Recent Developments and Reviews in Sentimental Analysis for Big Data

Dr. B. R. Prakash[1], Dr. D. Ramesh[2]

[1]*Assistant Professor, Dept. of Master of Computer Applications, Sri Siddhartha Institute of Technology, Tumkur, INDIA*
[2]*Professor and Head, Dept. of Master of Computer Applications, Sri Siddhartha Institute of Technology, Tumkur, INDIA*

***Abstract*- with the increasing need for understanding customer behavior and need for better buyer-seller relationships more than ever sentiment analysis has become one of the major tool in today's time. The growing data and the need for faster computation efficient and more reliable processes of SA (sentiment analysis) are preferred and are in great demand.SA as a field of science has grown a lot from its earlier days. With the advent of big data practices, this paper focuses on processes followed in performing SA on big data and how big data tools and frameworks go along with sentiment analysis and it also highlights the gaps and suggests future works that should be explored.SA studies need to be expanded into providing better scalability and velocity along with reliability.**

I. INTRODUCTION

The growing need for the powerful computational means and high speed analysis big data tools and frameworks gained recognition. Big data also enables companies to collect and analyze diverse and unstructured data which in past was mostly ignored. People realized the potential that different sources of data hold. Sentiment analysis (SA) which is basically the study of patterns and relationships that emerge from the data sources. This analysis helps businesses make informed predictions and access the people's response, attitude or emotion towards a particular product, campaign, service, agenda and ideas over the internet in the form of text, audio or video. The resultant output can be divided into three categories positive, negative or neutral. These categories comprise of many names and slightly varied tasks, such as subjectivity analysis, opinion mining, opinion extraction, sentiment mining, affect analysis, customer complaint, emotion analysis, review analysis and review mining. Many techniques of sentiment analysis are developed which can be characterized into: Application based, which consists of stock market predictions, supply chain assistance ,business strategy analysis etc. ; fundamental approaches, including word-level sentiment disambiguation, sentence-level Sentiment Analysis, aspect-level Sentiment Analysis, bigram trigram and multigram SA, concept-level SA, multilingual SA and linguistic features analysis; and social aspect, which exploits the online content generation to analyze such inputs as epidemic spreading, emotion and responses towards events, campaigns and products. However, this article aims at understanding the synergy between sentiment analysis and big data and the issues faced in SA by big data framework and tools. In SA the focus is mainly on content analysis andfinding patterns and meaningful information and in big data the idea of velocity, storability, variety and usability are stressed upon. Several papers have mentioned that SA on big data is related to the speed and volume drawback, however a study that reviews the relation between big data problems and SA is unavailable. Existing review-based studies on SA are centred on techniques, applications and net services, however none have centred on the ability of SA approaches in big data. This paper addresses this drawback and reviews whether or not the SA techniques, that have been introduced before big data was made standard, are appropriate, economical and effective for big data infrastructure. The main contribution of this paper lies in distinguishing challenges and making suggestions to resolve the gaps. This paper is organised as follows: the first part briefly introduces SA and its relevancy big data. The second half introduces the overall problems associated with big data. The third part details the approaches of SA, whereas the fourth part describes the future opportunities to resolve the problems of SA in big data. The fifth part contains the conclusion.

II. PROBLEMS FACED IN SENTIMENT ANALYSIS ON BIG DATA

Sentiment analysis is the main focus of big data, but still a considerably less amount of studies have been conducted to understand the suitability of Sentiment analysis for big data frameworks. This section focuses on this aspect by starting discussing the general scenario and challenges of big data analysis, followed by an exposition about the general SA framework. The volume and amount of data are main features of big data. Another important feature is the speed of computation which is essential when dealing with real time streaming data. Uncertainty of result and accuracy also needs addressing. The above mentioned facts also bring in new kinds of data analysis processes along with new storage mechanisms. Big data analysis is a continuous process having many steps and not an isolated one, which involves mining knowledge from data which is varied and in its raw form. It lacks a data model to define meaning to each element and in the right context. Useful information is usually hidden deep into the data and is difficult to unearth. Furthermore, the

