

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

Concepts and Methods of Sentiment Analysis on Big Data

M. Edison ¹, A. Aloysius ²

Research Scholar, Dept. of Computer Science, St. Joseph's College (Autonomous) Tiruchirappalli, TN, India.¹

Assistant Professor, Dept. of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, TN, India.²

ABSTRACT: Data analysis is the biggest challenging issue in the world expressly huge volume of data. The rapid fruition of big data is becoming a problem and to solving it is another big issue in the industry. Exclusively, Social Media Networks (SMN) are generating a massive tome of data which contains structured, semi-structured and unstructured data. In this era, these data are having audio, video, text, numbers, hashtag and URLs. Consequently, the data need to be extracted and analysed from the variety regarding big data. Big data consists of different analyses, this paper focuses on Sentiment Analysis. To perform Sentiment Analysis, the concepts of approaches, techniques, models and features have been proposed. To solve the text and web based issues in big data the sentiment analysis (SA) is useful. Therefore, the concepts of the SA and Hadoop ecosystems are explored in the paper.

KEYWORDS: Big Data, Sentiment Analysis, Machine Learning, Analytics.

I. INTRODUCTION

This paper represents a basic concepts of big data. It focusses on the characteristics, metrics define size and exist fine potential of big data. The sudden rises in the issues of big data which presents in the academics, municipals and in the industries need a development of new technologies to resolve it. Big data are stretching enormous data set so it's very complex to process traditional data. The application process these data are also insufficient. Challenges are included for analysing data, sharing the information, storage, creation, visualization and querying updating information. Sentiment analysis is spoken of the computational study of text analysis to find and extract the subjective study in big data. Generally, SA is applied to analyse the user's review from social media networks such as: Facebook, Twitter, Amazon, etc., from the marketing customer's review about the product.

Sentiment analysis intentions to decide the attitudes, appraisals and moods of the user with respect to some of the reviews or complete contextual isolations of a document. The basic task of sentiment analysis is to classify the polarity in different levels like Document level, Sentence level and Aspect level or entity level. The sentiment or opinion expressed emotions are classified in different classes as positive, negative and neutral. Emotional states are expressed like "angry", "happy", "sad" etc. This paper is organized as follows. In Section 2, background details related with Big Data and Sentiment Analysis are discussed. Section 3 describes the Big Data Analytics. Sentiment Analysis Approaches and Techniques are described deeply in section 4. Section 5 summaries the conclusion.

II. RELATED WORK

Big data concepts have been described and proposed new technological issues and challenges. This area focuses on the business organization to enrich the business value [2]. Many applications and organizations are considered to categorize and summarize the sentiment algorithms and techniques have given the contribution [3]. Sentiment Analysis is challenging to forecast the movie review, especially, document level sentiment classification is difficult to determine the sentiment whether it is positive, negative or neutral. Bag of Words (BOW), one of the techniques of SA, improves the algorithm to enhance the accuracy [8]. The aspect oriented sentiment analysis comes under the product reviews to analyse and extract the data. The data have measured polarity about the product [10] entity. The particular investigation in to big data focuses technical algorithm or system development which could be considered its adoption by way of the growing potential of the big data paradigm [9]. Lexicon and learning based sentiment analysis has proposed by the

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

Andrius. However, the paper comprises a novel approach for learnings [12]. The psenti techniques have been used to predict the reviews whether they are positive or negative.

III. BIG DATA ANALYTICS

“Big data analytics is the process of examining large data sets containing a variety of data types -- i.e., big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information.”

The main objective of the big data analytics [4] is helpful to the companies to take supplementary informed business decisions through qualified data scientists. The big data analytics framework is illustrated below in fig. 1.

