Study on Sentiment Analysis in Big Data

[1]S. Gandhimathi, Research Scholar, PG & Research Department of Computer Science, Government Arts College (Autonomous), Karur – 639005, Tamilnadu, India.

[2]Dr. V. Baby Deepa, Assistant Professor, PG and Research Department of Computer Science, Government Arts College (Autonomous), Karur – 639005, Tamilnadu, India.

Abstract- The biggest challenging issue of data analysis concept is expressly bulky volume of data. Another big issue of big data is becoming the rapid function to solving it in an efficient manner. In recent years, SMN (Social Media Networks) is generating a huge tone of data that includes structured, semi-structured and unstructured data. The data which is mentioned in different formats such as audio, video, text, numbers, hash tag and URLs. Accordingly, the data require be analyzing and extracting from the variety regarding big data. Big data contains different analytical techniques. In this paper, the elaborate discussions focus Sentiment Analysis. Sentiment (Opinion) study on big data streams from the continuously generates text streams on SMNs to hundreds of millions of online consumer analysis provides several organizations in all the fields with opportunities to find out valuable intelligence from the enormous user generated text streams. An overview of Sentiment Analysis, various techniques for sentiment analysis and detailed on the various applications and challenges in sentiment analysis is presented.

IndexTerms- Big Data, Big Data Analytics, Social Media, Sentiment Analysis, Opinion Mining, Social Media Network (SMN), MapReduce, Apache Hadoop, Social Media Analytics (SMA)

I. INTRODUCTION

With rapid improvements and surge of internet popular companies like Yahoo, Amazon, Google, eBay and a speedily growing internet savvy population, today's enterprises and advanced systems are producing data in a multi-structured format and in a massive volume with great velocity including, videos, weblogs, images, sensor data etc. from various sources. This has to a new type of data known as Big Data that is unstructured sometime semi structured and also unpredictable in nature. This data is mostly generated from social media websites that has increasing exponentially on a daily basis. Internet usage among masses is increasing drastically mainly due to popularity of social media. In such that user-generated contents even if "any" form of user created content including: wikis, blogs, posts, forums, tweets, chats, or podcasts have to convey people's opinion. The huge packages of data generated by businesses, individuals, research agents and government, have undergone exponential growth. Different Big data technologies are provided such as, Hadoop, schema-less databases, Pig, Hive, PLATFORA aims at collection, storing and processing of big data in an effective and cheaper way. Big data does not only indicate that storing the bulky amount of data although analyzing of data and predicting some pattern in business intelligence. Sentiment analysis is known as opinion mining, is the field of revision that analyzes people's opinions, emotions, evaluations, appraisals, sentiments, and attitudes towards entities like services, products, individuals, organizations, issues, attributes and their events. Sentiment analysis is divided into three categories that include document level, sentence level and aspect based. The aim of document level sentiment classification is obtaining the overall sentiment of a given review document [1]. Sentence level analysis focuses on classifying the text at the level of objective and subjective nature [2]. Aspect based approach pinpoints and partitioned the entire document into different aspects (entities) and sentiment analysis is performed on each entity to discover the overall polarity [3]. In this paper, **t**he general scenario and challenges of Sentiment analysis are focused to explore the methodologies of opinion analysis of social network data.

II. BIG DATA ANALYTICS

Big Data refers as the Big Data Analytics to recognize the "hidden patterns, business information, unknown correlation, trends in market, user preference, and social network, and unknown statistical relations". The major part of Big Data Analytics is to combine massive amount of statistical data for utilizing a precise output within a prescribed time limit. The Big Data analytics is classified into three types such as Prescriptive, Predictive, and descriptive analytics [4]. In first stage of Descriptive Analytics, the beginning stage of data processing provides some idea to make utilization of historical data for prediction. In Prescriptive Analytics, refers to the process of examining the abstraction of an exact data similar to a particular field to improve the classification result. Using predictive Analytics technique, it uses the historical data or the past to provide the reasonable accuracy in future prediction. Google introduced a MapReduce programming model for processing and generating huge data sets on large number of computing nodes [6]. Two phase of a MapReduce program execution is provided map and reduce functions. In the map phase, map functions perform as input a list of key-value pairs and produce a set of intermediate output key-value pairs, which are stored in the intermediate storage. The reduce function handles each intermediate key to generate a final dataset of key-value pairs. Using this method, the map functions achieve data parallelism; at the same time as reduce workers perform parallel reduction. The MapReduce runtime usages are handled the resource management, parallelization, fault tolerance and other related problems. Apache Hadoop [5] is the open source implementation of the MapReduce programming model. Hadoop prevents effectively storing the computational capabilities over substantial packages of data. Generally, hadoop includes three layers such as a resource manager layer (YARN), a data storage layer (HDFS), and a data processing layer (Hadoop MapReduce Framework). HDFS refer as a block-oriented file system which depends on the effective data processing pattern that means a write-once, read-many-times pattern. YARN is a new framework that helps to writing a random distributed processing structures and applications.

