# OPINION MINING: INSIGHT

Abhilasha Singh Rathor[1], Dr. Amit Aggarwal[2], Dr. Preeti Dimri[3]
[1]PHD CSE Scholar, UTU, Dehradun,
[2]Associate Professor, UPES, Dehradun
[3]Associate Professor, GBPEC-Pauri, Garhwal

**Abstract**
**There are two main types of textual information available on Web i.e. Facts and Opinion. Search engine currently available searches only for facts and assumes it to be true with the help of keywords. Search engines available do not search for opinion expressed. Since opinions cannot be expressed with few keywords. This paper focuses on opinion mining and provides an insight into its techniques, processes and applications in real life. It also explains why Opinion Mining is of high importance in current scenario when each and every kind of data is available on Web and how it helps in analysis and decision making in fields of education, corporate and research.**
**Keywords: Web Content Mining, Opinion Mining, Sentimental Analysis**

## I. Introduction

With the increased digital information and advancement in automated systems in every field the data is also growing rapidly. Therefore large amount of data is generated in each and every field i.e engineering, science, medical etc. Automated systems are designed to automate analysis, summarization and classification of data and numerous efficient ways to store large amount of data.

Text mining is most common way of extracting information, although it is not an easy task. Data Mining is looking for patterns in data while text mining is looking for patterns in text (1). Text mining is most popular in several fields like statics, machine learning, computational linguistics and information retrieval. One of its sub disciplines is Web mining, which deals with mining the semi structured web data. It has three parts (1) (2):

a) Web Content Mining: In this useful data, information and knowledge are extracted from web page content. In this useful information is discovered from web documents. In this mining the content can be image, text, video, audio, hyperlinks and metadata etc. it also differentiates personal home pages and other web pages (2). Research in this field includes resource discovery, document categorization and clustering and information extraction from web pages(3).

b) Web Structure Mining: In this mining focuses on the data that describes the structure of content. There are two types of structure intra-page structure and inter-page structure. Intra-page structure focuses on the links that exist within the same page while Inter-page structure focuses on links from one page to another (4).

c) Web Usage Mining: In this user access patterns are discovered from the web usage logs (4). There are many data mining techniques available to analyze and understand search patterns (5).

Opinion mining is a procedure of finding users opinion about specific product, topic or problem. With the help of opinion mining we train computer to understand, recognize and express emotions (1) (2)(6) (7).

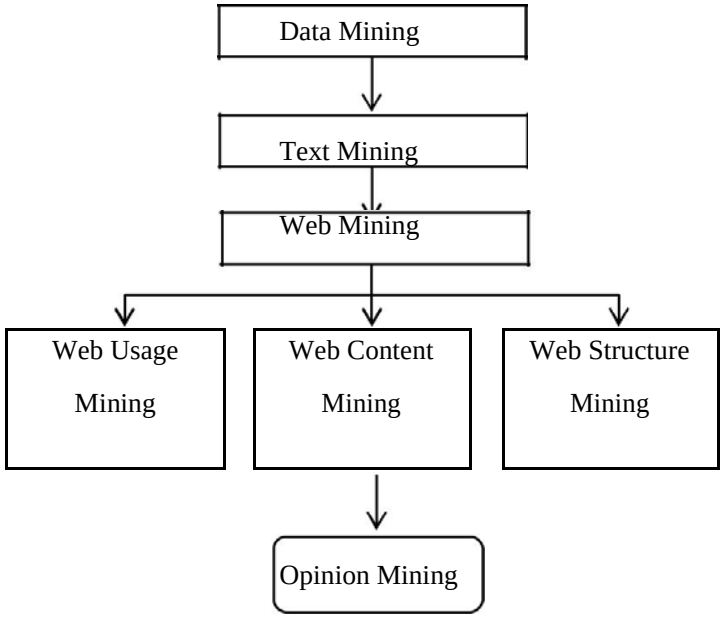


**Figure 1. Data Mining Hierarchical Model** (7)

**II. Web Content Mining**

Web content mining involves techniques for classification, clustering and summarization of the web contents, and offers interesting and useful patterns about needs and contribution of users and his behavior. Knowledge discovery is used to extract information from text documents and other multimedia documents such as videos, images, audios which are linked or embedded inside Web pages. The focus lies on information retrieval and text mining including information extraction, text clustering and classification and information visualization. Main web content mining techniques are as follows (8) (9)

