RESEARCH ARTICLE OPEN ACCESS

Sentiment Analysis of Unstructured Data, Applications - A Survey

Rekha Sunny T
Assistant Professor
SCMS School of Technology and Management
Cochin
Kerala - India

ABSTRACT
The extensive popularity of Internet resulted in the proliferation of textual data particularly in an unstructured form. More than 80% of data generated through blogs, tweets, customer reviews, emails etc. are unstructured. The rapid increase of unstructured data makes the analysis and decision making more and more challenging. Sentiment analysis is one of the efficient techniques adopted to mine the emotions, opinions and attitudes from various data sources. This paper presents a survey on the analysis of unstructured data, sentiment analysis, different approaches and its diverse applications.

Keywords **:-** Big Data, Data Analysis, Unstructured Data, Sentiment Analysis

I. INTRODUCTION

The concept of big data has grabbed peoples' attention since the time it was introduced and continues to be of high value in various aspects. TechAmerica Foundation defines big data as "a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information" [1].

The big data is a collection of heterogeneous mixture of data containing structured, semi-structured and unstructured data. Over 90% percentage of data was generated over last 3 years. A leading industrialist Merrill Lynch has published, out of which "80% of business-related information originates in unstructured form basically text" [2]. Recent study reveals that 85% of the Fortune 500 organization may fall behind the other 15% of the organization because they failed on leveraging the information by exploiting the unstructured data [3].

Data analytics helps us to find the hidden information in unstructured data through the computational process of Knowledge Discovery on Database (KDD). But firstly, the unstructured data has to be refined from its crude form to more structural one which is done through various analytical procedures. Various challenges are faced while combining the structured and unstructured data. The resultant data will be helpful in the predictive analytics [4]. Sentiment analysis- which is the automated mining of attitudes, opinions, and emotions from text, speech, and database sources through Natural Language Processing is one of the efficient techniques to retrieve useful information. Opinions in text can be classified into categories like "positive" or "negative" or "neutral" through Sentiment analysis. Sentiment analysis is also referred to as subjectivity analysis, opinion mining, and appraisal extraction [17].

The objective of this paper is to introduce different approaches to analyse unstructured data, especially the concept of sentiment analysis and its wide range of applications. The paper is organized as follows.

Section 2 describes about unstructured data, text analysis, audio analysis and video analysis. Sentiment analysis and different granularity levels are explained in section3. The related works in sentiment analysis are briefed in section 4. Various applications of sentiment analysis are described in section 5 and section 6 concludes the paper.

II. UNSTRUCTURED DATA

Data analysis involves data inspection along with data cleaning and applying transformation technique. Later the data is modelled to recover some valid knowledge that helps in decision making. Knowledge discovery methods such as data mining, business intelligence which deals with business data in decision making, explorative Data Analysis which is used to find new features of data, Confirmatory Data Analysis which is used to check the validity of hypothesis, Predictive analytics used to forecast some event or scenarios are all part of Data Analysis.

Most of the available data, which is in unstructured form have to be cleaned and modelled into an acceptable form from which the knowledge can be discovered. Therefore, unstructured data has to go a hefty compilation and time-consuming process which is really expensive. Hence all strata of business or any other fields handling immense amount of data find it beneficial for them to find least expensive methods to analyze the unstructured data.

There are various techniques to analyze unstructured data according to the type content that the data has such as text analysis, audio analysis, video analysis and sentiment analysis.

A. TEXT ANALYSIS

In text analysis, the information or insights are extracted from various sources of text data such as feeds, documents, emails, advertisements, blogs, news content, logs, website contents, social media content etc. Techniques such as statistical analysis, Computational linguistics and machine learning are made use for the same. These meaningful insights benefit the organizations in better decision-making processes.

Information Extraction or IE techniques are used to retrieve structured data from the unstructured data. It particularly involves two subtasks which are 'Entity Recognition' and 'Relation Extraction' [5]. In the process of entity recognition various entities such as person, organizations and things are found out and are classified into separate classes. In the relation extraction phase of information extraction various semantic relations that lies between the classified entities are found out. For example, from a medical prescription various entity such as the person, the hospital or organization, drug dosage information can be found out and classified in entity recognition phase. While the relation extraction leaves us with a clear idea of the various relations between the entities that are found out.

Another method of text analysis is 'Summarization'. Summarization involves the techniques that creates a summary from the text content of the information provided. The summarization can be of two types which are 'Extractive' or 'Abstractive' [6]. The extractive summarization involves creating a summary of information by adopting the actual original units from the original content. The abstractive summarization involves the technique of analyzing the contents of information and extracting the summary by learning the semantic relations in data. Hence the abstractive method demands the use of s Natural Language Processing (NLP) to parse the text and produce summary. Abstractive summary may need not include the original units from the contents. However for big data, extractive summarization is better since it is easier to adopt.