actual relationship between the huge data set is also unknown. So an iterative approach is to be followed in order to reach a conclusive and useful result. Many a times you could go on with a series of steps and end up with no worthy knowledge and reach a dead end. One constant issue with big data analysis is the need for a flexible capacity which can deal with different amount of data. Although cloud computing concepts are widely in use, the iterative characteristic of analysis leads to problems and utilises varied amount of storage capacity. The results generated are also not always straightforward positive or negative they are somewhere in between. Identifying and predicting the relationships between different data set is an exhaustive process. To minimize the cost and effort and make the process efficient different decision management techniques are also being considered nowadays.

III. FRAMEWORK OF BIG DATA

Sentiment analysis focuses on the opinion of the people. The approach followed can be specific to the content or unspecified. Opinion mining was introduced earlier than SA but because of the industrial and commercial usefulness of sentiment analysis it has grown in stature. In the digital world of today the producers and manufactures need to know the sentiments and opinion of the people to gain higher profits and remain relevant in the huge market in which today, people are not short of choices. Enterprises often indulge in marketing advocacy to generate positive review about their product and also listen to grievances of people and consider the reviews when taking decisions about the new products. In this aspect, the focus is on the sentiment orientation of the data or a message online and the reason for the popularity of messages of a particular theme. Over the years the research in the field have grown but not in context with big data. Several platforms provide sentiment analysis on big data because of its Applications in social media marketing and monitoring. The data which is collected should first of all be relevant to the user's objectives and is directed at specific hashtags and message containing important keywords which are known in advance. Hadoop plays an important role in data preprocessing and helps to identify the trends and information which sometimes gets unnoticed like the out of scope values. Enterprises generally use Hadoop for initial steps and later on perform advanced data analysis and mining. It is used for filtering, reduction, sorting and arranging data so that denser data is obtained which contains more information and knowledge can be generated from it. The aim of Pre-processing is to provide filtered data and make a data warehouse from dataset which contains customised and data which is relevant for analysis. All the irrelevant data is filtered out.

Table 1: Big Data tools and their uses

Structured query language	It allows data to be inserted, manipulated and transformed into the Hadoop file system.
HIVE	A data warehouse software which allows data summarization and is based upon apache Hadoop. It Produces HQL from SQL.
Hadoop Distributed File System (HDFS)	Provides high performance data storage capacity for Hadoop and is widely used in big data analytics.
Map Reduce	This framework is used for parallel processing on different nodes and basically involves distribution and aggregation after processing.
Flume	Used in deriving data directly from social media websites like twitter and Facebook. Sometimes restrictions are there so aggregation tools are used.
Jaql	It is used to simplify high end queries into simpler ones and aids parallel processing used in map reduce task.
HBase	It uses non SQL technique and provides compression.It follows column based approach and sits on top ofHDFS.
Zookeeper	It maintains parallel processing on big sized clusters and provides various central services like synchronisation.
PIG	PIG programming language contains the language named as PIG Latin and is used to assimilate structured and unstructured data.

Table 2: Big Data tools for analysis

Tools	Description
Rapid Miner	Provides a single platform for machine learning, business analytics and maintains high value data science and delivery.
Google analytics	Free analytics service which can be used by the businesses to analyse website traffic and is provided by google
MongoDB	It is a new and dynamic database management tool which does not use SQL and takes help of JSON type documents.
Tableau .	It is a tool for formatting and visualisation of data and is used with Hadoop hive. It optimises the queries and memory cache
Python	It is a high level new generation language used in data analysis and many different fields. It reduces the complexity and makes the steps easier.
Alpine data labs	an advanced visualisation and ollaboration tool used with apache Hadoop and big data analytics algorithms to enable the use of different models

MapReduce is a 2 step process: Mapping, which involves transforming unstructured data and deriving something meaningful from it; reducing, which involves taking the data received from step one and aggregating the result for reporting. MapReduce concepts are used by many social networking sites like Facebook and twitter for providing features like who visited your profile on LinkedIn and who read your posts on Facebook. After generation of information or meaningful data stored in Hadoop, it is provided to the business intelligence platforms or analyses using tools like PowerPivot. Many companies use SQL as their Business intelligence platform. Many options are available to analyse and access data such as Polybase and Sqoop.

