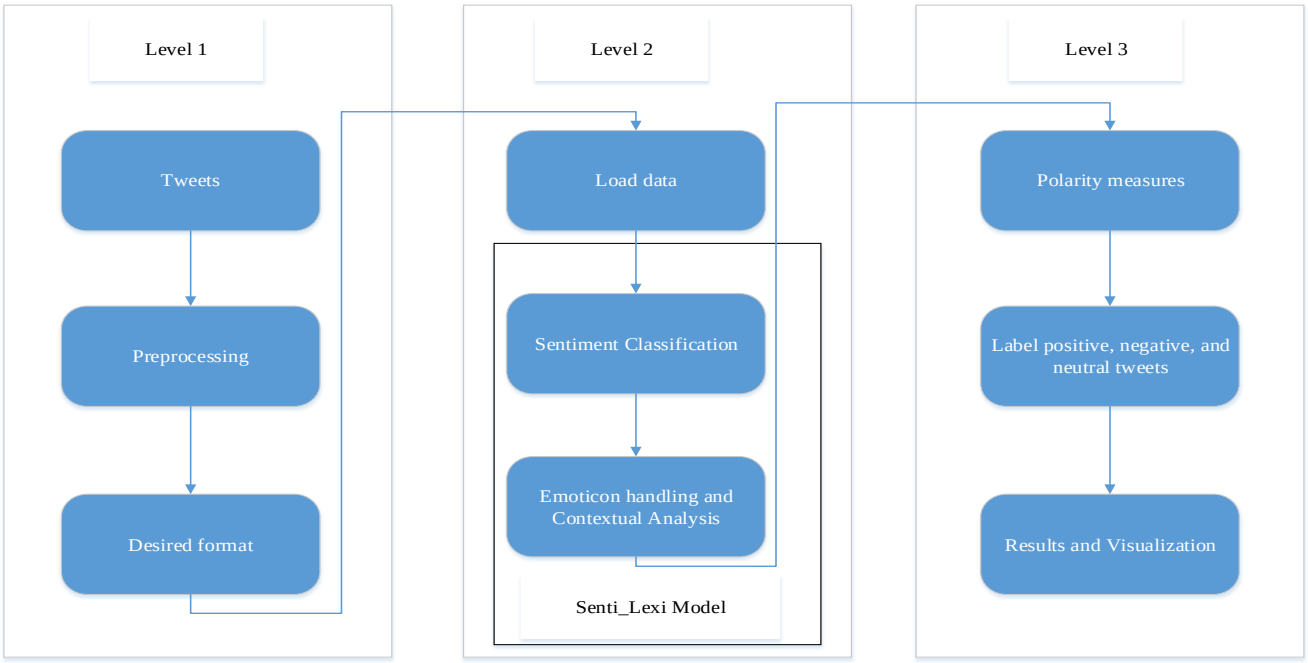

Fig 1. Methodological diagram for Sentiment Analysis on Tweets

VI. PROPOSED WORK

The overall process of the proposed algorithm has been represented in the Fig 2. The concept of the proposed algorithm is to evaluate the sentiment knowledge from both emoticon and contextual words. Analyzing emoticon and contextual word along with the popular sentiment dictionaries (Sentinet, WordNet,etc.) will produce greater accuracy results than the existing works. Consider each tweet as a sentence. The proposed algorithm will check the presence of emoticon and contextual words in each sentence. Considering the emoticons for polarity calculation will give more accuracy. If a tweet contains only positive emoticons and no negative emoticons, it is classified as positive. If a tweet contains only negative emoticons and no positive emoticons, it is classified as negative. If a tweet contains no emoticons, then it checks the presence of contextual word. The value for contextual word will be calculated by using the process Contextual Valence Shifter (CVS) [9]. If a tweet contains no contextual word, then the condition, map the opinionated word in the sentiment dictionary. Based on the calculation the positive, negative and neutral tweets will be classified. The algorithm for the proposed work is represented in the Fig 3.