Another way of text analysis involve the Question and Answer method. It is mostly adopted in academic and healthcare sectors. There are basically three types of QA methods, they are Information Retrieval (IR), Knowledge-based approach and Hybrid approach. In information retrieval method the question and answer way of analytics are done in three phases. First phase consists of question processing from which a query is made. In second phase the pre-written documents are analyzed and the relevant information is extracted. The third phase involves the answer processing where the answer is matched to the previously extracted information and are ranked. The answer which ranks top are given out as solution. Knowledge-based approach is widely accepted in fields where absence of large volume of pre-written contents are present. It can be effectively adopted in the restricted domains such as medicine or tourism. This approach make use of semantic information of the question and then it is used for querying. In Hybrid approach, both IR and Knowledge-based approaches are used. The querying is based on knowledge based approach and the answers are taken by information retrieval approach.

B. AUDIO ANALYSIS

Audio analysis deals with analyzing, extracting or retrieving data from unstructured audio content. When applied to human speech, the audio analytics can be called as speech analytics [8, 9]. This kind of information is very useful in fields which consists of lot of information in the form of audio. This method can be successfully implemented in fields such as call centers and healthcare sectors. Audio analytics can even be applied in analyzing the emotional conditions of a person [8].

There are basically two approaches in audio analysis which are transcript based and phonetic based. Transcript based analytics can also be called as Large Vocabulary Continuous Speech Recognition (LVCSR) which is further divided into two sub process: Indexing and Searching. In indexing phase, Audio Speech Recognition (ASR) algorithms are used to match sounds with words in pre-defined words in dictionary. If failed to do so, the most similar word is returned. The output file contains sequences of words which were identified during the analysis. In second stage, simple text based operations are applied to search terms. In phonetic-based analysiss, the analysis is done based on phonemes instead of words. It also involves two phases which are phonemic indexing and phonemic searching.

C. *VIDEO ANALYSIS*

Process of monitoring, analyzing and gaining meaningful insights from video streams also called as Video Content Analysis [10]. This type of analysis can be used in both real-time as well as pre-recorded videos. These techniques are still at its initial stages. The main source of information is video streaming sites such as YouTube and are Closed Circuit TVs (CCTVs). There can be two approach while parsing the information which are Server-based and Edge based [11]. In Server based approach the information captured are sent to a central server at which further analysis of the data occur. The server based approach will be more efficient if the network bandwidth is higher and server processing capacity is higher. If the bandwidth of network low, the data which has to be sent will meet a need to be compressed hence making analysis of data less reliable. The edge-based approach involves analysis of data locally hence reducing the use of bandwidth requirement since no chance of data loss. However the edge-based approach may have lesser processing capacity and costs more due to the fact that the processing is done locally. Implementing video analysis can become very crucial in near future. For example, a store CCTV can be used to monitor the customer and their purchase pattern, items bought together, pattern of searching items…etc.

III. SENTIMENT ANALYSIS

Sentiment analysis or Opinion analysis consists of natural language processing, computational processing, text mining and biometrics to identify, retrieve , visualize and to learn various affective states and information on the emphasized subject. Sentiment analysis can be applied to review the voice of the customers, responses to surveys, social medias, costumer services and even targeted marketing[17].

In other words, sentiment analysis involves identifying the attitude of a writer, speaker or any other actors on any topic or even the overall polarity of the context or sentiment in the context. The view of the actor can be evaluation, judgement or intended emotional communication or even the affected state of actor.

Sentiment analysis can be categorized into three types based on the level - document level, sentence level and feature/aspect based as shown in the figure [1].

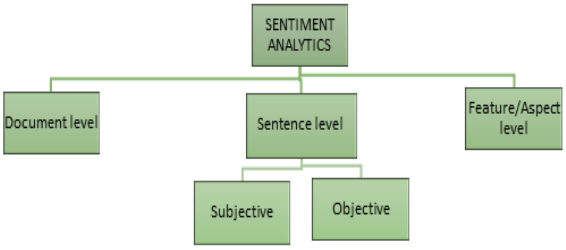

Fig. 1 Different Levels of Sentiment Analysis

These levels indicate whether the opinion is expressed within a sentence, document or based on aspects such as 'negative', 'positive' or 'neutral' opinion about an entity. Apart from these polarity-based opinion mining, going beyond polarity can lead to mining of emotional states such as 'sad', 'angry' or 'happy'.