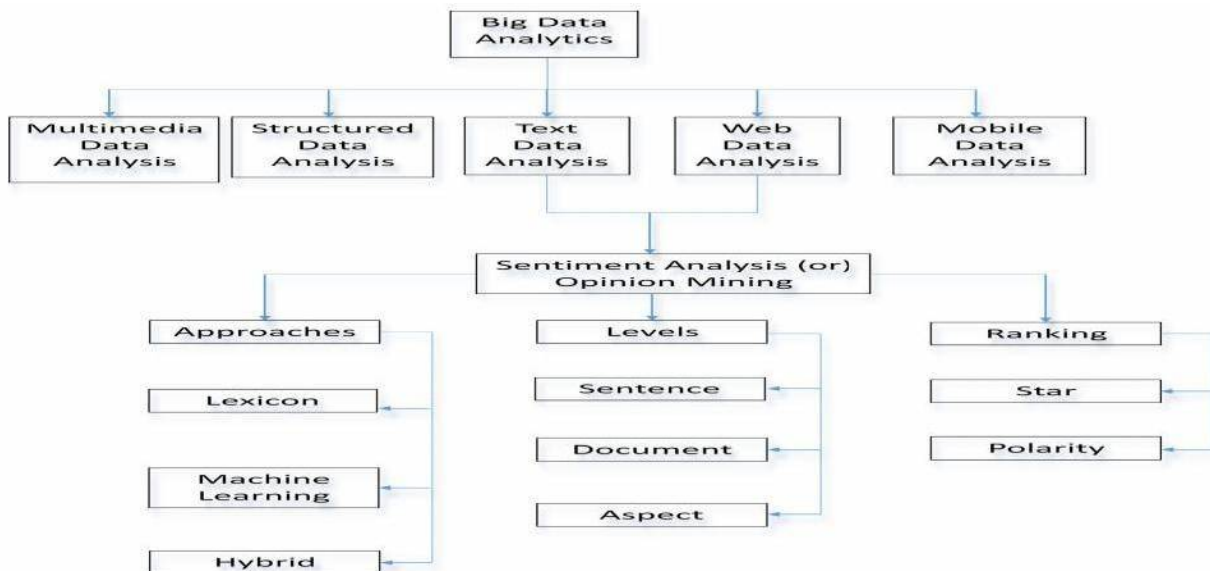


Fig. 1. A Framework of Analytics and Sentiment Analysis on Big Data

A. Big Data Definitions

“Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” (“Gartner IT Glossary, n.d.”) [5]. Similarly, TechAmerica Foundation defines big data as follows:

“Big data is a term that describes large volumes of high velocity, complex and varied data that require advanced techniques and technologies to enable the capture, storage, distribution, Management, and analysis of the information.” (TechAmerica Foundation’s Federal Big Data Commission, 2012).

B. Multimedia Data Analysis

Multimedia analysis is a processing of text retrieval, image processing, audio analysis and video analysis. Through the internet, vast amount of multimedia information are emerging. Concurrently, digital camera, video camera and android mobiles are used to record multimedia items, those information have been taken and analysed.

a. Audio Analysis

Audio analysis is to analyse and extract information from unstructured audio data. The analysis of the human spoken language is known as audio analysis which is also referred as speech analytics or speech analysis. Nowadays, the customer service centre, health care service centre etc., is often doing audio analysis for promoting a business. The speech analysis involves two approaches such as: transcript-based approach and phonetic-based approach. The

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

transcript based approach is known as large vocabulary continuous speech recognition (LVCR).

b. Video Analysis

Video analysis is called as video content analysis (VCA), it comprises the techniques to analyse, monitor and extract the information from the video stream. Already most of the techniques have been developed to process and record the video data like video streaming. The surveillance of prevalence closed circuit television (CCTV) cameras is successfully booming popular video sharing websites. In a big data, video analysis is a big challenging problem for forecasting and leverage the automatic evolution of video analysis.

The initial application of video analysis in modern years has been fashionable computerized security and surveillance systems. Totally, their cost of the labour surveillance system is less effective than the automatic system. There are two existing approaches are available in video analysis, namely server based approach and edge based approach.

C. Structured Data Analysis

Structured data analysis is the computational study of the mathematical and statistical data analysis. This can arise either in the form of preference for the structure such as multiple-choice questionnaires or in situations with the need to search for structure that fits the given data, either exactly or approximately.

D. Text Data Analysis

Text analysis refers to containing a feed, microblogs, blogs, tweets, and emails. It has been extracted from the textual data. Text data analysis consists of computational study, machine learning and statistical analysis. Generally, user's generating a large amount of data from business to convert expressive things which is based on the decision making to predict the business intelligence.

Information Extraction (IE) technic helps to extract the data from unstructured to structured data. Mainly IE algorithm is working as extract structured information. Text summarization techniques have produced concise instant of the single or multiple document which monitors two approaches like extract approach and abstract approach.