Amazon Web services are used my many companies like Expedia, for Expedia amazon EMR is used. The data that

arrives is generally user interactions and supply chain data. This arriving data is collected in amazon S3. Everyday over 200 requests are processed on request by Expedia using AWS. The advantage of using AWS is the data can be scaled and manipulated to meet load demands. Amazon EMR cluster is created which is later on used for sentiment analysis in NLTK program. The whole process is implemented in Hadoop framework and the output is used to analyse the sentiment of data.

Oracle has developed OAA which has extended oracle database to provide advanced analytical solutions using open source R language algorithms for opinion mining, text mining and numerical evaluations. This way the scale of data analysis is increased as the volume of data increases by bringing analytical algorithms close to the data storage site.

Main customers of sentiment analysis on big data platforms are enterprises which need to know about their brand value in the market. The process starts with a research on social media data about the responses and acceptability of the brand on digital platforms. After this the marketing and promotion campaigns are organised based on the research and information obtained. Different SA tools are used to monitor social media. These tools either check mentions or check the whole content. The trend analysis of twitter does not follow any deep digging into data but just follows the mentions and hashtags to provide the trending hashtags. Content based analysis which involves detailed examination of the data is a costly process because of its interactive and multilingual properties. Some of the tools are Hootsuite, Brand watch, Sprout social, NUVi, adobe social etc. Their properties range from in-depth competitor analysis, Reports customisation, Social traffic and source analysis etc. All these applications and tools have been developed using social media data. However, the true power of big data tools has still not been put to use. The traditional processes included gathering information from digital sources and applying opinion mining algorithms to calculate sentiment score based on the subjective data. Recently tools like HIVE and FLUME are used for gathering and storing data. Basically sentiment analysis applications like Horton works are used and power View is used for customised visualisation. Although, because it does not have varied capability due to the use of only python.

IV. FEATURES OF SENTIMENT ANALYSIS AND APPROACHES

Sentiment analysis which is focus only on finding the polarity (positive, negative or neutral). In content analysis many of the traditional kinds of analysis may not be relevant now. Therefore, it is now a days used mainly on internet friendly forms. There are several tasks which are carried out during the whole process like, finding the polarity and the orientation of the text i.e. positive or negative, finding out the classification range and depth and also identifying the target or the source. Problems that arise are mainly based on the classification front. Some researchers tried to classify emotions like sadness, excitement, rudeness, happiness, horror and not the sentiment while some tried to classify messages based on their context and factual correctness. Different approaches are followed based on the requirement and the costs. All approaches are mentioned below:

- Document level, whole document is considered as a whole and total resultant polarity or sentiment of the article is displayed.
- Sentence level, each sentence is considered individually and polarity is calculated for each sentence.
- Phrase level, if in each sentence you have phrases with different orientation or meaning than this approach is followed for effective analysis. The machine learning and sentiment analysis are detailed in the following sections.

V. FEATURES OF SENTIMENT ANALYSIS

Four explicit features are known:

Feature	Description
Semantic	In this the semantic orientation of the word is considered and polarity is derived.
syntactic	Most common feature contains n-grams and POS tags along with punctuations. They follow simple rules like noun followed by positive word will denote positive sentiment and vice versa.
Link based	It is used to derive information from the digital platforms and social media documents. This is a rarely used feature so effectiveness in unknown.
Stylistic	It contains lexicon attributes. Structured style markers and not used in much of the studies present. Radical criteria have been used like classification in terms of length of reviews for a product.

