Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 05, May-2016

A Literature Review on Architecture, Classification Technique and Challenges of Sentiment Analysis

Pratishtha Parashar
M.Tech Student
MITS
Gwalior, MP

Sanjiv Sharma
Assistant. Prof
MITS
Gwalior, MP

***Abstract* -In the today's era on the development of web the consumers use to give their opinion on web rather than on being simple reviews .Better visualization is achieved by generating reviews which are visible to each and every user. Basically Opinion mining or sentiment analysis the field of data mining is done to present the exact reviews to users. It aims for distinguishing the emotions and feelings expressed within the reviews, classifying them into positive or negative or neutral and summarizing into the form that is quickly understood by users. This paper gives an insight on various works done in the area of opinion mining.**

Keywords - Data mining, Opinion mining , Sentiment analysis.

I.INTRODUCTION

Data mining refers to obtaining or "mining" knowledge from the vast amount of data. Data mining can be defined as the mining of useful patterns from a variety of data [1]. Data mining is the extraction of non trivial, implicit, previously unknown, potentially useful information and pattern from the data from large databases. Data mining is, basically, posing questions and finding the patterns related to them. The data can be found out by using various mining techniques. In the modern era the World Wide Web has impacted all aspects on our lives. The web has the same kind of attributes which make the mining of valuable information and knowledge a challenging task.

Data mining denotes to knowledge mining or extracting from huge amounts of data. Data mining is the process of discovering the intensive information about the huge amount of data which is stored in data warehouses and information repositories [2]. Data mining is the knowledge discovery method from the huge amount of data stored in various databases. Here the knowledge belongs to the valuable information which can be used further computation. The basic aim of the data mining is to mine higher-level unseen knowledge from raw data abundance. Data mining has been used in several domains of the data. Data mining can be regarded as an algorithmic procedure that takes information as input and yields different patterns, for example rules of the Classification, item sets, rules of association, or summaries, as output.
Data Mining involves an algorithmic process, which takes preprocessed input data and extracts patterns. Various techniques exist such as association rule mining, classification, clustering, etc. An important and widely used data mining technique is the discovery of association rules. Association rule mining aims at discovering frequent itemsets from market basket data and generating association rules.

Most association rule mining algorithms implicitly consider the utilities of the itemsets to be equal [Yao H. et al (2004)] [5]. A utility is a value attached to an item depending on its evaluation e.g. if coke has supported 20 and profit of 2%, cookies may have support 10 but with a profit of 20%.

Association Rule Mining (ARM) is a well-studied technique that identifies frequent itemsets from data sets and generates association rules by assuming that all items have the same significance and frequency of occurrence without considering their utility [3]. But in a number of real-world applications such as retail marketing, medical diagnosis, client segmentation, etc., utility of itemsets is based on cost, profit or revenue. Utility Mining aims to identify itemsets with highest utilities by considering profit, quantity, cost or other user preferences [4].

1.1Opinion Mining

Opinion Mining is a procedure to extricate the information from client assessment, surveys, emotions, and musings. Life of people is loaded with sentiments, feelings, musing and suppositions [6].We can't envision our existence without them. Suppositions assume a vital part in all human exercises. They lead the human life by affecting the way we think, what we do, how we do and how we act with the things. Having an entrance to the immense measure of information through the web and its change into a social web is no more a matter, as there are terabytes of information produced on the web every day that is accessible to any individual. The client of the data don't just expand the accessible information on the web, yet thus, effectively translate this substance and deliver new bits of data. Today, individuals not just give feeling on the data, additionally give comments however they likewise share their musings, data and learning with the extensive group. Along these lines, the whole group turns into a per user, notwithstanding being an essayist. The current mediums like Social Networks, Blogs, Wikis, where clients can share data, give audits and get input from different clients on various

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 05, May-2016

subjects, going from governmental issues and wellbeing to item surveys and voyaging. The exponentially expanding notoriety of data distribute on the web of various types proposes that stubborn data will turn into a vital part of the printed information on the web. As of late, numerous scientists have concentrated on this region. They are attempting to concentrate feeling data to break down the conclusions communicated naturally with frameworks. This new range of exploration is by and large called Opinion Mining and Sentiment Analysis. Presently, scientists have advanced different strategies to the arrangement of the issue. Ebb and flow day Opinion Mining and Sentiment Analysis is a region of examination at the junction of Information Retrieval and Natural Language Processing and impart some imperative attributes to different teaches, for example, Information Extraction , Text mining[7] .

