ISSN: 2321-7782 (Online)

Volume 3, Issue 3, March 2015

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Sentiment Analysis and Classification: A Survey

Shailesh Kumar Yadav
Department of Computer Science
Pondicherry University
Pondicherry, India

Abstract: Sentiment Analysis (SA) or opinion mining (OM) has recently become the focus of many researchers, because analysis of online text is beneficial and demanded for market research, scientific surveys from psychological and sociological perspective, political polls, business intelligence, enhancement of online shopping infrastructures, etc. Nowadays if one wants to buy a consumer product one prefer user reviews and discussion in public forums on web about the product. As a result opinion mining has gained importance. This online word-of-mouth represents new and measurable source of information with many applications, this process of identifying and extracting subjective information from raw data is known as sentiment analysis. This paper presents a survey on sentiment analysis or opinion mining.

Keywords: sentiment analysis; sentiment identification; sentiment classification; classification techniques

I. Introduction

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g. sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. However, they are now all under the umbrella of sentiment analysis or opinion mining. While in industry, the term sentiment analysis is more commonly used, but in academia both sentiment analysis and opinion mining are frequently employed. They basically represent the same field of study. The term sentiment analysis perhaps first used by Nasukawa and Yi in 2003[1], and the term opinion mining first used by Dave, Lawrence and Pennock, in 2003[2].

Initial research in text mining [7], [8] focused on extracting factual information from documents. In recent times, focus is shifting towards opinion mining - also called sentiment analysis. One of the drivers for this shift is availability of opinionated text in the form of reviews, blog posts, social media comments and more recently, tweets. Such documents are also called User Generated Content (UGC).

Sentiment mining from user generated content is tedious task, because it needs in-depth knowledge of syntactical and semantic, the explicit and implicit, regular and irregular language rules. The researchers involved in sentiment analysis faces challenge with NLP's unresolved problems. Some NLP's unresolved problems are anaphora determination, co-reference resolution, named-entity recognition, word-sense disambiguation and negation handling. Opinion mining is an exceptionally confined NLP issue, because the system only needs to understand the polarity of sentiment that is positive or negative sentiments of each sentence and the target entities, or aspect. Therefore, sentiment analysis is an open door for NLP researchers to make substantial progress on all fronts of NLP, and conceivably have a tremendous practical impact. Figure 1 shows sentiment analysis process on UGC. Here UGC is about product reviews.

II. LITERATURE SURVEY

In [1], Nasukawa and Yi. Illustrate a sentiment analysis approach to extract sentiments associated with polarities of positive or negative for specific subjects from a document, instead of classifying the whole document into positive or negative. The essential issues in sentiment analysis are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject. Powerful functionality for these kinds of issues is used.

In [3], Ding et al. proposed an effective method for identifying semantic orientations of opinions expressed by reviewers on product features. It is able to deal with two major problems with the existing methods, (1) opinion words whose semantic orientations are context dependent, and (2) aggregating multiple opinion words in the same sentence. For (1), a holistic approach is proposed that can accurately infer the semantic orientation of an opinion word based on the review context. For (2), a new function to combine multiple opinion words in the same sentence is proposed.

Taylor et al. [4] presented a generic design of a tourism opinion mining system that aims to be useful in many industries. They also used their proposals to successfully implement the system and solve a specific problem in the Lake District tourism industry.

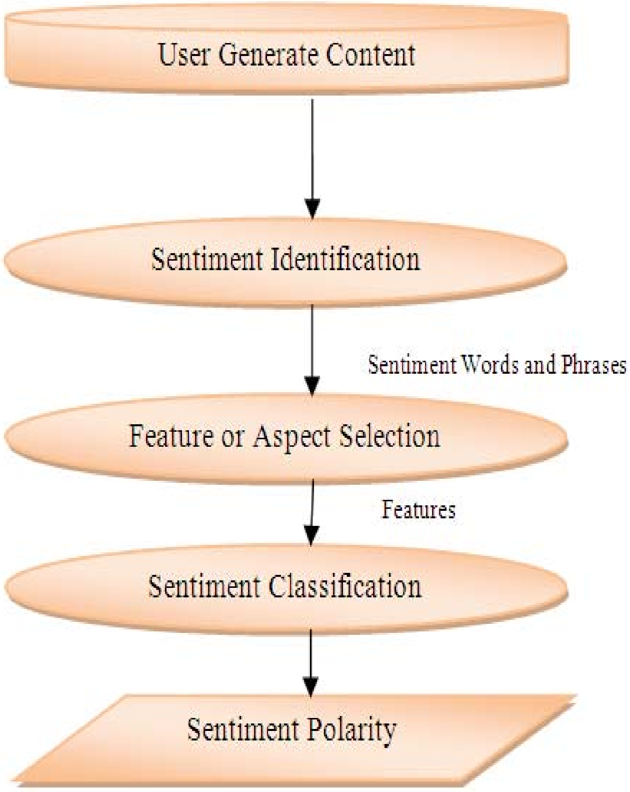

Fig. 1 Sentiment analysis process on user generated content.