YARN applications are not followed by MapReduce model, the original Apache Hadoop MapReduce and Hadoop YARN is try to take MapReduce for data-processing. The researcher's mostly selecting the research domain is Sentiment Analysis in Hadoop environment and collecting the data from the Social Media Network (SMN). For sentiment analysis data accuracy is the significant part and Apache Hadoop framework archieves an accurate results even if data generated from social media are abundant. The technologies like Hive, Apache Pig, HBase, Sqoop, Flume, Zookeeper, are integrated with Hadoop to boost the performance and efficiency of Hadoop.

III. SOCIAL MEDIA

Social media consists of mobile-based application and web-based application that allow the access ability, creation, maintain and exchange of user-generated data accessible at anytime [7]. Social media data is evidently most dynamic support base of human behavior, taking new prospects to identify with individuals, society and groups. At the time of writing, Facebook had 219.94 million monthly users (2018) whilst Twitter had more than 30.4 million tweets transmit each day (2018). Different categories of Social Media Data Different online social media generate different types of data. Several different categories of the data can be used such as image, text, video, audio, deleted non-posted content, mouse movement data, click data, etc. [8]. Social media data can be categorized into seven categories:

1) Demographic Data: It means that the publicly open and shared information, comprising race or ethnicity, age, education, gender, income, and geography

2) Product Data: product data are generated via social media users' mentioning of a specific product or brand on social media.

3) Psychographic Data: Psychographic data refer to data that can notify consumers' personality, attitude, values, lifestyle and interests related to a brand or a product.

4) Behavioral Data: behavioral data notify consumers' past buying behavior, like buying record, in social media platform

5) Referrals Data: Referrals data provide a clearer picture that supports companies or organizations to identify reason for sharing the information.

6) Location Data: location data analysis can be useful for resort and event management organization to better organize the business by targeting consumers based on their current geographical locations

7) Intention Data: It refers as data that can help companies and organizations to predict consumers' prospect with a product or a brand and future activities similar to them.

The term "SMA (Social Media Analytics)" has achieved a great deal of attention. SMA is defined as "an emerging interdisciplinary research field that aims on combining, extending, and adapting methods for analysis of social media data". SMA can be classified into several categories based on the objectives.

- **Topic Modeling:** Detecting dominant themes/topics by separating through large body of captured text. It aids to recognize latent themes/topics.
- **Opinion Mining**: This mining is related to sentiment analysis, but it focuses on the user views, judgment and trustfulness rather than referring positive or negative sentiment at first place. Opinion Analysis asses the user views, trustfulness which depends on the criteria for the purpose of analysis.
- **Trend analysis**: Prediction market strategy or customer activities using historical data. Market share, forecasting sales, customer growth or movements in stock market based on time series and regression analysis.
- **Popularity Prediction**: Popularity prediction techniques of gathering positive shares and negative feedbacks/ranks/opinion on certain events or subjects and to recognize the level of current popularity and forecast the future based on the current evidence. The prediction allows organization to forecast the future demand of services, and product.
- **Sentiment analysis**: Sentiment analysis is related to opinion mining however its refers to more in-depth interpretation of data of public/consumer/user sentiments, attitudes, evaluations, appraisals, and emotions towards entities like services, products, Individuals, organizations, events, issues, and their attributes. Sentiment analysis determines the individuals, communities, group, emotions towards any types of events, products, services, brand etc.
- **Customer engagement analysis:** This engagement analysis is processed to prolong the conversation or activities or events with social media users. Without proper incentives it is very hard to create engagement for long time, therefore proper incentives will understand the online consumer insights/behavior is significant. The purpose of the consumer engagement is to quantify the success of the online activities check whether it is a commercial campaign on non-profit activities. It aids organization to understand the current scenario and next action required to be successful in online environment.
- **Social Network Analysis**: Analysis of the social network that consists of individuals call nodes and connected with other nodes with similar opinion, interest, knowledge, etc. Data analysis technique contains frequency of edges, eigenvectors (i.e., page rank algorithm), and number of nodes. Social Network Analysis measures the types and depth of relationship between the networks. Many scholars analysis that considered the Social Network analysis as foundation of Social Media analytics.
- **Visual analytics**: Visual analytics is the most popular in the era of big data. It is an iterative process which involves information gathering, processing and decision making with the flow of execution in a respective manner. The purpose of visual analytics is to utilize graphical interfaces (e.g., dashboards) to present, explore and confirm relationships among variables.