Implicitly, the research carried out has focused on sentiment orientation (SO) and lexicon rules to find the meaningful information in the data by using a dictionary of keywords along with their sentiment scores. The semantic approaches tries to provide intensity to the polarity result. This approach is a modern approach where a data repository containing keywords along with their sentiment score which are manually given and coded. Semi

auto-generated tools or lexicon tags have also been used which provide information such as weak or strong subjective words. Other features used include bag of words which is used to tell if a document is subjective or objective. Semantic attributes also contain contextual features (Lexicon based approach) and represent the SO of text in its vicinity. This type of approach is useful in sentence level semantic analysis and also to find subjectivity or objectivity of a sentence.

CELEX database

This database contains ASCII versions of English, Dutch and German languages. The databases have not been customised for any particular DBMS. The information is in ASCII files in a UNIX directory that can be fetched using various defined tools like, AWK or ICON. The files are identified uniquely by the serial code given to them and helps in linking of information stored in different files. AWK functions have been given so that when the need arises for online computation, it can be easily carried out.

WordNet

It was developed in 1986 at Princeton University. It is a large lexicon database which is electronic and updates and developments are constant and are carried out on a regular basis. WordNet consists of synsets which contains common categories like nouns, verbs, adjectives etc. The current version of WordNet contains around 117,000 synsets, comprising over 81,000 noun synsets, 3,600 verb synsets, 19,000 adjective synsets and 3,600 adverb synsets. WordNet has been used for finding similar words and synonyms of the words, whereas SentiWordNet, an advanced data repository has been used to identify the SO of sentences and information which is derived.

CSLI database

The database named verb semantic ontology is developed by centre for the study of language and information, Stanford. It aims at ontology and interoperability of verbs. Its focus is joining the orientation to programming languages to modern standards, in particular, the ISO Common Logic standard.

VI. ROLE OF MACHINE LEARNING IN SA

Machine learning uses machine learning algorithm on data and trains the sample data to produce models then the data which is to be classified is compared with the trained dataset and classification is done based on the trained data values and trends. The big data framework contains various plugins to enable easy application of machine learning algorithms for sentiment analysis. The examples are Pentaho and Mahout. The classification methods in machine learning can be divided into 3 types:

- Supervised learning, it is also called labeled data training method and the class labels are already known to us before classification.
- Unsupervised, it is known as classless approach and the class labels are not known before hand and clusters are formed based on the algorithm output.
- Semi-supervised, it basically contains large amount of unlabelled data but a small amount of labelled data to give direction to the classification. Supervised learning methods are mostly effective and accuracy is also high but it is mostly subjective data. In unsupervised data is not contextual (no labels) which hinders the efficiency. The demand for unsupervised learning methods is high because in real world scenarios the data is mostly vibrant and non-labelled. This was the reason for introduction of semi-supervised learning.

There are a lot of machine learning algorithms which are adopted to perform classification task in SA, to name a few SVM, naive Bayes, ID3, k-NN, maximum entropy. Naïve Bayes It is based on Bayes theorem which is a probabilistic theorem used for classification. Its simplicity is the reason for its widespread use and is mostly used in document classification jobs.it can be improved when used with other methods like Senti-lexicon. For a small dataset it is better than other algorithms but becomes time consuming with larger datasets and effectiveness decreases.

K-NN

It is one of the lazy learning algorithms and process moves forward only when classification is complete for computation. It is the simplest one to understand as only clusters are formed containing values which are nearby each other. It is important to assign a value to each mapping during regression and classification as near neighbours should contribute more to the total average of the cluster than thedistant data values. It is sensitive to structure of the data.

ID3

In this algorithm decision trees are formed from the dataset and this algorithm output is further used in C4.5 algorithm. It moves forward by entering through the root node and finding unused attributes.At every steps it goes by calculating information gain and entropy. Iteration. Recursion can stop when, every subset belongs to the same class, no more attributes remain to be selected or the end of subset.