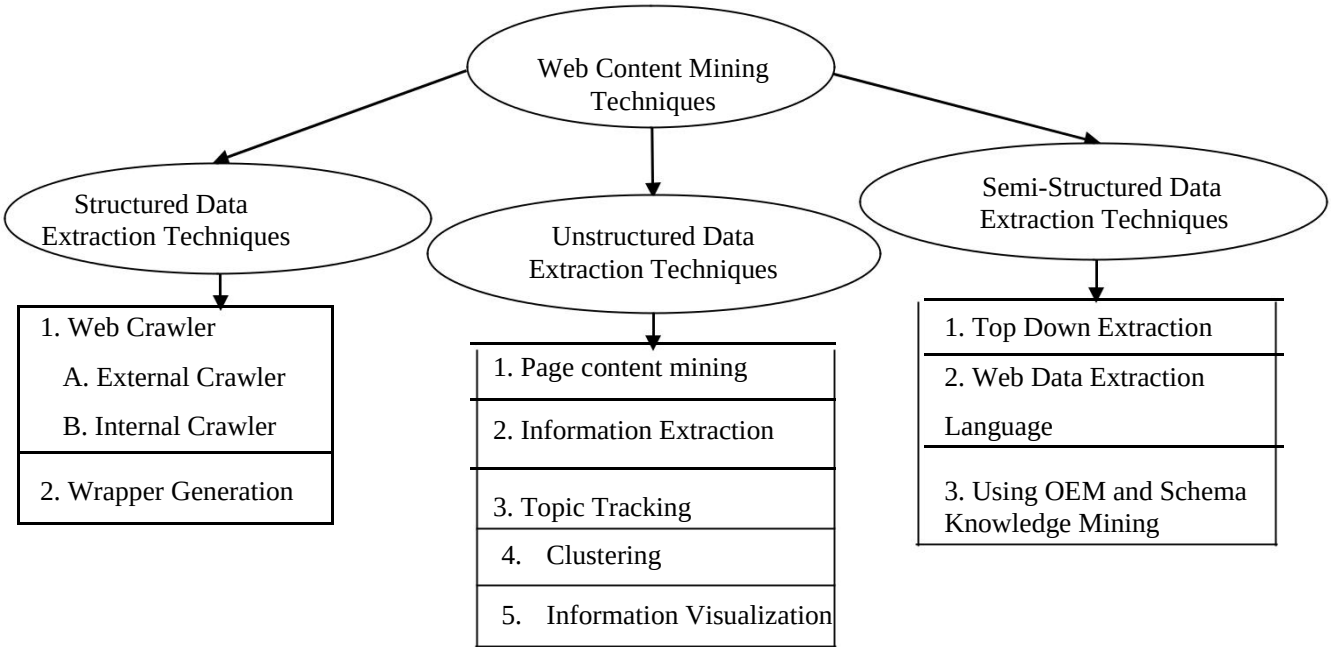


**Figure 2. Web Content Mining Techniques** (8)

A) Unstructured data mining techniques: Large amount of data is available on web some of it is in structured form while some in unstructured form. Numbers of the web pages are in form of text. The data retrieved is not only meaningful data, it may also be unknown information. Extracting information from HTML web page is challenging task, because HTML web pages contain multiple tags and data lies inside these tags. However with the advancement in the field and invention of modern tools like IEPAD, Decision tree, SVM it became easy to perform mining task (8) (5) (9).