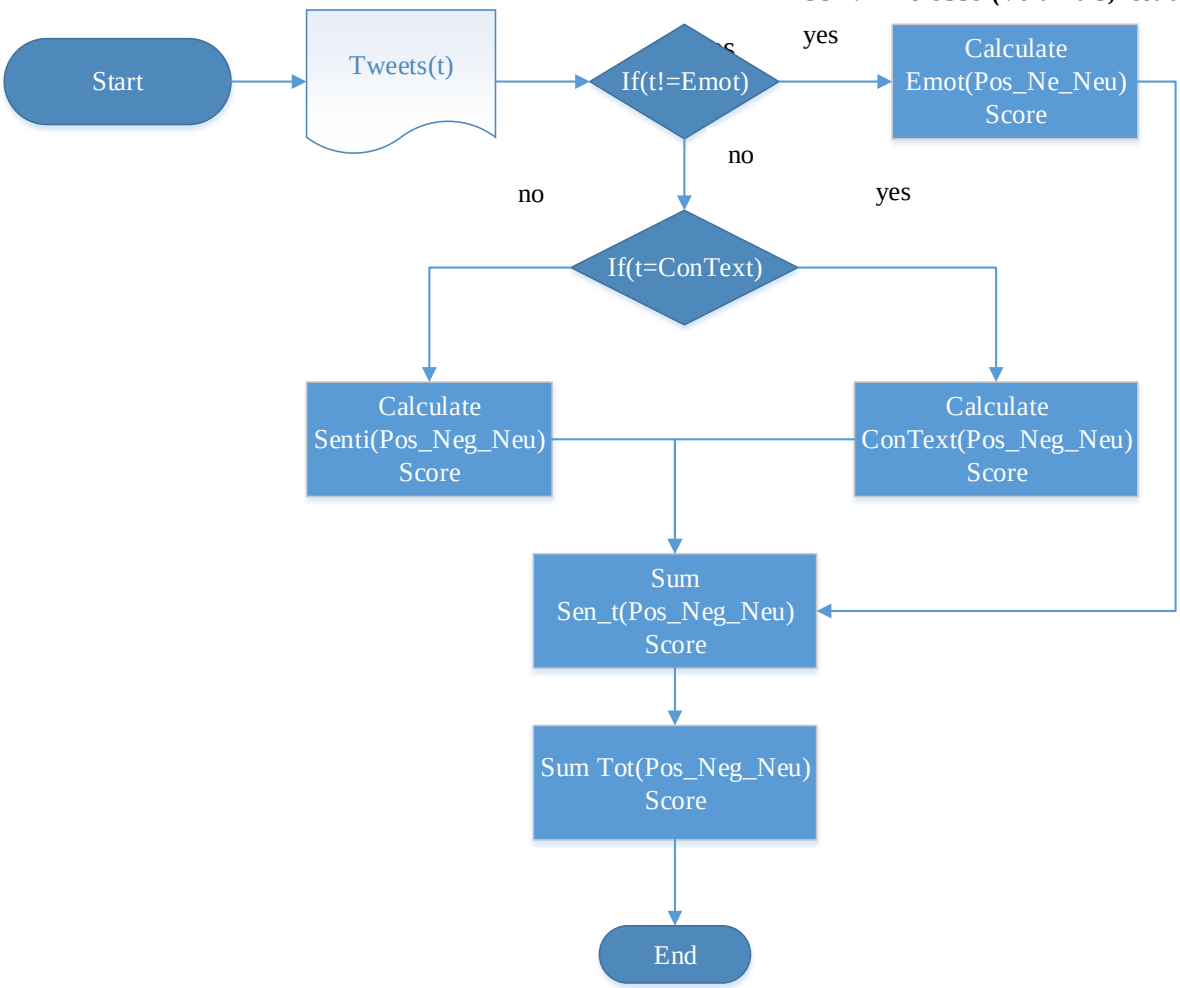

Fig 2. Process of Senti_Lexi Model

Algorithm: *Senti_Lexi*
Input: Tweets
Output: Positive, Negative, Neutral tweets
1. Begin
2. Each tweet
3. Each word in tweet
4. Find each word in a tweet in sentiment dictionary
   If emoticon found
      Calculate positive, negative, neutral.
   elseIf contextual word found /*Contextual Valence Shifter [9]*/
      Calculate positive, negative, neutral valence.
   elseIf sentiment word found
      Calculate positive, negative, neutral.
   Sum positive, negative, neutral for each sentence.
5. End
6. Count total positive, negative, neutral tweets.
7. End

Fig 3. Senti_Lexi Algorithm

VII. SOFTWARE PACKAGES

Choosing the best tool is the most difficult task for the industrialists and academicians. This section provides a brief introduction for establishing big data analytical platform for undertaking effective sentiment analysis. From the existing related works and social blogs the top three open source tools has been discussed. Fig 4 represents the tools used for the sentiment analytical process on big data. Flume is the appropriate tool to collect the tweets and loads in the HDFS (Hadoop Distributed File System). The tools are described below.

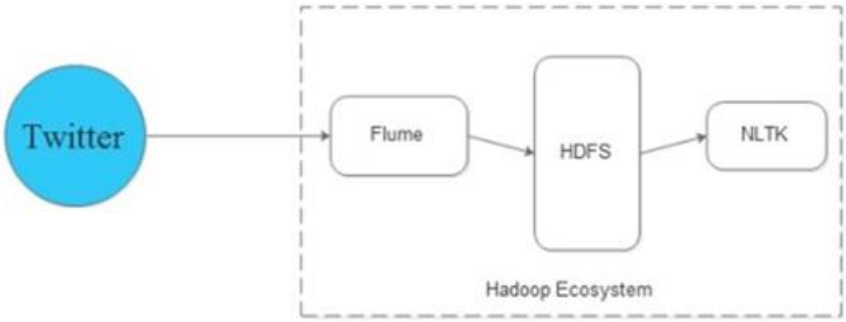

Fig 4. Big Data Platform for Sentiment Analysis

A. Flume[5]

Flume is a tool built by Cloudera that acts around the Hadoop cluster [15]. It ingests log data from multiple web servers into a centralized store (HDFS, HBase) efficiently. It supports a large set of sources and destinations types. It can be scaled horizontally. Flume acts as a channel to import huge volumes of event data produced by social networking sites like Facebook and Twitter, and e-commerce websites like Amazon and Flipkart into HDFS at a higher speed.