IV. SENTIMENT ANALYSIS

The beneficial concept of sentiment analysis refers as the opinionated text for decision making based on its analysis. As an input for the opinion mining systems collects information from the opinionated text; the basic terms associated with the opinion mining. The main conception of sentiment analysis is parsing the text. The basic term of the sentiment analysis can be defined as detecting the polarity of the text (positive, negative or neutral). It refers to as opinion mining as it derives opinion of the user. Opinions differ from customer to customer and sentiment analysis greatly aids to understand users' perspective. There are mainly three types of opinions, known as direct opinion, indirect opinion and comparative opinion as given below.

Direct opinion: It refers to an opinion expressed directly on an entity or an entity aspect, for example, "The picture quality is great."

Indirect opinion: It expresses indirectly on an entity or of an entity aspect based on its effects on some other entities. This sub-type often arises in the medical domain. For instance, the sentence "After injection of the drug, my hand and knee joints felt worse" express an undesirable effect of the drug on "my joints", which indirectly provides a negative opinion or sentiment to the drug. In such scenario, the entity is the drug and the aspect is the effect on joints.

Comparative opinion: A comparative opinion shares a relation of differences or similarities between two or more entities and/or a preference of the opinion holder based on some shared aspects of the entities. For instance, the sentences, "Coke flavors better than Pepsi" and "Coke flavors the best" express two comparative opinions.

In general, sentiment analysis is performed mainly at three levels:

- **Document level:** In this document level, the way of express the customer opinion conveys a positive or negative sentiment about the particular product. For example, the overall positive or negative feedback determines a quality of the particular item. This habitation is called as document level sentiment classification. Hence, it cannot applicable to documents that evaluate or compare with multiple entities.
- **Sentence level:** The task perform at this level, the customers determines that each sentence expressed as positive opinion, negative opinion or neutral opinion. Neutral opinion means no opinion expressed by user. This analysis level is similar to subjectivity classification, which differentiates sentences (called objective sentences) and (called subjective sentences). Both the sentence representations express factual information from sentences and other express opinions and subjective views. To imply opinions refer as several objective sentences, however the subjectivity is not comparable with sentiment. For example, "we brought the car last month and the windshield wiper has fallen off."
- **Aspect level:** Using this approach handles both the document level and sentence level analyses. It does not reveal what exactly customer like or dislike however it makes customer realize the opinion about the product entity. Aspect level performs fine-grained analysis is also known as feature level. This approach directly expresses the sentiment analysis that includes positive, negative and neutral opinion. For instance, the sentence "The Smart phone call quality is good, but its battery life is short" clearly has a positive feedback for the first part, the second part represents a negative opinion of the product [8]. This can be analyzed in two aspects, how the aspect level is performed as qualitative analysis and quantitative analysis and both the sentence and document level classification is challenging task and the aspect level is difficult**.**

The sentiment analysis is a difficult process which carries the sequence of steps to evaluate sentiment data. These steps are: a) data collection, b) text preparation, c) data pre-processing, d) feature extraction, e) sentiment detection, f) sentiment classification, and g) presentation of output [9].

