Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2015; 3(1B):71-75 **ISSN 2347-9523 (Print)**

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Review Article

Opinion Mining and Sentiment Analysis in Data Mining

Pragati Vaidya
Apaji Institute, Banasthali University, Rajasthan, India

***Corresponding author**
Pragati Vaidya
Email: shimpi27jan@gmail.com

Abstract: An essential part of our information-gathering behavior has constantly been to discover "what other people think". With the explosion of Web 2.0 platforms such as blogs, discussion forums, peer-to-peer networks, and various other types of social media and actively use information technologies to find and appreciate the opinions of others. This paper covers techniques that ability to directly facilitate opinion-oriented information-seeking systems and complete workflow of Opinion mining and sentiment analysis in which those techniques can be implement and also discusses about the application of opinion mining and sentiment analysis and discuss about tools through which are used to track the opinion or polarity from the user generated contents.
Keywords: Opinion Mining, Sentiment Analysis, Mining

INTRODUCTION

An essential part of our information-gathering behavior has constantly been to discover "what other people think". View of other people on a particular data set. With the continuously developing availability and admiration of opinion-rich resources such as online review sites, social networking's sites, social media and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others.

OPINION MINING AND SENTIMENT ANALYSIS AND DATA MINING

An *opinion* is a subjective statement, view, attitude, emotion, or appraisal about an entity or an aspect of the entity from an opinion holder [12]. Sentiment orientation of an opinion: positive, negative, or neutral (no opinion).It is also known as orientation, semantic orientation sentiment polarity. The term "sentiment" used in reference to the automatic analysis of evaluative text and tracking of the predictive judgments[13].

The term "Data mining" was introduced in the 1990s, but data mining is the evolution of a field with a long history. In the early 1960s, data mining was called statistical analysis, and the inventors were statistical software companies such as SAS and SPSS. By the late 1980s, the traditional techniques had been augmented by new methods such as fuzzy logic, heuristics and neural networks[14].

Opinion mining can be defined as a sub-discipline of computational syntax that concentrations on extracting people's opinion from the web. The current growth of the web encourages users to contribute and express themselves via blogs, videos, social networking sites, etc. All these platforms provide a vast quantity of valuable evidence that we are interested to analyze.

"Data" means information in a raw or unorganized form and "Mining" extract significant information. Data Mining is the nontrivial process of identifying valid, novel potentially useful, and ultimately understandable patterns in data.- Fayyad. [15].

Data Mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. – Zekulin.[16].

Data Mining is the set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data. – Ferruzza[17].

Data Mining is the process of discovering advantageous patterns in data. –John[18]

Data Mining is the decision support process where we look in large data bases for unknown and unexpected patterns of information.- Parsaye[19]

"Data Mining" is the process of extracting knowledge hidden from large volumes of raw data. The knowledge must be new, not apparent, and one must be able to use it. Knowledge Discovery database finding useful patterns in data. There is a

dynamic requirement for a new generation of computational models and tools to influence humans in extracting useful information (knowledge) from the promptly developing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD). Data Mining is the assessment and analysis of huge amounts of data in order to discover meaningful patterns and rules. Data Mining is a collection of powerful techniques intended for analysing large amounts of data. Basic need of data mining is that there are too much data and little information there is a need to extract useful information from the data and to interpret the data. There are various techniques in data mining such as Association Rule, Clustering, Decision Trees, and Neural Network.

Opinion Mining is a Big Business if Someone who wants to buy smart phones they definitely look for comments and reviews. Someone who just bought a smart phone on particular brand they comment on this and share their experience and smart phone manufacture they get feedback from customer and improve the quality of their products and adjust marketing strategies. Web 2.0 nowadays provides a great medium for people to share things. This provides a great source of unstructured information

(Particularly opinions) that may be useful to others (e.g. Companies and their rivals, other consumers...)

With the explosion of Web 2.0 platforms such as blogs, discussion forums, peer-to-peer networks, and various other types of social media . . . consumers have at their disposal a soapbox of unprecedented reach and power by which to share their brand experiences and opinions, positive or negative, regarding any product or service. As major companies are increasingly coming to realize, these consumer voices can wield enormous influence in shaping the opinions of other consumers — and, ultimately, their brand loyalties, their purchase decisions, and their own brand advocacy. . . . Companies can respond to the consumer insights they generate through social media monitoring and analysis by modifying their marketing messages, brand positioning, product development, and other activities accordingly. — Zabin and Jefferies[1].