Opinion mining is a sort of normal dialect handling for following the state of mind of people in general around a specific item or point. Assumption investigation, which is additionally called sentiment mining, includes in building a framework to gather and look at assessments about the item made in blog entries, remarks, surveys or tweets. Feeling investigation can be helpful in a few ways. For instance, in showcasing it helps in judging the accomplishment of a promotion crusade or new item dispatch, figure out which renditions of an item or administration are famous and even distinguish which demographics like or abhorrence specific components.

There are few difficulties in this field. The first is a conclusion word that is thought to be sure in one circumstance might be viewed as negative in another circumstance. A second test is that individuals don't generally express assessments samy. Most customary content preparing depends on the way that little contrasts between two bits of content don't change the significance in particular.

II. DATA SOURCE

User's feeling is a noteworthy rule for the change of the nature of administrations rendered and improvement of the deliverables. Websites, audit locales, information and small scale web journals give a decent comprehension of the gathering level of the items and administrations.

A. *Blogs*

With an expanding utilization of the web, blogging and blog pages are becoming quickly. Blog pages have turned into the most well known intends to express one's individual feelings.

B. *Review Sites*

Online websites are the most popular source of review data. User gives their opinion on these websites and these website analyze the opinion for future use.

C. Micro blogging

Micro blogging sites are the important and most popular source. Twitter is a micro blogging site on which people comments and shares their views in the form of "tweets".

III. ARCHITECTURE

Opinion has more importance in today's era. Opinion of any individual is important for others in different sense. If anyone can buy a product and he/she is not satisfied then they express their views and information in websites and other people get the information without physical interaction. Opinion mining process collects reviews and summarizes it[7]. There are three main steps, Opinion Retrieval, Opinion Classification, and Opinion Summarization.

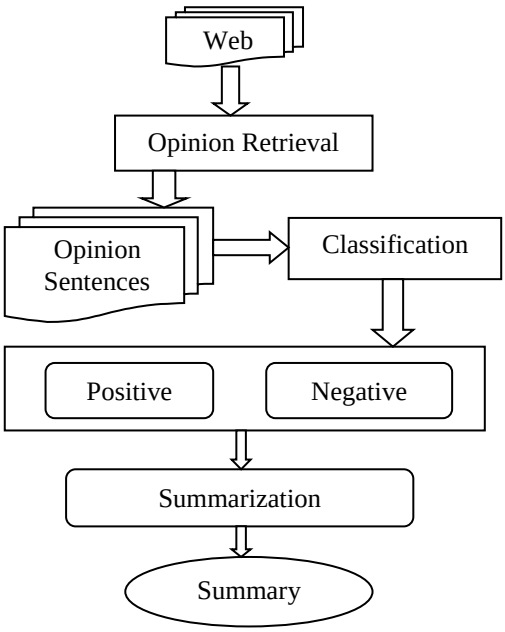

Figure1. Architecture of Sentiment Analysis

Opinion Retrieval

Opinion Retrieval is a first step of opinion mining process. Basically, in this step reviews or opinion sentences are collected from review websites. Various websites contain the review sentences about different products. There are also micro blogs and social sites on which opinions and views are shared by people. For opinion retrieval different information retrieval technique are used. Reviews are collected from various source and collected in a database.

Opinion Classification

In opinion classification step opinion sentences are classified into a pre-classified class such as positive and negative[7]. There are two type of classification technique named as Supervised and Unsupervised.

Opinion Summarization

Opinion Summarization is a most important and final step of this process. Summary of reviews, opinion and user reviews gave ought to be taking into account highlights or sub-topics that are specified in surveys. Various works done on opinion summarization. Basically, there are two techniques in opinion summarization,[8]. *Feature Based Summarization* work on the frequent word called "Feature", which is shown in many sentences. Summary is derived by collecting reviews that contain these feature word. *Term Frequency* is a count of word occurrences in a review document. Figure No1 Shows the architecture of sentiment analysis process.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 05, May-2016

IV. CLASSIFICATION TECHNIQUE

There are various algorithms for classification of opinion sentences. These algorithms classify opinion into class called positive and negative. Some of the algorithm is listed here.