- **Data collection:** The beginning stage of sentiment analysis includes collection of data from user generated content contained in social networks, forums, product reviews and blogs. These data are mess up and expressed in several ways by using different slangs, vocabularies, context of writing etc. But, the manual analysis process is impossible to gather the review information from these sources. A specific web crawling mechanism is used to fetch the data and then store it in a database considering the format of data.
- **Text preparation:** After collecting data in a database, the review data requires to be extracted within a set of heterogeneous data fields. Non-textual contents and irrelevant contents for the analysis are recognized and removed.
- **Data Pre-Processing:** In the third stage of processing, different tasks consist of splitting sentences extracted into words, filtering of stop-words, tokenization, POS (Part-of-Speech) tagging, stemming, and the transformation to lower/upper cases are performed on the reviews in the pre-processing step to train them for the next step (i.e. feature extraction). A significant pre-processing task forms a part of sentiment analysis by assigning each word to a particular label in POS tagging (e.g., noun, verb, and adjective).
- **Feature extraction:** Feature extraction is called the process of obtaining a set of informative, discriminative, and non-redundant values to numerically signify a review sentence or text. The feature extraction techniques depend upon term occurrences, known as TF (term frequency) or TF-IDF (term frequency-inverse document frequency).
- **Sentiment detection:** This detection process is considered as a multiclass classification issue in which a text is classified into an appropriate topic class depending on its application and content. The extracted sentences of the reviews and opinions are analyzed. Sentences with objective communication (factual information) are discarded and Sentences with subjective expressions (opinions, beliefs and views) are retained.

- **Sentiment classification:** In this stage, the subjective sentences are separated such as negative, positive, good, bad; but classification can be made by using multiple points.
- **Presentation of output:** The main objective of sentiment analysis is converted into unstructured text format into meaningful data. When the analysis is completed, the text evaluation results are displayed on graphs such as bar chart, pie chart, and line graphs. Moreover, the time calculation can be examined and graphically displayed constructing a sentiment time line with the chosen value (frequency, percentages, and averages) over time.

V. SENTIMENT ANALYSIS TECHNIQUES

The sentiment analysis consists of two main techniques; machine learning based and lexicon based. These two techniques gain relatively better performance for research analysis. To perform the sentiment Analysis using three ways: 1) Machine learning technique, 2) Lexicon based Technique and 3) the combination of two approaches (Hybrid).

A. Machine Learning-Based Approach

This learning approach becomes an important task in many application areas. Machine learning accomplishes throughout recent years have magnificently created algorithms for handling volumes of data to unravel real world problems. Machine learning algorithms are grouped into supervised learning and unsupervised learning algorithms. Supervised learning algorithms will aid users train and learn from the training example which is tested and evaluated using the test data. The main disadvantage of supervised machine learning algorithms is the responsibility to create a training example. The training example should be comprehensive enough to make the algorithm effective and reliable enough to classify the instance in test data. Another type of machine learning is unsupervised learning algorithms. To identify the hidden associations in unlabeled data performs by the working mechanism of that algorithm. The unsupervised learning techniques are based on computing similarity differences between data. For instance, it analyzes k-means when similarity between data is evaluated based on proximity measures, like Euclidean distance. The following important steps are capable to construct the machine learning-based method. The initial step in Supervised Machine learning technique is used to collect the training set and then choose the appropriate classifier. Once the classifier is selected, the classifier provides trained using the collected training set. The key step in the Supervised Machine learning technique is feature selection. The classifier selection and feature selection concludes the classification performance. The most common techniques used for feature selection are:

- Opinion words and phrase: The most of the opinion words can be extracted from the document by considering adverb and adjectives, however sometimes verbs or nouns can also express opinion. For example, fantastic, amazing, good, bad and boring are all adjective or adverb that convey the emotions while rubbish is a noun but it express a sentiment similarly hate and like are verb but it express opinion. Once opinions are collected, its polarity can be calculated using lexicon based or statistical-based techniques.
- zTerms and frequency: n-grams or uni-grams with their frequency of occurrence are considered as features. Many research studies achieved better result. [10,11].
- POS tagging: To determine the feature that POS tag of words is used. It tags each word by regarding its position in the grammatical context [12].
- Negations: Negation word reverses the meaning, and it is very important feature in polarity calculation [13].