E. Web Data Analysis

Web data analysis is analysing the data from the web sources. It has to be concentrated collection, measurement, analysis and reporting. Moreover, web analytics application measure the traditional advertisement. It helps to identify the users and provide the visitor information, this is very popular in the marketing research. Web analytics have a basic analytic process and advanced techniques.

F. Mobile Data Analysis

Mobile Analytics is a tool which analyses the features of the mobile channel to enhance higher performance. The platform is especially absorbed on those metrics associated with mobile navigation, among which are: conversions, campaigns, visit times, origin, bounce and many others. Mobile Analytics can have monitored mobile user behaviour on sites, enabling to check and analyse how the smartphone users behave when visiting classic web pages.

IV. SENTIMENT ANALYSIS APPROACHES AND TECHNIQUES

Sentiment Analysis

"It is the computational study of people's opinions, appraisals and emotions toward entities, events and their attributes. Opinions are important because whenever we need to make a decision – "we listen to other's opinions"." (Bing Liu). The sentiment classification techniques described figure [5] illustrated below in fig. 2.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

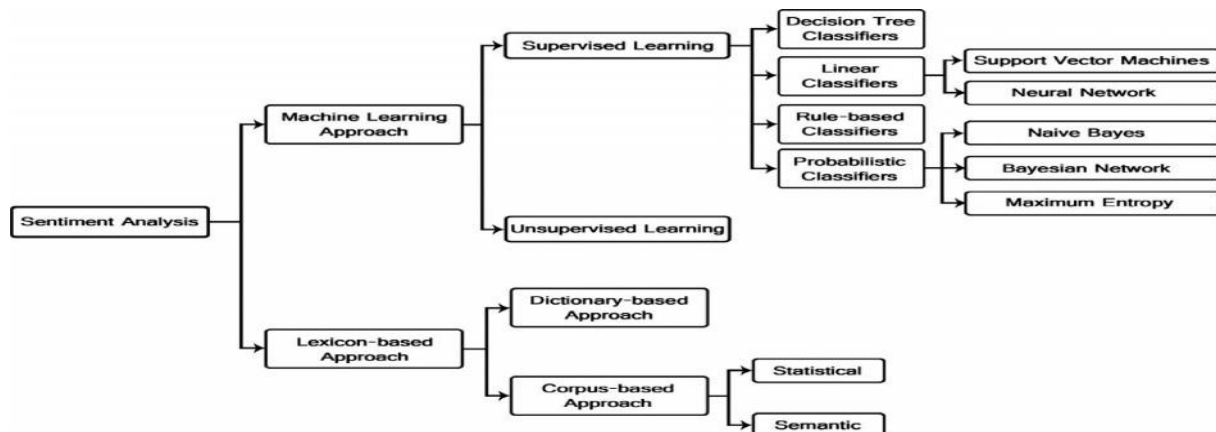


Fig. 2. Sentiment Classification Techniques.

A. Approaches

a. Lexicon Based Approach

The lexical or lexicon approach is a method for teaching dictionary based described by Michael Lewis in the early 1990s [11]. The basic concept of this approach respites an idea that the significant part of education involves understanding and produce lexical phrases as chunks. In this pattern of language like grammar as well as consume meaningful set of words at their dumping.

b. Machine Learning Approach

Machine learning is a subset of computer science that developed from the study of pattern acknowledgement and computational study theory in artificial intelligence [12]. Machine learning gives study for computer learning ability with explicit program. Which explores the study and creation of algorithms that can be made a prediction on data. The algorithms function of constructing a model from an example training set of input clarifications in order to make data-driven predictions or decisions stated as outputs. Machine learning is associated to computational statistics. It has a strong mathematical optimization. There are two different types of machine learning is applicable [1].

1) Supervised Learning

Supervised learning is the machine learning task of deducing a function from labelled training data. The training data contain a set of training examples. Fashionable supervised learning, every examples are pair comprising of input object and a desired output value. This process required wide practices in the machine learning. Different types of techniques are obtainable in supervised learning such as classification and regression.

Here the goal of the classification is to learn the mapping input from x to output y , where $y \in \{1, \dots, C\}$, with C being the number of classes. If $C = 2$, this is called binary classification (in which case we often assume $y \in \{0, 1\}$); if $C > 2$, this is called meticulous classification. Regression has been just like classification excluding the response variable is unremitting. Which is a single real-valued input $x_i \in \mathbb{R}$, and a single real-valued response $Y_i \in \mathbb{R}$. This is to consider fitting two models to the data a straight line and a quadratic function.