In Zhu et al [5] an aspect-based opinion polling system takes as input a set of textual reviews and some predefined aspects, and identifies the polarity of each aspect from each review to produce an opinion poll.

In [6] Haddi, Lui and Shi investigated the sentiment of online movie reviews. They used a combination of different pre-processing methods to reduce the noise in the text in addition to using chi-squared method to remove irrelevant features that do not affect its orientation. Authors have reported extensive experimental results, showing that, appropriate text pre-processing accuracy achieved on the two data sets is comparable to the sort of accuracy that can be achieved in topic categorization, a much easier problem.

In Moraes, Valiati & Neto [9] they have focused on comparing SVM and ANN in terms of the requirements to achieve better classification accuracies. In this, experiments evaluated both methods as a function of selected terms in a bag-of-words (unigrams) approach. Regarding the sentiment learning literature, the main findings/contributions are in the two points. First point is in terms of classification accuracy on the benchmark dataset of Movies reviews and second point as an overall comparison in the context of balanced data.

III. Sentiment Classification

There is an important assumption about sentiment classification. Sentiment classification assumes that the opinion document express opinions on a single entity or object and opinions are from a single opinion holder. Opinionated documents contain information which can be broadly categorised in two categories: facts, which are typically objective statements about some entity (object) or event and sentiments, which are subjective in nature expressing sentiments and feelings of the opinion holder about the entity. Both facts and opinions are useful in decision making.

In the literature, sentiment classification is based on polarity. Polarity can be positive, negative or neutral. That is opinions can be classified as positive, negative or neutral (Table I). In [10] there is also a fourth type - a constructive opinion. In this opinion holders give suggestions to improve or make the product or service better. Constructive opinions need not imply that the opinion holder is negatively inclined about the entity.

Opinions can also be classified into three types: direct opinions, comparative opinions and indirect opinions. In direct opinion, opinion holder directly attack to target. Indirect opinions are either implied as in idioms or expressed in a reverse way as in sarcasm. In comparative, opinion holder generally compare among entity.

Basically, researchers have studied opinion mining at three levels of granularity, namely, document level, sentence level, and aspect level.

a) Document Level Sentiment Classification

Document-level sentiment classification aims to automate the task of classifying UGC, which is given on a single entity or aspect. In this, overall sentiment of UGC can be determined by the polarities of different sentiment words used in the UGC. However, other studies have focused on developing sentiment dictionaries (lexicon) .This classification does not work with forum and blog posting because in such a posting the author may express opinions on multiple products, and compare them using comparative sentences.

The main task of document-level sentiment classification is to identify the polarities of UGC.

Two type of classification techniques have been used in document-level sentiment classification, supervised method and unsupervised method.

1) Supervised Methods: Sentiment classification can be formulated as a supervised [19], [11], [22], [23] learning problem with four classes, positive, negative, neutral and constructive. User generated contents mostly are used as training and testing data. Any existing supervised learning techniques can be used to sentiment classification, such as naïve Bayes and support vector machines (SVM). Pang et.al. [11] were one of the very first to perform SA on online movie reviews. They tested machine learning (LM) approaches, namely, SVM, MaxEnt and NB classifiers, and trained them on different feature sets including unigrams. Their findings showed that an SVM trained on a unigram bag-of words feature set, outperforms all other approaches presented in their work. Drawbacks of standard ML approaches are that for opinion polarity detection they are both domain and temporally dependent [20]. In most cases, SVMs performed well over Naïve Bayes classifiers. Whenever, the set of training data is small, a Naïve Bayes classifier might be more appropriate. Several techniques and features are used by researchers in learning process. One of the most fundamental tasks in sentiment classification is selecting an appropriate set of features. Some of the important features are:

Terms and their Frequency: These features are individual words (unigram) and their n-grams with associated frequency counts. They are also the most common features used in traditional topic-based text classification.