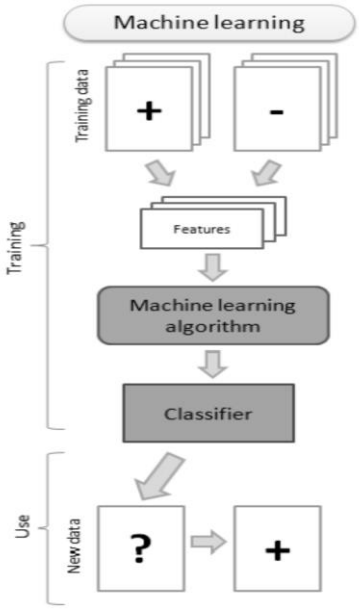

Figure 1: Machine learning approach to sentiment analysis

B. Lexicon-Based Classification

Lexicon-based approaches make lists of words manually labeled as having positive and negative polarity, and a polarity score for each word is generated. This constructed lexicon is utilized to evaluate the overall sentiment score of a given post or text. The notable benefit of the lexicon based technique does not require any training data (as the supervised machine-learning method does). The lexicon-based method is broadly utilized in conventional text such as forums, blogs reviews, [15], [16]. However, big data extracted from social media websites they are less likely to be utilized [16]. The key point is the unstructured format and nature of social media websites (the data includes informal and dynamic nature of language, textual peculiarities, abbreviations, new slang, and new expressions) [16]. As it performs at the word level, negated posts and posts with other meanings trick the lexicon polarity score measurement. Second, lexical dictionary and polarity scores are usually biased toward the text of a specific type, dictated by the linguistic corpora source [14]. Hence, it is challenging to build a generalized model regardless of the application domain. The key steps of lexicon based sentiment analysis are the following [17]:

- Preprocessing: The preprocessing method handles the document by eliminating noisy characters and HTML tags present in the document, by correcting grammar mistakes, spelling mistakes, and incorrect capitalization and punctuation errors and replace non-dictionary words like acronyms or abbreviations of general terms with their actual term.
- Feature Selection: POS tagging is used to extract the feature present in the document.
- Sentiment score calculation: For each extracted sentiment word initialize with Zero, check whether it is present in the sentiment dictionary, if it presents with negative polarity, w then s = s-w or else present with positive polarity, w then s = s + w.
- Sentiment Classification: If s is below a particular threshold value then classifying the document as negative otherwise classify it as positive.

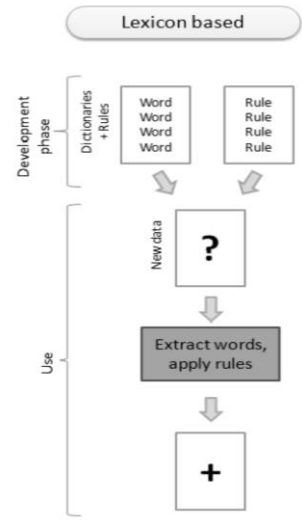

Figure 2: Lexicon approach to sentiment analysis

D. Hybrid approach

Hybrid method is the combination of both the lexicon based and machine learning approach. This approach has to prove the combination whether it is enhanced results or performance of classification. The main benefits of this approach are symbiosis stability uses of learning and lexicon algorithms for sentiment deduction [18].

VI. TOOLS FOR SENTIMENT ANALYSIS

Many research studies elaborate the methods and tools used for sentiments analysis.

- **Emoticons:** The most used tools for detecting the feelings polarity (positive and negative affect) of a message depends on the emoticons. Emoticons are face-based expression and happy feelings or symbolize sad expression, even though there are a wide range of non-facial variations. To extract the feelings polarity from emoticons, different set of ordinary emoticons can be used. Thus, emoticons have been frequently used for building a training dataset in supervised machine learning techniques [19].
- **Linguistic Inquiry and Word Count [20]**: It allows analyzing not only positive and negative but also emotional, structural and cognitive, components of a text based on the utilization of a dictionary containing words and their classified categories.
- **Happiness Index:** This Index [20] is a sentiment analyses property that utilizes the **ANEW** (Affective Norms for English Words) [21]. It provides scores for a given text between 1 and 9, point out the amount of happiness. The frequency is computed that each word from the ANEW emerges in the text format and then calculate a weighted average of the valence of the ANEW study words.