2) Unsupervised Learning

Unsupervised learning is a type of machine learning algorithm, which is used to draw inferences from data sets containing of input data without labelling reactions. The most common unsupervised learning method is cluster analysis, this is used for examining the data analysis to discover hidden patterns or grouping data.

c. Hybrid Approach

Hybrid approach is used for both combinations of machine learning and lexicon based approach. This approach has to prove

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

the combination whether it is enhanced results or performance of classification. The main advantages of this approach are symbiosis stability uses of learning and lexicon algorithms for sentiment deduction [7].

B. Levels

a. Sentence Level

Sentence level approach goes to the sentences to determine whether the sentence communicated as positive, negative, or neutral opinion. Usually neutral means no opinion. This level of analysis is closely related to subjectivity classification which decides sentences called objective sentences [6]. The factual information from sentences called subjective sentences that prompt subjective views and opinions. However, the subjectivity is not equivalent to sentiment as many objective sentences can imply opinions.

b. Document Level

The document level approach is to classify whether the whole document expresses as positive or negative sentiment. The system concludes whether the review states an overall positive or negative opinion about the product. This task is commonly Sentiment Analysis and Opinion Mining known as document-level sentiment classification. This level of analysis assumes that each document expresses opinions on a single entity. Thus, it is not appropriate for documents which appraise or compare numerous entities.

c. Aspect Based Level

The aspect based level approach carries both the document level and the sentence level analyses. It does not disclose what exactly people like or dislike but it makes people realize the opinion about the product entity. Aspect level performs fine-grained analysis is also called feature level. This approach directly expresses the sentiment which consists of positive, negative and neutral. For example, the sentence “*The iPhone’s call quality is good, but its battery life is short*” clearly has a positive tone for the first part, the second part is having a negative opinion [8]. It cannot be assumed that the sentence is entirely positive or entirely negative. This might be evaluated in two aspects, how the aspect level is working, qualitative and quantitative analysis and both the sentence and document level classification is very challenging and the aspect level is more difficult.

V. HADOOP ECOSYSTEM

Apache Hadoop is an open source framework for distributed storage and processing the massive data set. Hadoop permit vast amount of data which are broken into smaller components and deals to store huge files across different machines. Therefore, the analysis might be done quickly and cost-effectively. Generally, the hadoop modules are designed as a distributed file system like Hadoop Distributed File System (HDFS) which is consider the part of processing called MapReduce [13]. HDFS files system store the massive data information and there is no loss of data. Suppose the data have lost, then the system send a file to another distributed environment for running the application. There is another file system is working in Hadoop environment named as HBase. HBase is a column based data (Data Base (DB)) storage system. Especially, this is worked as a distributed environment but it comprises the data in column format which means structured format. Fig. 4 is illustrated below for the processing procedure of Hadoop Ecosystem [14].