B) Structured data mining techniques: it is the process of extracting structured information from web pages. Wrapper is the program for extracting such information. It includes data records retrieved from database and displayed in WebPages with the help of templates (table or form). Extracting such data is easy and useful. Crawlers or Web Spiders look for information across WWW. The create a copy of all visited pages for later processing by search engine which indexes downloaded pages to provide fast searches (9) (10).

C) Semi structured data mining techniques: Semi structured data are not full and grammatical text. Semi structured data is Hierarchical structured. They do not have a predefined structure (7). There are several techniques to extract semi structured data some of them are NLP techniques, wrapper generation, ontology, TINTIN. To extract such data common approach is to build a specific grammar which details the surrounding of each piece of data to be extracted (10) (9).

## III. Opinion Mining

Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language (9). It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining. The research has spread outside of computer science to the management sciences and social sciences due to its importance to business and society as a whole. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks (11). With the help of opinion mining we can find and extract particular information from text documents.

There are two types of Opinions (11) (12):

a) Direct Opinion: It provides positive or negative opinion about a product directly (9). For example
"This food is tasty" expresses a positive direct opinion of the user.

b) Comparison: In this user expresses his opinion by comparing one object with other similar objects. For example "Mango is tasty than Apple" expresses a comparative opinion of user that one product is better than other (13).

A) Pre-processing

It is process of extracting features by preprocessing the raw data. There are many sub phases in it. Some of them are discussed here (14):

a) Tokenization : In this text document is divided into tokens by removing commas, white spaces and other symbols etc. in some cases articles(the, a, an like) are removed.

b) With Stemming relevant tokens are shrunk into single type.

c) Case Normalization leads to conversion of entire document into one case either uppercase or lowercase.

B) Feature Extraction

It deals with the following (11) (15):

a) Feature Type: This identifies type of features used for opinion mining. It can be (16):

i. Term Frequency
ii. Term Co-occurrence
iii. Part of speech information
iv. Opinion words
v. Negations
vi. Syntactic Dependency

b) Feature Selection: Feature selection methods provide a criterion for eliminating terms from document corpus to reduce vocabulary space (11). It has been done in the following ways (15) (16):

i. Information Gain

ii. Odd Ratio

iii. Document Frequency

iv. Mutual Information

c) Feature Weighting Mechanism: It computes weight for ranking the features. The top n feature which gets sorted will be recommended. The various feature weighting mechanisms are (11) (16):

i. Term Presence and Term Frequency

ii. Term frequency and inverse document frequency (TF-IDF)

d) Feature Reduction: It reduces the feature vector size to optimize the performance of a classifier (4) (16).

C) Sentiment Analysis

Information composed from the reviews can be classified in the following ways to provide a good recommendation system (16) .

a) Document Sentiment Classification: in this the review document is taken as a whole and is trained with the labeled samples, as either positive or negative as a whole. The entire process is typically composed of two steps (16) (17):

i. Extracting the subjective features from the training data and converting them as feature vectors.

ii. Training the classifier on the feature vectors and classifying its subjectivity.

b) Sentence Level Sentiment Classification: They are just short documents, where the documents obtained from reviews is parsed into sentences (16). After that the sentences containing opinion words are extracted and classified into (17):

i. Subjective sentences: that holds opinions

ii. Objective sentences: that holds only factual information.

c) Word or Phrase Sentiment Classification:

In this the word level consolidation of sentiments is done. Mostly adjectives or adverbs words have semantic orientation which classifies the given word into positive, negative and neutral classes (17). The model of feature-based opinion mining and summarization is proposed in (18), which extracts the sentiment words from the reviews and classifies them (16). The approaches used to classify sentiments at word level could be grouped into two (17) (18):

i. Corpus based approaches

ii. Dictionary based approaches

The word level sentiment classification provides a fine grained sentiment classification (16).