- **SentiStrength**: It is considered by [22] "the most popular stand-alone sentiment analysis tool". It utilizes a sentiment lexicon for allocating scores to positive and negative phrases in text.
- **NRC Hashtag Sentiment Lexicon**: A list of words (associations to positive and negative sentiments) is provided by the NRC Hashtag Sentiment Lexicon (version 0.1). The NRC Emotion Lexicon consists of frequent English adjectives, nouns, verbs, and adverbs explained as eight emotions (joy, anger, sadness, fear, surprise, disgust, anticipation and trust) in addition to negative and positive sentiments.
- **SailAil** Sentiment Analyzer (SASA): SASA considers an open source tool was calculated with 17,000 labeled tweets on the 2012 U.S. Elections [23]. It was evaluated also by the AMT (Amazon Mechanical Turk), where "turkers" were invited to label tweets as positive, negative, neutral, or undefined.
- **EWGA** and **FRN:** In EWGA and FRN tools are used [24,25]. The EWGA tool exploits an entropy-weighted genetic algorithm for performing an efficiently selection of features for sentiment classification utilizing a wrapper-model. While the FRN utilizes a feature relation network considering two syntactic n-gram relations: parallel relations and subsumption [26]. This tool uses classifiers built from machine learning algorithms. Unlike other tools that show aggregated numbers which makes it difficult to assess how accurate their classifiers are, this tool is able to classify individual tweets.
- **SenticNet**: SenticNet tool explores semantic Web techniques and machine intelligence. It utilizes NLP (Natural Language Processing) methods to infer the polarity of common sense concepts from natural language text at a semantic level, rather than at the syntactic level. SenticNet was tested to compute the level of polarity in opinions of patients concerning the National Health Service in England [19].
- **SentiWordNet**: The SentiWordNet tool is used as a lexical resource publicly available for supporting opinion mining applications and sentiment classification [27]. It depends on an English lexical dictionary known as WordNet [28] that ghaters adjectives, verbs, nouns, etc. into synonym sets known as synsets.
- **iFeel**: To develop a new sentiment analysis technique provides the competitive agreement and best coverage, which is performed by the combination of the different described approaches. It implements a public Web API, called as iFeel that affords comparative results among the different sentiment techniques for a given text.
- **PANAS-t**: This tool comprises an adapted version of the PANAS (Positive Affect Negative Affect Scale) [30], technique used in psychology. The PANAS-t tracks increases or decreases in sentiments over time; it depends on a large set of words associated with eleven moods: assurance, joviality, serenity, shyness, surprise, sadness, fear, hostility, guilt, attentiveness and fatigue. This technique computes the score for each sentiment for a given time period as values between [−1.0, 1.0] to indicate the change.

VII. APPLICATION AREA OF SENTIMENT ANALYSIS

Sentiment Analysis has been used and applied in several application and diversified areas such as the financial sector, politics, tourism, healthcare professionals, sports analysis, and consumer behavior, some of the emerging application areas are described in the following section.

- **Government:** Sentiment analysis aids government in assessing their power and weaknesses by analyzing public opinions. For instance, "If this is the state government, how do you expect truth to come out? The Member of Parliament (MP) who investigates 2g issue scam himself is intensely corrupt." [31]. this will clearly show the negative sentiment about government. Whether it should be tracking the public opinions on a new 108 system to take immediate action to spot out power and weaknesses in a recruitment campaign in government job, assessing achievement of electronic submission of tax returns, or several other areas, notice the potential for sentiment analysis.
- **Crime Analysis:** A preliminary work has been carried out in predicting crime using sentiment analysis techniques [32]. Also, it spacio temporal mining is to identify the crimes where happening in various fields. Linguistic analysis and statistical topic modeling is mainly used to automatically recognize across a major city in the United States, and then incorporated them in the crime prediction model.
- **Online Commerce:** The most common usage of sentiment analysis is provided in ecommerce proceedings. Websites allows their users to suggest their experience about product qualities. It can be useful for customers believe different features of the product by assigning and analyzing scores or ratings. Customers can effortlessly view recommendation and public opinions information on whole product and specific product features.
- **Business Intelligence:** In recent times, the business intelligence has been examined that people tend to look upon ratings and reviews of products that are available online before on the market sale. Many businesses, the online opinion decides the success or failure of their product. Accordingly, Sentiment Analysis plays a major role in businesses. Businesses also wish to extract sentiment from the online reviews to improve their products and in turn their reputation and help in customer satisfaction.
- **Disaster Recovery:** To analyze the current situation of the people during crisis period and disasters recovery has been carried out immediately action. Few of the actions analyses how could be the social media networking sites where utilized during disaster period. Such analyses are supportive to reach out people to solve the issue and help them. Voluntary organizations can render help to people who are in need. Some of the disasters that are analyzed are typhoons, earthquakes [33].
    - **Smart Homes:** In future, smart homes are supposed to be the technology of the future. In future entire homes would be networked and people would be able to control any part of the home using a smart phone or a tablet device. Recently, it has been several researches going on IoT (Internet of Things). Sentiment Analysis would also discover its way in IoT.