Part of speech: The part-of-speech (POS) of each word can be important too. Words of different parts of speech (POS) may be treated differently for example adjectives carry a great deal of information regarding a document's sentiment.

Sentiment words and phrase: Sentiment words or opinion words are words in a language that are used to express positive or negative sentiments. For example, good, awesome, and nice are positive sentiment words, and defective, poor, and risky are negative sentiment words.

Rules of opinions: Apart from sentiment words and phrases, there are also many other expressions or language compositions that can be used to express or imply sentiments and opinions.

Sentiment shifters: These are expressions that are used to change the sentiment orientations, e.g., from positive to negative and vice versa or from negative to constructive.

Negation words are the most important class of sentiment shifters. For example, the sentence "I don't like this smart phone" is negative.

Syntactic dependency: Words dependency-based features generated from parsing or dependency trees. These are also tried by researchers.

2) Unsupervised Methods: Opinion or sentiment words and phrases are the dominating indicators for sentiment classification. Thus, using unsupervised learning based on such words and phrases would be quite natural. For example, the method in [21] uses known opinion words for classification, while [16] defines some phrases which are likely to be opinionated.

Turney [16] displayed a straightforward unsupervised learning calculation for characterizing a review as suggested or not suggested. He figured out whether words are positive or negative and how solid the assessment is by figuring the words' point wise mutual information (PMI) for their co-occurrence with a positive seed word ("excellent") and a negative seed word ("poor"). He called this value the word's semantic orientation. This technique checked through an audit searching for expressions that match certain grammatical feature designs (descriptive words and intensifiers), computed the semantic orientation of those phrases, and added up the semantic orientation of all of those phrases to compute the orientation of a review. He accomplished 74% accuracy classifying a corpus of item reviews.

Harb et al. [17] performed blog classification by beginning with the 2 sets of seed words with positive and negative semantic introductions separately, as in [23] and utilized Google's web search engine to make association rule that find more. They counted the number of positive versus negative adjective in a document to classify the documents. A lexicon-based method to sentiment classification was presented by Taboada et al. [18]. They used dictionaries of positive or negative polarized words to this classification task. A semantic orientation calculator was build based on these dictionaries by incorporating intensifiers and negation words.

Table 1shows Techniques used for Document Level Classification in previous research papers

References	Dataset	Features	Techniques	Classifications Approach
[10]	Restaurant Reviews	Unigram, Bigrams, Trigrams	SVM, Naïve Bayes	Supervised
[12]	Reviews to travel destination	Unigram Frequency	SVM, Naïve Bayes, character based N-gram model	Supervised
[13]	Movie reviews, Product reviews, MySpace comments	POS tag, N-grams	SVM , Rule based Classifier	Supervised Learning and Rule-based Classification
[14]	Movie Reviews, MPQA dataset	Unigram, bigram and extraction pattern feature	SVM	Supervised
[15]	Movie Reviews	Adjective word frequency, percentage of appraisal groups	SVM	Supervised
[16]	Automobile bank, movie, travel reviews	adjectives and adverbs	PMI-IR	Unsupervised
[17]	Movie Reviews	adjectives and adverbs	Association Rule	Unsupervised
[18]	Movie Reviews, Camera Reviews	Adjectives , Nouns, verbs , Adverbs, Intensifier, Negation	Dictionary based approach	Unsupervised

Table 1.

b) Sentence level Sentiment Classification

The task of classifying a sentence as subjective or objective is often called subjectivity classification. The resulting subjective sentences are also classified as expressing positive or negative opinions, which is called sentence-level sentiment classification. In the sentence level sentiment analysis, the polarity of each sentence is calculated. This is similar to a document level sentiment analysis but done at a sentence level [24]. It assumes each sentence contains an opinion for one entity and aspect, and some of the sentences may not be opinionated (objective). The subjective sentences contain opinion words which help in determining the sentiment about the entity. A two stage inference is done for each sentence: first, each sentence is classified as subjective or objective and then the polarity of each of the subjective sentences are inferred. There may be complex sentences also in the opinionated text. In such cases, sentence level sentiment classification is not useful.

c) Aspect level Sentiment Classification

In a typical opinionated document, the author writes both positive and negative aspects of the entity, although the general sentiment on the entity may be positive or negative. Document and sentence sentiment classification does not provide such information. To obtain these details, we need to go to the aspect level. It assumes that a document contains opinion on several entities and their aspects. Aspect level classification requires discovery of these entities, aspects, and sentiments for each of them.