Using sentiment analysis at the document level to identify the polarity of various medias such as reviews can provide us with the predictive quality of the entity. However, use of aspects based sentiment analysis can provide us with more detailed study as well as with more insights to the entity since it covers the entity from various aspects than from just analysing the polarity of the document.

In sentiment analysis normally, the neutral words are ignored. But in a different method of sentiment analysis, there uses a scaling system where the words denoting positive, neutral and negative words are given within a scale from +10 to -10. This eliminates the possibility of ignoring the neutral words nearly to an extent. Later the whole text is rated using the score obtained during the analysis. Since such a sophisticated and accurate method is used to evaluate the document in its textual level than in document level, this will ensure the correctness of the system to an extent.

The sentence level analysis involves classification of the text to classes containing subjective as well as objective. This can be difficult when compared to document level analysis. The subjective sentences may sometimes depend on the context and objective texts may contain subjective texts within them such as quotes containing opinions.

The feature/aspect-based analysis involves identifying and extracting opinions about various features of an entity. The various features of an entity can be found out by topic modelling or deep learning [16]. Sentimental expressions on different aspects of an entity helps getting the insight of in what it excels and what feature it has to develop more.

IV. RELATED WORKS

Commonly adopted approaches for sentiment analysis are Text mining approach, Naïve Bayes algorithm, machine learning techniques, support vector techniques and hidden markov models.

A method for automatic sentiment analysis of Twitter messages is proposed by Ana C.S.E Lima et al. [16] uses the Text mining approach, Naïve Bayes algorithm, word-based approach and emotion-based approach.

Another proposal by Bo Pang et al. [17] gives beliefs and perceptions of reality and the choices we make largely conditioned on how others see and evaluate. Techniques such as machine learning techniques and NLP techniques such as BOW and Sentiwordnet are used.

Huosong Xia et al. [18] proposed sentiment text classification of customer reviews on the web based on SVM. The influence of different stop word removal methods on the result of text classification and represent more effective stop word removal list are analysed in this paper.

A novel treatment of HMM model to use the result of sentimental subjectivity analysis in syntactic level task ie, POS tagging is proposed by Shichang Sun et al.[19].

An automatically dictionary construction approach and sentiment analysis of stock market news with semi supervised learning is proposed by Mizumoto.k et al[20].

Aurangzed Khan et al [21] proposed a method that classifies subjective and objective sentences from reviews and blog comments using SentiWordNet dictionary.

V. APPLICATIONS OF SENTIMENT ANALYSIS

Sentiment analysis is beneficial in various disciplines such as business, medical, social & politics, finance and education.

A. *BUSINESS*

Sentiments of customers is a key factor which has to be analysed in order to succeed in business. Customers are allowed to share their shopping experiences and reviews about product quality. Analysis of these customer sentiments help organisations to enhance their business by adopting necessary

steps to improve their customers' satisfaction level. They can design new marketing strategies based on the real -time information from customers to improve product features and predict chances of product failure. Application of aspects based sentiment analysis is detailed in a proposal by Zhang et al. [14] - a weakness finder system that help manufacturers to find their product weakness from Chinese reviews .

Another area where business organisations can make benefit out of sentiment analysis is Brand Reputation management (BRM). Opinions about advertising of products, public relations and corporate messaging are gathered and analysed to judge how company's product name, product or service is being perceived by people online.

B. *MEDICAL*

Analysis of the opinions of doctors and patients about medicine, treatments and hospital facilities are really beneficial. Information about a patient's health status in terms of observations and descriptions of examination outcome, diagnoses and interventions can be gathered from clinical documents. Analysing this information appropriately, assessing helpful or unhelpful clinical outcomes or judging the impact of a medical condition on patient's well-being are essential.

C. *SOCIAL AND POLITICS*

Analysing the opinions of public about Government's different policies helps the authorities to assess the strengths and weaknesses. For example, assessing success of electronic submission of tax returns, forecasting the implications of forthcoming rules , regulations and policies, etc. are potentials of sentiment analysis.

D. *FINANCE*

Financial markets are highly sentimental and emotional in relation to various events in the world. Sentiment analysis of the financial news which flood across social medias is highly useful to monitor the trend of the financial market. It is very helpful for the investors to take their financial decisions to choose the best portfolio from the multiple alternatives. Economic and political reports, price records, seasonal fluctuations and other unexpected events are some of the important factors that can be considered for the analysis of financial marker where opinion mining or sentimental analysis can be utilized.