For instance, the emotion of the user or current sentiment, the home could change its ambiance to create a soothing and peaceful environment.

VIII. CHALLENGES IN SENTIMENT ANALYSIS

Sentiment analysis is an upcoming field that requires to be addressed to face many research challenges. Some of the open challenges are provided as follows.

- **Sarcastic sentences:** The positive feedback information which may have Sarcastic and ironic sentences. For example, "If the valuable feedback expresses about the particular product, on the next day it might be stopped working". In such case, positive words can have negative sense of meaning, so that can be difficult to identify the erroneous opinion mining of that.
- **Sentiment is Domain Specific:** The meaning of words modifies based on the context. For example, the phrase "go read the book" would be considered favorably in a book review, however if expressed in a movie review, it hints that the book is preferred over the movie, and thus have an opposite result [34].
- **Implicit sentiment and Sarcasm:** In case the Sentences might have an implicit sentiment devoid of the presence of any sentiment bearing words. For example, "How can you do this?" In this sentence, none of the words express negative feedback, but the meaning of the sentence is negative prospect. Therefore recognizing semantics is significant in semantic analysis.
- **Spam Detection:** To convey the views and aspects of anyone from any location, can express their views in social media analysis without disclosing their true individuality. Many fake reviews are written and spread in order to promote the sales of the product. Such an activity is known as opinion spamming. Apart from individuals, there are also commercial companies, this business spreading fake information. It is a challenging task to recognize such opinion spams to extract the exact sentiment.
- **Negation:** Negation handling is a complicated task in sentiment analysis as it reverses the polarity. In traditional text classification small differences between two pieces of text do not change the meaning. In Sentiment analysis, however, "the film was great" is different from "the film was not great". Negation also expresses by implicit sentences and that does not contain any negative words.
- **Conjunctions:** Presence of conjunctions in a sentence changes the entire meaning of the sentence. For instance, "The restaurant was very nice, but the service was poor". This sentence is divided into two parts. When it analyzes the first part to provide a positive sentiment. But the second part of the sentence reverses the entire meaning so that the must be considered for sentiment analysis.
- **Multiple Opinions in a Sentence:** Single sentence include multiple opinions with subjective and factual portions. For example, "The picture quality of this camera is amazing and also the battery life is good, but the view coverage area is too small for such a great camera", conveys both positive and negative opinions in same the same sentence [34].

IX.CONCLUSION

Huge collection of unstructured data that are accumulated on the web can be effectively analyzed and extracted by using Sentiment Analysis. Most of the business organizations consider that their business success, which depends on the satisfaction of the customers. So they encourage researchers and academicians have to be developing better solutions for Sentiment Analysis. Although some existing solution performs to find a better solution that overcome all the challenges that are being faced by Sentiment Analysis. In this paper, a brief survey presents on a variety of aspects of sentiment analysis. The sentiment analysis concept faces more challenges and issues using big data analytics for the further research.

X. REFERENCES

[1] Moraes R, Valiati JF, Neto WPG, "Document-level sentiment classification: An empirical comparison between SVM and ANN", Expert Systems with Applications, 2013; 40(2):621–33.
[2] Alexandre T, Alias F, "Sentence-based sentiment analysis for expressive text-to-speech. IEEE Transactions on Audio, Speech, and Language Processing", 2013; 21(2):223–33.
[3] Zheng-Jun Z, Yu J, Tang J, Wang M, Chua TS, "Product aspect ranking and its applications. IEEE Transactions on Knowledge and Data Engineering", 2014; 26(5):1211–24.
[4] Chun-Wei Tsai, Chin-Feng Lai, HanChieh Chao and Athanasios V. Vasilakos, "Big data analytics: A survey", 2015.
[5] "Apache Hadoop home page", http://hadoop.apache. org/, [Online; accessed February, 2016].
[6] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters", in Symposium on Operating System Design and Implementation, 2004, pp. 10-10.
[7] Kaplan, Andreas M, & Haenlein, Michael, " The challenges and opportunities of Social Media. Business horizons", Users of the world, unite 53(1), 59-68, (2010),.