VII. GAPS AND FUTURE SCOPE FOR SENTIMENT ANALYSIS IN BIG DATA

The use of big data analytics in combination with sentiment analysis is widely in use in industries so it is very important to identify gaps in implementing SA on big data framework. The enterprises utilise these concepts and techniques in day to day analytical tasks. The sentiment analysis techniques came into existence long before the big data framework emerged so most of the studies are research on SA are limited to text mining, content mining etc. Big data is not review in association with SA which is rather surprising because the main aim of big data concepts has been performing sentiment analysis. The reviews currently available on SA have been mostly related to web services, techniques etc. There is a need to understand the adaptability of sentiment analysis in big data.

The typical approached followed in sentiment analysis in big data era should be reviewed and modified for efficiently. Some literatures started exploring the issues faced in big data

while performing SA, like distributed parallel processing , scalability issue is also addressed along with few reviews dealing with new data tools for analysis and improvements in machine learning models. Most of this research can be found to be around the year 2014, which indicates the start of a new era in the world of computation. When it comes to the case of volume or amount of data, the SA techniques are designed to work with flexible volumes big or small in fact, when data sets are huge the accuracy should improve as in case of supervised learning as we will have a large amount of data for training. Scalability can be studied in depth when NB classified is calculated for sentiment analysis of scalable type rather than using standard external libraries. Altogether, Scalability is not a big issue in SA on big data. Velocity and variety are possibly the most relevant issue in sentiment analysis as most of the data is real-time data with streams of data continuously flowing in from digital platforms like review sites, social media and blogs. It gets closely related with the volume features of big data. New linguistic features are created almost every week in today's time like new emoji and phrases and slangs, which requires regular updates on the models. Also people on social media tend to make a lot of grammatical mistakes and use short forms which hinders the effectiveness of sentiment analysis. So a new and improved model has to be developed which can deal with new features and erroneous words. A fuzzy logic based approach is in existence for mining on twitter data. In this approach the sentiment of text is found using the attributes given to the words beforehand. To reduce the time period of processing the MapReduce feature was introduced which distributes the work to various nodes so that the process quickens up. A method based on deep learning has also been introduced for dealing with unsupervised data effectively. A method called Hierarchical Dirichlet Process-Latent Dirichlet Allocation also known as HDP-LDA is used for unsupervised data. Statistical weightage is shared between different clusters and a Dirichlet approach is used. This approach has been hugely successful in finding the correct meaning of the words and text. Although, it has only been used in experiments and real life examples are few. So a less developed and stagnant approach which requires modification and one temporary ad-hoc approach will not help in improving the data mining on social platforms. Also the requirement for new models and improved mechanism to deal with misspelt words and grammatically incorrect sentences. In the future, a model which takes use offeatures of various other models will be required for computation and helping out the enterprises. The data gathering process also needs to get more relevant and trustworthy. This means that the data gathered should be verified and the activities of the person posting it should also be considered. It may be the case that the user is a fake user and posts unreliable content with no particular sentiment to it. This is where filtering of data comes in. The analysts and researchers think that applying SA to big data is in hit and trial approach with many iterations until you get the desired results. But the issues mentioned in this section cannot be neglected in order to increase the accuracy and ensure high speed of computation for the industrial use.

VIII. CONCLUSION

With the advent of sentiment analysis in past 10 to 15 years, the industries and enterprises have utilized all its properties and methods in order to gain benefits. The advancements in the field of storage, computational speed and introduction of big data frameworks. Big data was developed as an individual data management mechanism and deals with all the analytical processes in the same way. It is just there to help out various techniques with its high value features. Although research is available to study issues in big data when performing sentiment analysis, the features like volatility and variability are still to be explored and is acceptable considering the fact that big data is a new concept. This paper deals with adaptability of sentiment analysis to big data framework. The effectiveness and extent of big data in customer-producer relationship management is still not known although the system is widely in use. Most of the companies nowadays have realised that sentiment analysis is the way to go forward in future in order to stay relevant in the market and deal with market problems logically and effectively.