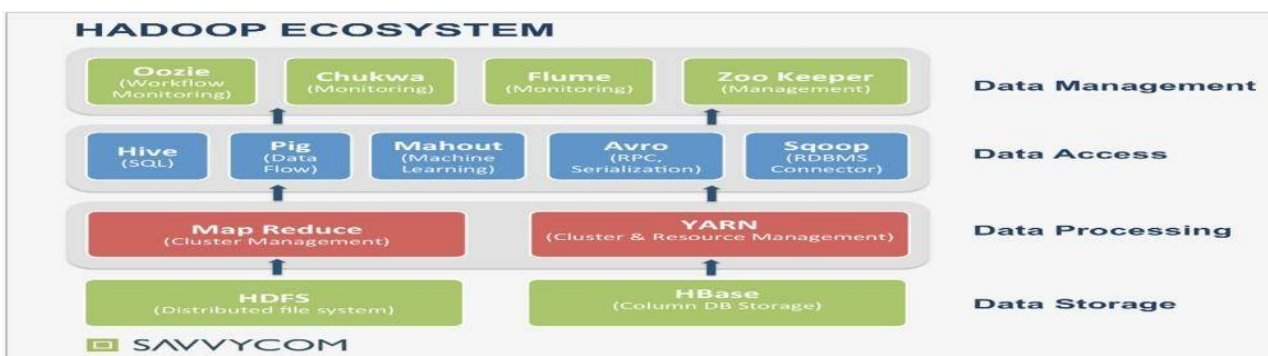


Fig. 3. Hadoop Ecosystem

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

B. Data Processing

a. MapReduce

MapReduce is one of the programming model for processing enormous data set with a corresponding distributed algorithm on a cluster. The MapReduce algorithm contains two different tasks such as Map and Reduce. Map task, taking a set of data and it converts into another set of data, where separate elements are smashed into tuples. Then, reduce task, taking the output and takes input from map for combine individual data tuples into a smaller set of tuples. This process is working as sequential of the term MapReduce. The Map is perform as first always secondly, the Reduce can perform job. Apache MapReduce paradigm is created over Apache YARN Framework. The fig. 4 is shown below for the MapReduce function methodology [15].

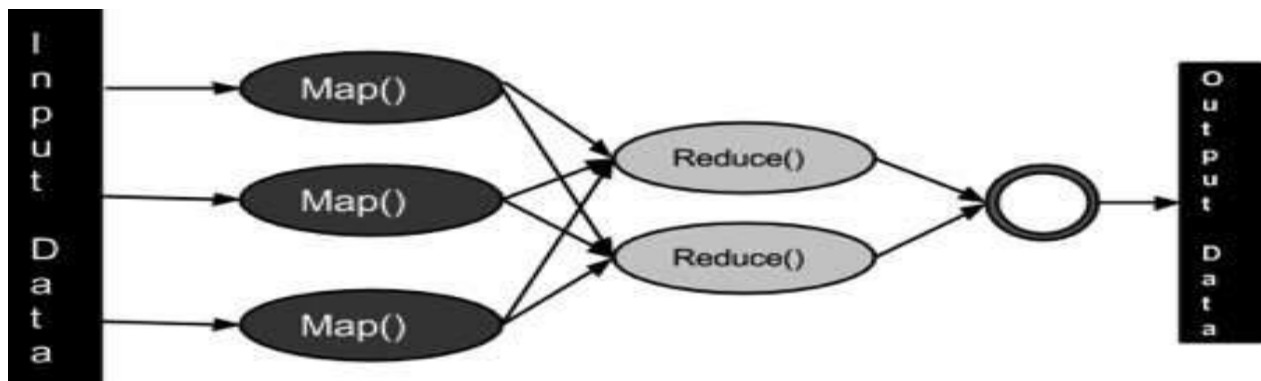


Fig. 4. Methodology for MapReduce Function.

b. Yarn (YARN)

YARN is denoted by “Yet-Another-Resource-Negotiator” which is a new framework that helps to writing a random distributed processing structures and applications. The execution model of MapReduce operation is more generic than the earlier. YARN applications are not followed by MapReduce model, unlike the original Apache Hadoop MapReduce and Hadoop YARN is try to take MapReduce for data-processing.

C. Data Access

Data analysis is a big task in this world because the Social Media Networks (SMN), Medical reports, Traffic data and Weather reports are generating more and more ambiguous data. These medium data are considered as 1 Peta Byte (PB) approximately in one hour. The data have to be evaluated and isolated, whether the data are consider to format unstructured data into structured data. In this part, most of the tools are available to segregate the data in Hadoop environment. Hadoop framework has accomplished many tools such as Pig or Pig Latin, Hive, Mahout and Sqoop for the purpose of data accessing and the data management tools are also available in Hadoop environment named as Flume, Oozie, Chukwa and keeper.

a. Pig

Pig is one of the tool adapted by Hadoop framework. This is a high level language platform engine to analyse enormous data set. It provides an arbitrary access for analysing and implementing data flow in Hadoop which includes many operators for legacy data operations like sort, concatenation, filtering and etc. The users can create an own function namely method for writing, reading and processing. Pig language can working with both a support of Hadoop Distributed File System (HDFS) and Hadoop Processing System (HPS). Pig setup consist of own compiler to yield MapReduce program sequentially. Currently, Pig language consist of textual language called Pig Latin.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

b. Hive

“Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. While developed by Facebook, Apache Hive is now used and developed by other companies such as Netflix and the Financial Industry Regulatory Authority (FINRA)” [18].

c. Mahout

“Apache Mahout is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification. Many of the implementations use the Apache Hadoop platform. Mahout also provides Java libraries for common maths operations (focused on linear algebra and statistics) and primitive Java collections. Mahout is a work in progress; the number of implemented algorithms has grown quickly, but various algorithms are still missing” [19].