Companies cannot ignore consumer reviews; they have to pay attention to it. In large grade they can influence other people about that product it will affect the company's reputation. Most of the companies contribute to online discussions on a frequent basis; they continuously interact with their consumer through mailing, chat, talks and improve the quality of the product, services according to the consumer.

LITERATURE SURVEY

J. Zabin and A. Jefferies [1] discuss the "Social media monitoring and analysis: Generating consumer insights from online conversation

Ayesha Rashid et al,[2] presented a survey of different Opinion Mining Techniques and their Challenges.

G.Vinodhini et al[4] presented an overview of different opinion mining techniques with approaches used.

ArtiBuche et al [5] proposed the work on how text is classified by Navie Bayes algorithm and also Hidden Markov Model to calculate the Entropy and Purity measure.

S.Chandrakala et al [6] proposed a work on recent papers on sentiment analysis and its related tasks with future challenges.

Nidhi Mishra et al [7] proposed the classification of opinion mining techniques. M Caraciolo. [8] it proposed Working of sentiment analysis on Twitter with Portuguese language.

Raisa Varghese et al [10] proposed the different levels of sentiment analysis and the major challenges involved in sentiment analysis.

OPINION MINING AND SENTIMENT ANALYSIS TECHNIQUES

Opining Mining is a relatively recent discipline that studies the extraction of opinions using Information Retrieval, Artificial intelligence and/or Natural Language Processing techniques. More informally, it's about extracting the opinions or sentiments given in a piece of text. There are various techniques used to extract information and knowledge are generalization, classification, clustering, association rule mining, data visualization, neural networks, fuzzy logic, Bayesian networks, and genetic algorithm, decision tree.

SUPERVISED MACHINE LEARNING:

Classification is most frequently used popular data mining technique [2]. Classification used to expect the potential effects from given data set on the basis of well-defined set of attributes and a given prognostic attributes. The given dataset is found to be the training dataset comprise on independent variables (dataset related properties) and a dependent attribute (expected attribute). A training dataset produced model test on test corpus contains the same attributes but no expected attribute. Correctness of tested model that how correct and efficient it is to make expectation. A Naive Bayes Classifier is a simple probabilistic classifier based on Bayes' theorem and is particularly suited when the

dimensionality of the inputs are high. Naïve Bayes classification is an approach to text classification that assigns the class c to a given document *d.*

$$c * = \text{argmax}_c P(c|d) \quad (1)$$

The *Naive Bayes* (NB) classifier uses the Bayes rule given in eq (2)

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \quad (2)$$

Where P(c|d) is the probability of instance d being in class c , P(d|c) is the probability of generating instance d given class c , P(c) is the probability of occurrence of class c and P(d) is the probability of instance d occurring [20].

UNSUPERVISED MACHINE LEARNING

In compare of supervised learning, unsupervised learning has no logical targeted output associated with input. Class label for any instance is unknown so unsupervised learning is about to learn by observation. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). Clustering has long been used for feature construction. The idea is to replace a group of "similar" variables by a cluster centroid, which becomes a feature.

WORKFLOW OF OPINION MINING AND SENTIMENT ANALYSIS

To carry out sentiment analysis are necessary several steps, in which are applied various techniques and methodologies:

Direct Opinion

Direct opinion gives positive or negative opinion about the object directly [7]. For e.g. the voice quality smart phone music system expresses direct opinion.

Comparison

Comparison means to compare the object with some other similar objects [7]. For e.g. "the quality of music system of smart phone-X is better than smart phone music system-Y" expresses the comparison

Data Collection and Pre-Process

In this stage it is acquired the raw data from different resource's such as blogs, social media etc., that will be analyzed for revealing of opinions. It is essential, to remove all substances that not express opinions. In this phase, pre-processing is done to eliminate redundant words or irrelevant opinions. It is compulsory to extract useful keywords from the raw data which can provide accurate information. These useful keywords are generally stored as an array of features A = (A1, A2, ..., An). Each element of array is a word from the original text, called aspect (feature).In short we take raw data and is pre-processed for feature extraction. this phase has been sub-divided into number of sub phases such as : Tokenization is that a text document in which it store collection of sentences, It split the sentence into tokens remove white space, comma ,unwanted symbols etc..