D) Sentiment Classification techniques (11) (15) (16) (19) : It can be divided into three parts:

a) Machine Learning Approach (ML) : applies renowned ML algorithms and uses linguistic features. ML approach can be divided into following (11) (16) (20):

i. Supervised Learning Approach: it makes use of large number of labeled training documents (21).

ii. Unsupervised Learning Approach: it is used when it is difficult to find labeled training documents (22).

b) Lexicon Based Approach: relies on sentiment lexicon, which is collection of known and precompiled sentiment terms (23). It is divided into two parts, which uses statistical and semantic methods to find sentiment polarity:

i. Dictionary- based approach: it depends on finding opinion seed words and then

searching synonyms and antonyms in the dictionary (24).

ii. Corpus based approach : it starts with seed list of opinion words and then finding other opinion words in large corpus to help in finding opinion words with context specific orientations, which can be done by i) statistical methods or ii) semantic methods (16) (20).

c) Hybrid approach: it combines both approaches and is commonly used with sentiment lexicons, playing key role in the majority of methods (2) (22).

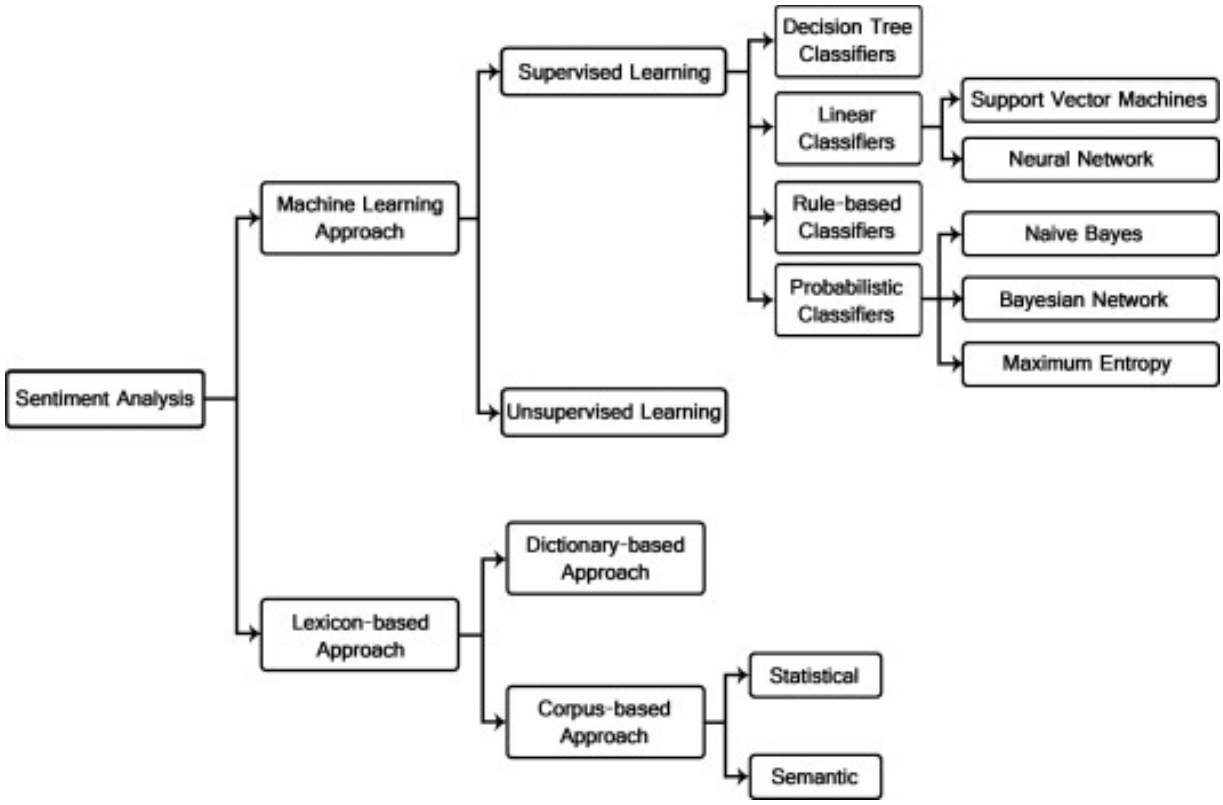


**Figure 3. Sentiment Classification Techniques**

## IV. Sentiment Tools Used For Opinion Mining Framework

A variety of open source text analytics tools used for natural language processing such as information extraction and classification can also be applied for opinion mining. Tools are as follows:

A) Ling Pipe: used for linguistic processing of text that includes clustering, entity extraction, and classification, etc. It is widely used NLP toolkit that is open source. This tool is considered good for its stability, speed, and scalability (25).