REFERENCES

[1]. Lexicon-Based Methods for Sentiment Analysis-Maite Taboada, Julian Brooke, Milan Tofiloski,Simon Fraser Voll, Manfred Stede, -Computational Linguistics June 2011, Vol. 37, No. 2, Pages: 267-307 Posted Online May 26, 2011

[2]. Sentiment user profile analysis based on forgetting curve in mobile environments- Sang-Min Park ; Doo-Kwon Baik ,Young-Gab Kim22-23 Aug. 2016 IEEE

[3]. Sentiment analysis on bangla and romanized bangla text using deep recurrent models-Asif Hassan ; Mohammad Rashedul Amin ; Abul Kalam Al Azad ; Nabeel Mohammed 12-13 Dec. 2016 IEEE conference

[4]. Sentiment analysis of Twitter data: Case study on digital India-Prerna Mishra ; Ranjana Rajnish ; Pankaj Kumar6-7 Oct. 2016 IEEE conference

[5]. Analyzing Sentiments in One Go: A Supervised Joint Topic Modeling Approach- Zhen Hai; Gao Cong; Kuiyu Chang ; Peng Cheng ; Chunyan Miao 14 February 2017 IEEE

[6]. Effect of negation in sentiment analysis- Wareesa Sharif ; Noor Azah Samsudin ; Mustafa Mat Deris ; Rashid Naseem24-26 Aug. 2016 INTECH conference

[7]. An analysis of ICON aircraft log through sentiment analysis using SVM and Naïve Bayes classification-Harshit Sinha ; Rahul Bagga ; Gaurav Raj 6-7 Oct. 2016 InCITe

[8]. A proposed method for predicting US presidential election by analyzing sentiment in social media-Andy Januar Wicaksono ; Suyoto ; Pranowo-ICSITech 26-27 Oct. 2016

[9]. Rudy Prabowo, Mike Thelwall, "Sentiment analysis: A combined approach .", Journal of Informetrics 3 (2009) 143– 157 Gamgarn Somprasertsri, Pattarachai Lalitrojwong , Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization, Journal of Universal Computer Science, vol. 16, no. 6 (2010),938-955.

[10]. Qingliang Miao, Qiudan Li, Ruwei Dai , "AMAZING: A sentiment mining and retrieval system", Expert Systems with Applications 36 (2009) 7192–7198.

[11]. Long-Sheng Chen , Cheng-Hsiang Liu, Hui-Ju Chiu , "A neural network based approach for sentiment classification in the blogosphere", Journal of Informetrics 5 (2011) 313– 323
[12]. Kaiquan Xu , Stephen Shaoyi Liao , Jiexun Li, Yuxia Song, "Mining comparative opinions from customer reviews for Competitive Intelligence", Decision Support Systems 50 (2011) 743–754
[13]. Popescu, A. M., Etzioni, O.: Extracting Product Features and Opinions from Reviews, In Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, 2005, 339–346.
[14]. Gang Li , Fei Liu , "A Clustering-based Approach on Sentiment Analysis" ,2010, 978-1-4244-6793-8/10 ©2010 IEEE
[15]. ZHU Jian , XU Chen, WANG Han-shi, "" Sentiment classification using the theory of ANNs", The Journal of China Universities ofPosts and Telecommunications, July 2010, 17(Suppl.): 58–62
[16]. Melville, Wojciech Gryc, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification",
[17]. KDD¨09, June 28–July 1, 2009, Paris, France.Copyright 2009 ACM 978-1-60558- 495-9/09/06
[18]. Asur, S. and B.A. Huberman, 2010. Predicting the future with social media. Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, Aug. 31 2010-Sept. 3, IEEE
[19]. Cheong, M. and V.C.S. Lee, 2010. A micro bloggingbased approach to terrorisminformatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. Inform. Syst. Frontiers, 13: 45-59.