IV. Sentiment Classification Techniques

In general, Sentiment Classification can be done with three techniques machine learning (ML) approach, lexicon based approach and hybrid approach [25]. The ML approach applies the famous ML algorithms and uses linguistic features. The

Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. Lexicon is an important indicator of sentiments called opinion words. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity and determines the emotional affinity of words, which is to learn their probabilistic affective scores from large corpora. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods. The various approaches and the most popular algorithms of sentiment classification are illustrated in Fig. 2.

The text classification methods using ML approach can be roughly divided into supervised and unsupervised learning methods. The supervised methods make use of a large number of labelled training documents. The unsupervised methods are used when it is difficult to find these labelled training documents.

The lexicon-based approach depends on finding the opinion lexicon which is used to analyse the text. There are two methods in this approach. The corpus based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical or semantic methods. The dictionary-based approach which depends on finding opinion seed words, and then searches the dictionary of their synonyms and antonyms.

There is a brief explanation of both approaches' algorithms and related articles in the next subsections. A widely used approach for lexicon acquisition is by using a seed lexicon. In this approach, a vocabulary of words and phrases is used, each of these words and phrases have positive, negative or neutral sentiment. Acquisition of the lexicon is a two-step process. In step one, opinionated words are identified in the document and in step two polarity of the opinionated word is inferred.

V. Tools Used For Sentiment Classification

There are so many open-source text-analytics tools used for natural language processing such as information extraction and classification can also be applied for sentiment analysis.

Tools are listed below:

1. NTLK: The natural language toolkit is a tool for text processing, classification, tokenization, stemming, tagging, parsing etc. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as Word Net, along with a suite of http://www.nltk.org/

2. GATE: Useful if you want to develop a pipeline. Language analysis modules for various languages are contributed by developers are available to be used plugged in your pipeline.

3. OpenNLP: perform the most common NLP tasks, such as POS tagging, named entity extraction, chunking and co-reference resolution. http://opennlp.apache.org/

StanfordCoreNLP: If you need part of speech categories, syntactic analysis (phrase structure or dependency analysis), co-reference or named entities in text.

4. OpinionFinder: It aims to identify subjective sentences and to mark various aspects of subjectivity in these sentences, including the opinion holder of the subjectivity and words that are included in phrases expressing positive or negative sentiments. http://code.google.com/p/opinionfinder/

5. Ling Pipe: Ling Pipe is used for linguistic processing of text including, clustering classification and entity extraction etc. http://alias-i.com/lingpipe/