VI. SENTIMENT ANALYSIS WITH BIG DATA

The Hadoop framework can support the language is Pig/Pig Latin. Nowadays, the researcher's mostly selecting the research domain is Sentiment Analysis in Hadoop environment and collecting the data from the Social Media Network (SMN). Social Media Network is generating an ambiguous data, that data must be isolated and evaluated for the necessary actions. Mostly, the SMN textual data exclusively reflected unstructured data. Therefore, the data need to be processed and removed unwanted data. Hereafter need to use Machine Learning (ML) techniques and algorithms for removing unwanted data, there are many algorithms are obtained in ML. Mainly, Machine Learning algorithms are offered to pre-process the data it should be stored in a structured format. In future, the format can help the users to classify the data sequentially and get clear cut idea for implementing the processed data to get a result.

The World Wide Web (WWW) provides plenty of users liking SMN. Especially, Twitter, Facebook, WhatsApp, and Amazon etc. These applications are given that abstruse data by the users. The user opinions are express their view easily and quickly, it represents ubiquitous. Sentiment Analysis can be consumed related to gather the information from these distinct sources. For the collected data can be combined, converted, or reformatted and formerly stored in a Hadoop Distributed File System (HDFS) in Hadoop environment.

Sentiment Analysis is the process for using text analytics to find various sources opinions. For the different sources collected opinions are unstructured data and the opinions are represent different distinct meanings. Here, the data need to classify in a sentence level. Usually, all the researchers would like to use the tool and algorithm for the classification and implementation of sentiment analysis as Support Vector Machine (SVM), Naïve Bayes and Maximum Entropy ML algorithms and the tools very helpful to implement the SA process is Pig tool.

Mainly, the ML techniques and algorithms are help to classify the data which characterizes whether the sentiment is positive, negative or neutral. The main goal of the polarity measurement is to predict the altered dimension with a help of ML techniques. In that formation the users (researcher) to be stored the polarized [17] opinion mixed into new dictionary format. Then the results will help the user to implement easily in the Hadoop environment. Pig tool is very famous tool for the text analytics to get a result for those analytics.

A. Process of Sentiment Analysis

Sentiment Analysis is also known as opinion mining. It is referred as the use of Natural Language Process (NLP), Text Analysis (TA) and Computational Linguistics (CL). Commonly, to identify and mine the objectivity and subjectivity information from the related sources. Exactly, this dispute is very difficult than polarity classification. Formerly the classification need to pre-process the data and then the actual data will represent, after that find the accuracy of the particular work may be used different measurement techniques like evaluation metrics, similarity measurement and statistical analysis. With a help of these approaches can support with pig tool to analyse the textual experiment and predict the result for the sentiments whether the result is positive, negative or neutral. In that measure totally how much classes are predicted and it reflects the contribution finally.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

a. Evaluation Metrics

We will evaluate our experiment results by using following Information Retrieval matrices [17].

1. Precision = $TP/(TP + FP)$
2. Recall = $TP/(TP + FN)$
3. F-measure = $2 * Precision * recall / (Precision + recall)$
4. Accuracy = $TP + TN / (TP + TN + FP + FN)$

b. Similarity Measurement and Statistical Analysis

The similarities are represent their individual measurement and analysis [20].

1. Text Similarity
2. Semantic Similarity
3. Cosine Similarity
4. Jaccard similarity
5. Euclidean distance
6. Minkowski distance
7. Pearson correlation
8. Spearman correlation
9. Tanimoto coefficient

c. Feature Selections

Feature selection is also known as variable selection, attribute selection or variable subset selection. Three reasons for feature selection

- ✓ Simplification of models to make them easier to interpret by researches/users
 - ✓ Shorter training times
 - ✓ Enhanced generalization by reducing overfitting
1. Information Gain (IF)
 2. Chi-square
 3. N-gram (unigram, bigram, trigram) etc.,

The above mentioned techniques are very helpful to analyse the sentiments and the techniques have been used to many of the articles for the purpose of classifications, pre-processing and measurement of the accuracy. Some of the features are performed on the stage where it needed based on the data set.