B. HDFS

Hadoop is designed to scale up from single server to thousands of machines, each machine offering local computation and storage. HDFS (Hadoop Distributed File System)[6] is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable, extremely rapid computations.

C. NLTK[7]

NLTK is a leading platform for building Python programs to work with text data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, SentiNet, etc, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

VIII. CONCLUSIONS

The methodological diagram discussed in this paper delivers a big picture for processing sentiment analysis. This paper has presented a novel lexicon based model for analyzing sentiment analysis on big data. A new algorithm Senti_Lexi has been proposed to provide better accuracy while analyzing the emoticon and contextual text will produce better accuracy than the existing works. The big data platform for processing sentiment analysis was discussed and provided. In future the proposed model will be implemented using the software packages that are discussed in the above section. Less amount of research has been carried out in emoticon and contextual based analysis, will bring more issues and challenges to the industrialists and academicians. Improving the emoticon and contextual word dictionaries will produce better results.

REFERENCES

[1] M.R. Saleh, M.T. Martin-Valdivia, A. Montejo-Raez and L.A. Urena-Lopez, “Experiments with SVM to classify opinions in different domains”, Expert Systems with Applications, Volume 38, Issue 12, November–December 2011, Pages 14799–14804.

[2] W. Medhat et al. “Sentiment analysis algorithms and applications: A survey”, Ain Shams Eng J, Volume 5, Issue 4, December 2014, Pages 1093–1113

[3] O. Popescu, and C. Strapparava, “Time corpora: Epochs, opinions and changes”, Knowledge Based Systems, Issue 3, Vol-13, 2014.

[4] Jenie Arock X, Vaitheeswaran G and Dr. L. Arockiam, “Parallelized Contextual Valance Shifter Algorithm for Sentiment Analysis on Big Data”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 9, September 2015, Pages 590-595.

[5] Amir Gandomi and Murtaza Haider, “Beyond the hype: Big data concepts, methods and analytics”, Elsevier, Volume 35, Issue 2, April 2015, Pages 137–144.

[6] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers, “Big data: The next frontier for innovation, competition, and productivity”, Mckinsey global institute, June 2011.

[7] Bing Liu, “Sentiment Analysis and Opinion Mining”, Morgan & Claypool Publishers, May 2012.

[5] http://www.tutorialspoint.com/apache_flume/apache_flume_introduction.htm

[6] https://thebigdatainstitute.wordpress.com/2013/04/29/introduction-to-big-data-and-hadoop-ecosystem-for-beginners/

[7] http://www.nltk.org/

[8] Geeta.G.Dayalani, Dr.Seema and Prof.B.K.Patil, "Emoticon-based unsupervised sentiment classifier for polarity analysis in tweets", International Journal Of Engineering Research and General Science, Volume 2, Issue 6,October-November, 2014.

[9] Vo Ngoc Phu and Phan Thi Tuoi, "Sentiment Classification using Enhanced Contextual Valence Shifters", International Conference on Asian Language Processing, IEEE, 2014, Pages 224 – 229.

[10] Ramesh R, Divya G, Divya D and Merin K Kurian, "Big Data Sentiment Analysis using Hadoop", International Journal for Innovative Research in Science & Technology,) Volume 1, Issue 1, 2015.

[11] Ilkyu Ha, Bonghyun Back and Byoungchul Ahn, "MapReduce Functions to Analyze Sentiment Information from Social Big Data", International Journal of Distributed Sensor Networks, 2015.

[12] Hassan Saif, Yulan He, Miriam Fernandez and Harith Alani, "Contextual semantics for sentiment analysis of Twitter", Information Processing & Management, Volume 52, Issue 1, January 2016, Pages 5–19.

[13] R. Suresh Ramanujam, J. Nivedha, R. Nancyamala and J. Kokila, "Sentiment Analysis Using Big Data", IEEE 2015, International Conference on Computation of Power, Energy, Information and Communication, 6, 2015.

[14] Chetan Kaushik and Atul Mishra, "A Scalable, Lexicon Based Technique For Sentiment Analysis", International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.5, September 2014

[15] Anand Loganathan, Ankur Sinha, Muthuramakrishnana and Srikanth Natarajan, "A Systematic Approach to Big Data Exploration of the Hadoop Framework", International Journal of Information & Computation Technology, Vol-4, Issues. 9, 2014, pp. 869-878.