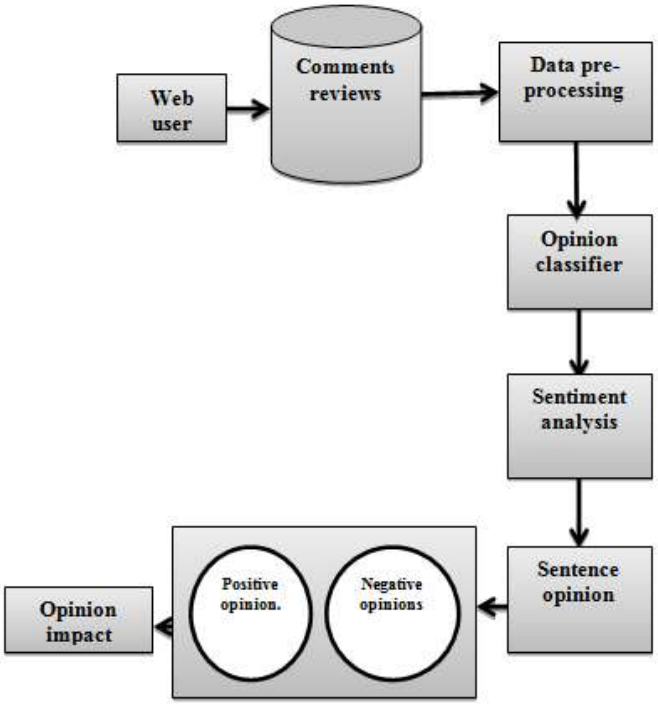

Fig-1: workflow of opinion mining and sentiment analysis

Feature Extraction

The feature extraction stage deals with feature types (which recognizes the type of features used for opinion mining), feature selection (used to select decent features for opinion classification), feature weighting appliance (weights each feature for good recommendation) reduction mechanisms (features for enhancing the classification procedure).

Feature Types

Types of features used for opinion mining could be features co-occurrence (features which occurs together like unigram, bigram or n-gram), features Part of speech information (POS tagger is used to separate POS tokens) features Negations (Negation words (not, not only) Opinion words (Opinion words are words which express positive or negative emotions)

Classification

This classification phase is subdivided into three classes (positive, negative and neutral). For sentiment analysis we use Classification algorithms which is dependent on two methods (supervised or unsupervised).

In this case we will use, for training, a collection of comments, this collection contains various sentences which is already classified as positive and negative opinion. Comments generally contain a number of sentences, but opinion will be determined at sentence level, and then later determine overall comment opinion. Obtained collection consists of two files, one for each set of positive and negative opinions, containing one sentence per line, making it easy to process. For further process ,to extract opinions we can use various algorithm but the Naïve Bayesian classifier one of the most frequently used for sentence classification , This type of classifier has the advantage that it is easy to implement, quickly and generate good results.

APPLICATION OF OPINION MINING AND SENTIMENT ANALYSIS

- Opinion mining is helpful to detect problems by listening, rather than by asking, thereby ensuring a more accurate reflection of reality.
- Opinion and sentiment analysis on products. A company is interested in customers' perceptions about its products. Information may be used to improve products and identifying new marketing strategies [8].
- Opinion and sentiment analysis on location. Tourist always curious to know about which place is best for vacation or famous restaurant or resorts. Through sentiment analysis it can be found appropriate information for planning a holiday trip.
- Opinion and sentiment analysis on election. Through sentiment analysis we can identify opinion of voter‘s about political party. Voting Advise Applications help voters understanding which political party is suitable for that position.
- Opinion and sentiment analysis on movies or software programs. We can identify users' sentiments from posted reviews on specialized sites. Different newspaper such as Times Of India ABP news etc.. also posts their reviews and comments on their own sites. Various social networking sites like Twitter and Facebook have become the boom of information source. People use these sites to express their opinion and sentiments about movies and software’s, these positive, negative and neutral comments or views are very important and appreciated for companies to improve their product.
- Opinion and sentiment analysis on financial markets. Through this sentiment analysis it gives the important information of stocks to investor in market and to identify price trends. Various stock market experts give their views about stocks which is very useful to investors.

TOOLS OF OPINION MINING AND SENTIMENT ANALYSIS

The tools which are used to track the opinion or polarity from the user generated contents are:

Review Seer Tool

This tool is used to automate the work done by aggregation sites. The Naive Bayes classifier approach is used to collect positive and negative opinions for assigning a score to the extracted feature terms. The results are shown as simple opinion sentence [4].

Web Fountain

It uses the beginning definite Base Noun Phrase (bBNP) heuristic approach for extracting the product features. It is possible to develop a simple web interface.

Red Opal

It is a tool that enables the users to determine the opinion orientations of products based on their features. It assigns the scores to each product based on features extracted from the customer reviews. The results to be shown with a web based interface [5].