B) OpenNLP: It performs the common NLP tasks, including named entity extraction, chunking, POS tagging and co-reference resolution (26).

C) Stanford Parser and Part of Speech (POS) Tagger: It is used for sentence parsing and part of speech tagging from the NLP group (27).

D) NTLK: It is natural language toolkit. It is a tool for teaching and researching parsing, classification, and clustering (28).

E) OpinionFinder :It aims to identify subjective sentences and then mark the various aspects of subjectivity in those sentences, including the source of the subjectivity and words that are incorporated in phrases expressing positive or negative sentiments (29) (30) (31).

## V. Conclusion

Opinion mining or sentiment analysis analyzes people's opinions and attitudes towards entities such as products, events, topics, etc. Aspect and sentiment extraction are among the keys tasks This paper focuses on the opinion mining, web content mining its techniques.

There are many challenges that exist in the field of opinion mining such as implicit feature identification,, dealing noisy input texts, finding opinions from comparative sentences, extraction of opinion phrases and features, extraction of multiple opinions from the same document etc. In future, these challenges could be handled in a better way for providing good recommendations to the user.

**REFERENCES**

1. Web Personalization Using Web Mining: Concept and Research Issue . Pooja Mehtaa, Brinda Parekh, Kirit Modi, and Paresh Solanki. October 2012, International Journal of Information and Education Technology , pp. Vol. 2, No. 5,.

2. Web Content Mining: Its Techniques and Uses. Govind Murari Upadhyay, Kanika Dhingra. Issue 11, s.l. : International Journal of Advanced Research in Computer Science and Software Engineering, November 2013 , Vol. Volume 3. 2277 128X.

3. Web Page Data Collection Based on Multithread. Liu, Wentao. s.l. : ICCSEE,Atlantis Press, 2013.

4. An Efficient Approach for Mining Web Links based on Priority. K.Swathi, M.Meghana,T.Subba Reddy. 5, s.l. : International Journal of Advanced Research in Computer Science, May 2013, Vol. 4. 0976-5697.

5. Usage Mining for and on the Semantic Web. Stumme, Gerd, Andreas Hotho, and Bettina Berendt. s.l. : National Science Foundation Workshop on Next Generation Data Mining, November, 2002, Vol. 143.

6. Detection and Summarization of Genuine Review using Visual Data Mining. Jagruti Prajapati, Malay Bhatt, Dinesh J. Prajapati. 11, s.l. : International Journal of Computer Applications, April, 2012, Vol.43. 0975 – 8887.

7. Opinion Mining, Analysis and its Challenges. Nidhi R. Sharma, Prof. Vidya D. Chitre. 1, s.l. : International Journal of Innovations & Advancement in Computer Science, April, 2014, Vol. 3. 2347 – 8616.

8. Web Mining: A Tool for information retrieval from online marketing. Santosh Kumar Rath, Smaranika Mohapatra, Jharana Paikaray. 2, s.l. : IJARIIE, 2016, Vol. 2. 2395-4396.

9. Web Content Mining: Tool, Technique & Concept. X.Leela Mary, G.Silambarasan. 5, s.l. : International journal of Engineering Science and Computing, May,2017, Vol. 7.

10. Survey on Web Content Mining and Its Tools. T. Shanmugapriya, . P.Kiruthika. 8, s.l. : International Journal of Scientific Engineering and Research (IJSER), August, 2014, Vol. 2. 2347-3878.