E. *EDUCATION*

Sentiment analysis helps to improve students' learning by monitoring their performance. Opinions about the course, academic environment, interest towards subjects taught, can be gathered and analysed to address teaching and learning issues in the most effective way. Sentiment analysis techniques can classify the students' positive, negative or neutral feelings collected via social medias such as Facebook and Twitter. This can improve teaching to a great extent [15].

VI. CONCLUSION

Nowadays, all enterprises work with large volume of unstructured data which is a big business asset. Applying Sentiment analysis to mine this huge amount of unstructured data is really challenging. This paper précises related works in sentiment analysis and explores its applications in various domains. The challenges associated with sentiment analysis can be explored and further researches can be done to extent its applications in different domains .

REFERENCES

[1] Amir Gandomi and Murtaza Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, Issue 2, pp 137-144, April 2015

[2] Christopher C. Shilakes and Julie Tylman, "Enterprise Information Portals", Merrill Lynch, 16,November 1998 (*references*)

[3] Stephen Prentice, "From Data to Decision: Delivering Value from 'Big Data,'" Gartner Inc., March 28, 2012.

[4] Mona Tanwar, Reena Duggar and Sunil Kumar Khatri "Understanding Unstructured Data: A Wealth of Information in Big Data" Amity Institute of Information Technology, Amity University Uttar Pradesh, Noida, India, pp. 2, 2015. (*references*)

[5] J. Jiang, "Information extraction from text," in C. C. Aggarwal, & C. Zhai (Eds.), Mining text data, Springer, pp. 11–41, 2012.

[6] U. Hahn and I. Mani, "The challenges of automatic summarization," Computer, vol. 33(11), pp. 29–36, 2000.

[7] B. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, vol. 5(1), pp. 1– 167, 2012.

[8] J. Hirschberg, A. Hjalmarsson and N. Elhadad, "You're as sick as you sound: Using computational approaches for modeling speaker state tgauge illness and recovery," A. Neustein (Ed.), Advances in speech recognition, Springer, pp. 305–322, 2010

[9] H. A. Patil, "Cry baby: Using spectrographic analysis to assess neonatal health status from an infant's cry," in A. Neustein (Ed.), Advances in speech recognition, Springer, pp. 323–348, 2010.

[10] B. K. Panigrahi, A. Abraham and S. Das, "Computational intelligence in power engineering," Studies in Computational Intelligence, Springer, vol. 302, 2010.

[11] G. Barbier, and H. Liu, "Data mining in social media," C. C. Aggarwal (Ed.), Social network data analytics,

Springer, pp. 327–352, 2011.

[12] Charu C. Aggarwal, "An Introduction to Social Network Data Analysis," Social Network Data Analytics, Springer, 2011.

[13] C. C. Aggarwal, "An introduction to social network data analytics," C. C. Aggarwal (Ed.), Social network data analytics, Springer, pp. 1–15, 2011.

[14] W. Zhang, H. Xu, W. Wan, "Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis," Expert Systems with Applications, Elsevier, vol. 39, 2012, pp. 10283-10291

[15] Nabeela Altrabsheh, Mohamed Medhat Gaber, Mihaela Cocea," SA-E: Sentiment Analysis for Education",5th KES International Conference on Intelligent Decision Technologies, 2013-06-26 - 2013-06-28, Sesimbra,2013.

[16] Ana C.S.E Lima and Leandeo N.de castro,"Automatic sentiment analysis of Twitter Messages"International conference on computational aspects of social networks(CASON)p.4673-4794(2012)

[17] Bo Pang and Lillian Lee,"Opinion Mining and Sentiment analysis" *Foundations and Trends in Information Retrieval* 2(1-2), p. 1–135, 2008.

[18] Huosongxia, Min Tao and Yiwang," Sentiment Text classification of customers Reviews on the web Based on SVM"sixth International conference on Natural computation(ICNC) p.3633-3637(2010).

[19] Shichang Sun and Hongbo Liu," Twitter Part of Speech Tagging Using Pre-Classification Hidden markovModel"IEEE International conference on systems, Man and Cybernetic(SMC)p.1118-1123(2012).

[20] Mizumoto.k, Yanagimoto,H and Yoshioka M." Sentiment Analysis of stork Market News with Semi-Supervised Learning"IEEE/ACIS 11th International conference on Computer and Information Science(ICIS)p.325-328(2012).

[21] Aurangzed Khan and BaharumBuharudin" Sentiment Classification using Sentence-level Semantic Orientation of Opinion Terms form Blogs"International journal computer science emerging Tech.p.539-552(2011).