[8] Batrinca, Bogdan, & Treleaven, PhilipC, " Social media analytics: a survey of techniques, tools and platforms", AI & SOCIETY, 30(1), 89-116. doi: 10.1007/s00146-014-0549-4, 2015.
[9] Hippner, H., & Rentzmann, R.,"Text mining." Ingolstadt: Springer-Verlag, (2006).
[10] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up: sentiment classification using machine learning techniques," Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, 2002, pp. 79-86.
[11] K. Dave, S. Lawrence, and D. M. Pennock,"Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," Proceedings of WWW, 2003, pp. 519–528.
[12] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach" , Journal of Informetrics, vol. 3, pp.143-157, 2009.
[13] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135
[14] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," IEEE Intell. Syst., vol. 28, no. 2, pp. 15–21, 2013.
[15] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Comput. Linguist., vol. 37, no. 2, pp. 267–307, 2011.
[16] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," ACM Comput. Surv., vol. 49, no. 2, p. 28, 2016.
[17] M. Annett, G. Kondrak, "A comparison of sentiment analysis techniques: Polarizing movie Blogs", In Canadian Conference on AI, pp. 25–35, 2008.
[18] 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.
[19] Tausczik, Y.R. & Pennebaker, J.W, "The psychological meaning of words: Liwc and computerized text analysis methods", Journal of Language and Social Psychology, 29(1):24-54, 2010.
[20] Dodds, P.S. & Danforth, C.M. 2009, " Measuring the happiness of large-scale written expression: songs, blogs, and presidents", Journal of Happiness Studies, 11(4):441- 456
[21] Bradley, M.M., Lang, P.J, "Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings", Technical Report C-1, Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.
[22] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas A, "Sentiment strength detection in short informal text", Journal of the American Society for Information Science and Technology, 61(12), pp. 2544- 2558, 2010
[23] Wang, C.J., Tsai, M.F., Liu, T., Chang, C.T, "Financial Sentiment Analysis for Risk Prediction", In Proceedings of the Sixth International Joint Conference on Natural Language Processing pp. 802-808, 2013
[24] Mayfield, E., & Rosé, C.P, "Light SIDE: Open Source Machine Learning for Text Accessible to Non- Experts. In the Handbook of Automated Essay Grading, Routledge Academic Press , 2012
[25] Hassan, A., Abbasi, A., Zeng, "Twitter Sentiment Analysis: A Bootstrap Ensemble Framework", Proceedings of the ASE/IEEE International Conference on Social Computing, pp. 357-364, 2013
[26] Abbasi , A, "Intelligent Feature Selection for Opinion Classification", IEEE Intelligent Systems, 25(4), pp. 75-79, 2010
[27] Esuli, A., & Sebastiani, F, "Sentiwordnet: A publicly available lexical resource for opinion mining", In Proceedings of LREC Vol. 6, pp. 417-422, 2006
[28] Miller, G.A, "WordNet: a lexical database for English", Communications of the ACM, 38(11), 39-41, 1995.
[29] Gonçalves, P., Benevenuto, F., Cha, M. , "Panas-t: A pychometric scale for measuring sentiments on twitter", arXiv preprint arXiv:1308.1857, 2013.
[30] Watson, D. & Clark, L, "Development and validation of brief measures of positive and negative affect: the panas scales", Journal of Personality and Social Psychology, 54(1):1063–1070, 1985.
[31] http://proudtobeindian.net/indian-media-exposed
[32] Xiaofeng W, Gerber MS, Brown DE, "Automatic crime prediction using events extracted from twitter posts", Social Computing, Behavioral-Cultural Modeling and Prediction, LNCS 7227, Springer Berlin Heidelberg; 2012. p. 231–8.
[33] Caragea C, Squicciarini A, Stehle S, Kishore N, Tapia A., "Mapping moods: Geo-mapped sentiment analysis during hurricane Sandy", Proceedings of the 11th International ISCRAM Conference – University Park, Pennsylvania, USA; 2014 May. p. 642–51.
[34] Pang, B and Lee L, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieva,l 2008, (1-2),1–135.