A. Naïve Bayes

Naïve Bayes algorithm works on the principle of Bayes rule. This algorithm introduced by Thomas Bayes. Naïve Bayes classifier is comes under the category of supervised algorithm[11]. It is a probabilistic model. This classifier classifies opinion into positive and negative class. In this assumption is that each word is independent of each other so the classification process is fast. Firstly by using training data classifier is trained and then used for classification. This classifier is easy and efficient. There is one problem in this classifier assumption of independency of word, but it is not valid.

B. Support Vector Machine

Support Vector Machine is a supervised technique of classification of opinion sentences[10]. This is a binary classifier. Firstly, SVM classifier is trained by using set of N-pairs of words along with its category such as +1 or -1. Now, SVM function takes an unknown input and classifies it into category of positive and negative. Performance of SVM is good and it is less dependent.

C. Decision Tree

Decision Tree is also a classifier. In this classifier hierarchal tree is created. A component of this tree represents different things. In decision tree, feature is represented by internal nodes; test done on feature is represented by edges, and the result i.e. positive or negative is represented by leaf node. This classifier work in up-down manner.

V. ISSUES IN SENTIMENT ANALYSIS [9]

1. A positive and negative word may have their inverse importance in a specific space so it is difficult to foresee by its watchword meaning.

2. An Interrogative sentence might not have neither positive nor negative estimation but rather the watchword utilized asa part of the conclusion might be sure or not.

3. Few sentences as athletes may abuse the importance of entire sentence such sort of sentence need a force full consideration towards the watchword and sentences. These entertaining sentences not just violet the sentence of specific sentence additionally demolish the estimation of the entire report.

4. Sometimes opinion does not utilize any notion word like great, awful, bad, best and so forth however the sentences may have its positive and negative input about anything.

5. Spam slants are those feelings which are posted by the inverse or contender association for expanding their item esteem or their association esteem among the clients. Some government official may utilize the same spam audit to only for their reputation.

VI. RELATED WORK

Many algorithms have been proposed in order to understand and implement opinion mining and sentiment analysis. Researchers have developed models for identifying the polarity of words, sentences and whole document [2].Various tools are also available now for opinion extraction, sentiment analysis and opinion summarization. There have been researches regarding development for better algorithms for such tools.

Kyu, Liang and Chen [13] proposed algorithm for opinion extraction, opinion summarization and tracking the opinion which may be used for multiple languages. The opinion extraction algorithm takes value of opinion holder into consideration whereas in this paper the value of opinion holder is taken to be one.

Kin and Hovy [12] in their first model selected a topic and analyzed sentiment of remarks using word sentiment classifier with word net. The second model used probability of sentiment words.

Along the same line, He, Lin and Alani (2011) used joint topic modeling to identify opinion topics (which are similar to clusters in the above work) from both domains to bridge them. The resulting topics which cover both domains are used as additional features to augment the original set of features for classification.

In (Gao and Li, 2011), topic modeling was used too to find a common semantic space based on domain term correspondences and term co-occurrences in the two domains. This common semantic space was then used to learn a classifier which was applied to the target domain.

In (Wu, Tan and Cheng, 2009), a graph-based method was proposed, which uses the idea of label propagation on a similarity graph (Zhu and Ghahramani, 2002) to perform the transfer. In the graph, each document is a node and each link between two nodes is a weight computed using the cosine similarity of the two documents. Initially, every document in the old domain has a label score of +1 (positive) or -1 (negative) and each document in the new domain is assigned a label score based a normal sentiment classifier, which can be learned from the old domain. The algorithm then iteratively updates the label score of each new domain document i by finding k nearest neighbors in the old domain and k nearest neighbors in the new domain. A linear combination of the neighbor label scores and link weights are used to assign a new score to node i. The iterative process stops when the label scores converge. The sentiment orientations of the new domain documents are determined by their label scores.

Mihalcea & Moldovan [21] argue that the reduced applicability of statistical methods in word sense disambiguation is due basically to the lack of widely available semantically tagged corpora. They report research that enables the automatic acquisition of sense tagged corpora, and is based on (1) the information provided in Word Net, and (2) the information gathered from Internet using existing search engines.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 05, May-2016

Martinez & Garcia-Serrano [15] and Martinez et al.. (2000) propose a method for the design of structured knowledge models for NLP. The key features of their method comprise the decomposition of linguistic knowledge sources in specialized sub-areas to tackle the complexity problem and a focus on cognitive architectures that allow for modularity, scalability and reusability. The authors claim that their approach profits from NLP techniques, first-order logic and some modeling heuristics (Martinez et al.. 2000).