VII. CONCLUSION

This paper has explored the concepts of big data and sentiment analysis. The concepts are allowed to know more issues and challenges in the area of sentiment analysis on big data for the further research. Generally, the machine learning approaches are very important to classify the text or reviews from an ambiguous data. However, the notions are helpful to the researcher and whoever new to the area of applications, approaches and techniques in Sentiment Analysis. The overall literature contains analytics of big data, measurement of big data, issues, approaches, methods for predict the accuracy with a help of evaluation metrics/statistical analysis and ecosystem of the sentiment analysis on big data.

REFERENCES

- [1] Zhaoxia Wang, Victor Joo Chuan Tong and Hoong Chor Chin "Enhancing Machine Learning Methods for Sentiment Classification of Web Data", Springer International Publishing Switzerland, 2014, pp. 394-405.
- [2] Avita Katal, Mohammad Wazid and R H Goudar "Big Data: Issues, Challenges, Tools and Good Practices", IEEE, 2013, pp: 404-409.
- [3] Walaa Medhat, Ahmed Hassan and Hoda Korashy "Sentiment analysis algorithms and applications: A Survey", Ain Shams Engineering Journal (ASEJ) ELSEVIER, 2014, pp: 1093-1113.
- [4] Min Chen, Shiwen Mao and Yunhao Liu "Big Data: A Survey", Springer Science Business Media New York, 2014, pp: 171-209.
- [5] Amir Gandomi and Murtaza Haider "Beyond the hype: Big data concepts, methods and analytics", International Journal of Information Management (IJIM) ELSEVIER, 2015, pp: 137-144.
- [6] Xing Fang and Justin Zhan "Sentiment analysis using product review data", Journal of Big Data (JBD) Springer, 2015, pp: 1-14.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 9, September 2016

- [7] Chauhan Ashish P and Dr. K. M. Patel "Sentiment Analysis Using Hybrid Approach: A Survey", International Journal of Engineering Research and Applications, January 2015, pp: 73-77.
- [8] K. Mouthami, K. Nirmala Devi and Dr. V. Murali Bhask "Sentiment Analysis and Classification Based On Textual Reviews", Information Communication and Embedded Systems (ICICES), IEEE, 2013, pp: 271-276.
- [9] Ohbyung Kwon, Namyoon Lee and Bongsik Shin "Data quality management, data usage experience and acquisition intention of big data analytics", International Journal of Information Management, ELSEVIER, 2014, pp: 387-394.
- [10] A. Nandhini, G. Vaitheeswaran and Dr. L. Arockiam "A Hybrid Approach for Aspect Based Sentiment Analysis on Big Data", International Research Journal of Engineering and Technology (IRJET), 2015, pp: 815-819.
- [11] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil and Bing Liu "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", HP laboratories, 2011, pp: 1-7.
- [12] Andrius Mudinas, Dell Zhang and Mark Levene "Combining Lexicon and Learning based Approaches for Concept-Level Sentiment Analysis", Journal of American Society, 2012, pp: 1-8. [13] https://en.wikipedia.org/wiki/Apache_Hadoop. [14] <https://hadoopecosystemtable.github.io>.
- [15] http://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm.
- [16] Charles Uye "Perform sentiment analysis in a big data environment", developerWorks IBM, ibm.com/developerWorks, 2015, pp: 1-10.
- [17] L. Jeba Sheela "A Review of Sentiment Analysis in Twitter Data Using Hadoop", International Journal of Database Theory and Application, 2016, pp: 77-86.
- [18] <https://www.google.co.in/#q=hive+hadoop>
- [19] https://en.wikipedia.org/wiki/Apache_Mahout
- [20] Saikat Bagchi "Performance and Quality Assessment of Similarity Measures in Collaborative Filtering Using Mahout", 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), ELSEVIER, 2015, pp: 229 – 234.

BIOGRAPHY



M. Edison is a research scholar in the department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He is doing his Doctor of Philosophy in the area of Big Data. He has published research articles in his research area and he has attended many workshops and conferences.



A. Aloysius is working as an Assistant Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 16 years of experience in teaching and research. He has published many research articles in the National / International conferences and journals. He has acted as a chairperson for many national and international conferences. Currently, eight candidates are pursuing Doctor of Philosophy Programmed under his guidance.