Opinion Observer

This is an opinion mining system for analyzing and comparing opinions [6] on the Internet using user generated contents. This system shows the results in a graph format showing opinion of the product feature by feature. It uses WordNet Exploring method to assign prior polarity.

Freely Available Tools

A comprehensive list of such tools is available in http://groups.diigo.com/group/crossoverproject/content/tag/argumentmapping and

http://groups.diigo.com/group/crossoverproject/content/tag/VAA

There are currently freely available applications that simply analyze terms based on a pre-defined glossary, and giver highly simplified and unreliable results. One example is
http://twitrratr.com/

OPPORTUNITIES AND CHALLENGES

The detection of spam and fake reviews, mainly through the identification of duplicates, the comparison of qualitative with summary reviews, the detection of outliers, and the reputation of the reviewer[11].

The asymmetry in availability of opinion mining software, which can at present, be afforded only by organizations and government, but not by public. In other words, government has the means today to monitor public opinion in ways that are not available to the average citizens. While content production and publication has democratized, content analysis has not.

The incorporation of opinion with behavior and implicit data, in order to authorize and deliver further analysis into the data beyond opinion expressed

The constant need for better-quality usability and user-friendliness of the tools, which are currently usable mainly by data analysts

CONCLUSION

Opinions are so essential that whenever one wants to make a decision to buy product, one wants to listen to others' opinions about the quality of product and services. This is true for both individuals and organizations. Opinion mining and sentiment analysis is based on information gathering from current available opinion rich resources such as blogs, online review sites, social networking's sites, social media and personal blogs.opinion mining and sentiment analysis can be very helpful for organization to know what is the actual requirement of customer and how to fulfil their need.

REFERENCES

1. Zabin J, Jefferies A; Social media monitoring and analysis: Generating consumer insights from online conversation. Aberdeen Group Benchmark Report, 2008; 37(9).
2. Rashid A, Anwer N, Iqbal M, Sher M; A Survey Paper: Areas, Techniques and Challenges of Opinion Mining. IJCSI International Journal of Computer Science Issues, 2013;10(6).
3. Domingos P, Pazzani M; On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning, 1997; 29(2-3): 103-130.
4. Vinodhini G, Chandrasekaran RM; Sentiment analysis and opinion mining: a survey. International Journal, 2012; 2(6).
5. Buche A, Chandak D, Zadgaonkar A; Opinion Mining and Analysis: A Survey, 2013;2(3):39-48.
6. Chandrakala S, Sindhu C; Opinion Mining and Sentiment Classification: A Survey. ICTACT Journal on Soft Computing, 2012; 3(1):420-425.
7. Mishra N, Jha CK; Classification of Opinion Mining Techniques. International Journal of Computer Applications, 2012;56(13):1-6.
8. Caraciolo M; Working on sentiment analysis on Twitter with Portuguese language. [Online]. 2012. http://aimotion.blogspot.com/2010/07/working-on-sentiment-analysis-on.html
9. Smeureanu I, Diosteanu A, Delcea C, Cotfas LA; Business ontology for evaluating corporate social responsibility. Amfiteatru Economic, 2011;29:28-42.
10. Varghese R, Jayasree M; A Survey on Sentiment Analysis and Opinion Mining", International Journal of Research in Engineering and Technology, 2(11).
11. Osimo D, Mureddu F; Research Challenge on Opinion Mining and Sentiment Analysis. Universite de Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment, 2012; 508.
12. Liu B; Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 2012; 5(1):1-167.
13. Pang B, Lee L; Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2008; 2(1-2):1-135.
14. Han J, Kamber M, Pei J; Data mining, southeast asia edition: Concepts and techniques. Morgan kaufmann. 2006.
15. Fayyad U, Piatetsky-Shapiro G, Smyth P; From data mining to knowledge discovery in databases. AI magazine, 1996; 17(3):37.
16. Zekulin AD, Busche FD; U.S. Patent No. 6,430,547. Washington, DC: U.S. Patent and Trademark Office. 2002.
17. Gatnar E, Rozmus D; Data Mining—The Polish Experience. In Innovations in Classification, Data Science, and Information Systems (pp. 217-223). Springer Berlin Heidelberg. 2005.
18. John GH; Enhancements to the data mining process (Doctoral dissertation, stanford university). 1997.
19. Parsaye K; OLAP and Data Mining: Bridging the Gap. Database Programming and Design, 1997; 10:30-37.
20. Raghavan P, Tompson CD; Randomized rounding: a technique for provably good algorithms and algorithmic proofs. Combinatorica, 1987; 7(4):365-374.