11. Liu, Bing. Sentimental Analysis and Opinion Mining. s.l. : Morgan & Claypool, 2012.

10. 2200/S00416ED.

12. Resource Construction and Evaluation for Indirect Opinion Mining of Drug Reviews. Noferesti S, Shamsfard M. 5, s.l. : PLoS ONE , May, 2015, Vol. 10. e0124993.

13. Feature-based opinion mining and ranking.

Magdalini Eirinaki, Shamita Pisal, Japinder Singh. s.l. : Journal of Computer and System Sciences, Elsevier, 2012, Vol. 78. 1175–1184.

14. Mining opinion components from unstructured reviews: A review. Khairullah Khan, Baharum Baharudin, Aurnagzeb Khan. s.l. : Journal of King Saud University –Computer and Information Sciences, Elsevier, March, 2014. 1319-1578.

15. A Study on Sentiment Analysis: Methods and Tools. Abhishek Kaushik, Anchal Kaushik, Sudhanshu Naithani. 12, s.l. : International Journal of Science and Research (IJSR), Dec,2015, Vol. 4. 2319-7064.

16. A SURVEY ON OPINION MINING. Blessy Selvam, S.Abirami. 9, s.l. : International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2. 2319-5940.

17. CLOpinionMiner: Opinion Target Extraction in a Cross-Language Scenario. Xinjie Zhou, Xiaojun Wan, Jianguo Xiao. 4, s.l. : IEEE/ACM TRANSACTION ON AUDIO, SPEECH, and LANGUAGE PROCESSING, April 2015, Vol. 23. 2329-9290.

18. Opinion Observer: Analyzing and Comparing Opinions on the Web. Liu .B, Hu. M and Cheng. J. s.l. : International World Wide Web Conference, 2005.

19. APPLICATION OF THE SENTIMENT CLASSIFICATION TECHNIQUES FOR WEBSITE MONITOR SYSTEM. CHAO-RONG LI, MING- QING HUANG, YONG-HAI DENG, JIAN-PING LI. 10, s.l. : IEEE, 2010, Vol. 5. 978-1-4244-8026

20. A Survey on Sentiment Analysis and Opinion Mining. Dudhat Ankitkumar M, Prof. R. R. Badre, Prof. Mayura Kinikar. s.l. : International Journal of Innovative Research in Computer and Communication Engineering, 2012. 2320-9798.

21. A Survey on Sentiment Analysis Algorithms and Techniques. Prajakta Gosavi, Vaishali Shirsath. 3, s.l. : International Conference On Emanations in Modern Technology and Engineering (ICEMTE-2017), 2017, Vol. 5. 2321-8169.

22. New avenues in opinion mining and sentiment analysis. Cambria, Erik. 2, s.l. : IEEE Intelligent Systems, 2013, Vol. 28.

23. A Lexical Approach for Tweets Sentiment Classification. Sneha Vishwakarma, Suresh Kumar. s.l. : National Conference on Recent innovations in Applied Sciences and Humanities' NCASH-2015, 2015. 2277-8179.

24. Sentiment Analysis Based on Dictionary Approach. Reshma Bhonde, Binita Bhagwat, Sayali Ingulkar, Apeksha Pande. 1, s.l. : International Journal of Emerging Engineering Research and Technology , 2015, Vol. 3. 2349-4395.

25. http://aliasi.com/lingpipe/. [Online]

26. http://opennlp.apache.org/. [Online]

27.http://nlp.stanford.edu/software/tagger.shtm/. [Online]

28. http://www.nltk.org/ . [Online]

29.http://code.google.com/p/opinionfinder/. [Online]

30. A survey on sentiment analysis challenges. Doaa Mohey, El-Din, Mohamed Hussein. s.l. : Journal of King Saud University – Engineering Sciences, Elsevier. 1018-3639.

31. Using Opinion Mining Techniques in Tourism. Bucur, Cristian. Prague, Czech Republic : 2nd Global Conference on Business, Economics,Management and Tourism, October, 2014. 1666 – 1673.