Cowie and Lehnert [16] reviewed the earlier research on IE and commented that the NLP research community is ill-prepared to tackle the difficult problems of semantic feature-tagging, co-reference resolution, and discourse analysis, all of which are important issues of IE research. Gaizauskas and Wilks (1998) reviewed the IE research from its origin in the Artificial Intelligence world in the sixties and seventies through to the modern days. They discussed the major IE projects undertaken in different sectors, viz., Academic Research, Employment, Fault Diagnosis, etc.

CONCLUSION

In this paper the survey of various techniques and challenges that are implemented for opinion mining are discussed. Opinion mining is a process which is used for auto extraction of knowledge from the opinion of others about some particular topic or problem. Various techniques like Naïve Bayes, SVM is discussed .The survey has been done on the opinion mining technique as user opinion has greater potential for knowledge discovery and decision support.

REFERENCES

[1] Han J and Kamber M: Data Mining: Concepts and Techniques. Second edition Morgan Kaufmann Publishers.
[2] Pang Ning Tan, Michael Steinbach, Vipin Kumar, “Introduction to Data mining”, 2009.
[3] Abhijit Raorane, R.V.Kulkarni, “Data Mining Techniques:A Source For Consumer Behavior Analysis”, International Journal of Database Management Systems, Vol.3,No.3,Aug. 2011, pp.45-56.
[4] Sudip Bhattacharya1 , Deepty Dubey, “High Utility Itemset Mining”, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 8, August 2012.
[5] H. Yao and H. J. Hamilton, “Mining itemset utilities from transactio databases,” Data and Knowledge Engineering, vol. 59, pp. 603-626 2006.
[6] Arti Buche, Dr.M.B.Chandak, Akshay Zadgoanakar “Opinion Mining and Analysis: A Survey”, International Journal on Natural Language Computing (IJNLC) Vol 2 No 3 June 2013Pg No 39-48.
[7] Blessy Selvam, A. Abirami, “A Survey on Opinion Mining Framework”, International Journal of Advanced Research in Computer and Communication Engineering, Vol 2, Issue 9, Sep 2013Pg No 3544-3549.
[8] Vijay .B. Raut et al, “Survey on Opinion Mining and Summarization of User Reviews on Web”, International Journal of Computer Science and Information Technologies (IJCSIT), Vol 5(2), 2014. 1026-1030.
[9] Ayesha Rashid et al, “A Survey Paper: Areas, Techniques and Challenges of Opinion Mining”, International Journal of Computer Science (IJCSI), Vol 10 Issue 6 No 2, Nov 2013.
[10] Dr. Ritu Sindhu, Ravendra Ratan Singh Jandail, Rakesh Ranjan Kumar, “A Novel Approach for Sentiment Analysis and Opinion Mining”, International Journal of Emerging Technology and Advanced Engineering (IJETAE), Vol 4, Issue 4, April 2014.
[11] Nidhi Mishra et al, “Classification of Opinion Mining Techniques”, International Journal of Computer Applications, Vol 56, No 13, Oct 2012Pg No 1-6
[12] Kim, S. and Hovy, E. Determining the Sentiment of Opinions. Proceedings of the 20th International Conference on Computational Linguistics (COLING’04), 2004.
[13] Lun-Wei Ku, Yu-Ting Liang and Hsin-Hsi Chen, “Opinion Extraction, Summarization and Tracking in News and Blog Corpora”, 2006 American Association for Artificial Intelligence.
[14] TheresaWilson, JanyceWiebe and Paul Hoffmann, “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis”, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 347–354, Vancouver, October 2005. c 2005 Association for Computational Linguistics.
[15] Martinez, P. & Garcia-Serrano, A. (1998). A knowledge-based methodology applied to linguistic engineering. In: R.N. Horspool (Ed.) Systems Implementation 2000. IFIP TC2 WG2.4 Working Conference on Systems Implementation 2000: Languages, Methods and Tools, 23-26 Feb. 1998, Berlin. London: Chapman & Hall pp. 166-179
[16] Cowie, J. & Lehnert, W. (1996) Information extraction. Communications of the ACM, 39, 80 – 91.