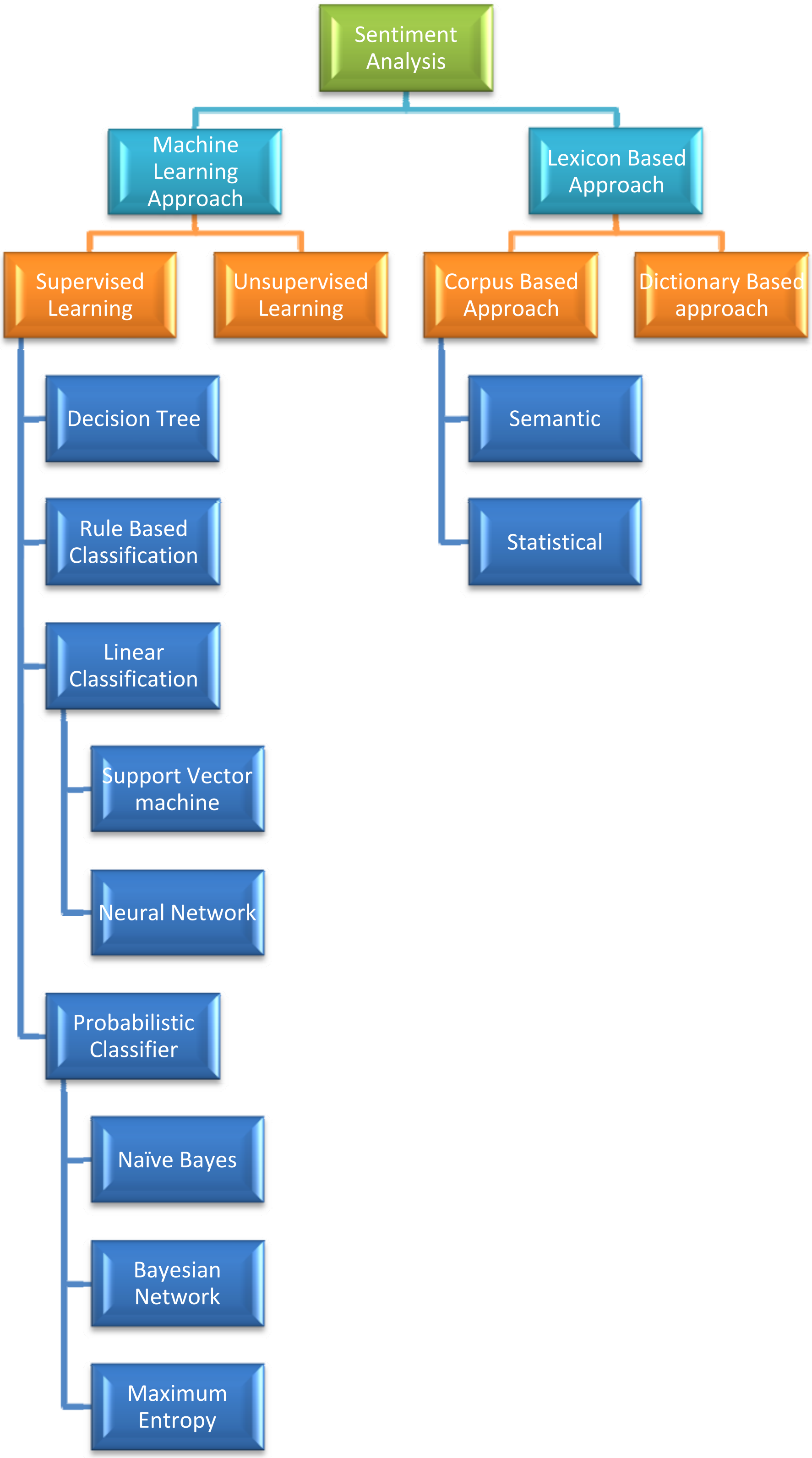

Fig. 2 Sentiment Classification Techniques

VI. Application, Challenges And Issues

OM has various applications in different fields. It can be used in online advertising, hotspot detection in forums, search engines, recommendation systems, email filtering, questioning/answering systems, etc. OM application in daily life is most interesting as OM can be used to improve human–computer interactions, business intelligence, opinion poll that is voter can see opinion of other people before going to poll government intelligence, citation analysis etc. The following sample questions could be helpful in better understanding the applications of OM.

- » What do individuals think about government strategies?
- » Which feature of an item are loved or hated by overall population
- » Who is a solid competitor for the general race body?
- » Why has a sale of product declined?

OM suffers from several challenges, such as determining which segment of text is opinionated, identifying the opinion holder, determining the polarity strength of sentiment. Sentiment analysis is concerned with the human reviews, emotions and sentimental discussion. Everyone has their own perception and concern about a particular problem, issue, or topic. Opinionated text may be fake, irrelevant and or ambiguous information. Opinions are far harder than facts to describe.

VII. Conclusion

Sentiment analysis is vast research are with several challenges. It has a wide variety of applications in e-commerce. It helps in classifying, summarizing reviews and in other real time applications. This paper focuses on sentiment classification, classification techniques and what tools are available for sentiment analysis. There are still some open challenges exist in this area such as discovering of sentiment and their polarity in complex sentences, implicit aspect identification, extraction of opinion phrases and features from different corpora, extraction of multiple opinions from the same document etc. The vocabulary of natural language is very large that things become even hard. Therefore, several challenges exist in the field of machine learning. These problems have to be tackled separately and those solutions can be used to improve the methods to do sentiment analysis and classification

References

1. T. Nasukawa, “Sentiment Analysis: Capturing Favorability Using Natural Language Processing Definition of Sentiment Expressions,” pp. 70–77, 2003.
2. K. Dave, I. Way, S. Lawrence, and D. M. Pennock, “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews,” 2003.
3. X. Ding, S. M. Street, B. Liu, S. M. Street, P. S. Yu, and S. M. Street, “A Holistic Lexicon-Based Approach to Opinion Mining,” pp. 231–239, 2008.
4. E. Marrese-Taylor, J. D. Velasquez, and F. Bravo-Marquez, “Opinion Zoom: A Modular Tool to Explore Tourism Opinions on the Web,” 2013 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol., pp. 261–264, Nov. 2013.
5. J. Zhu, H. Wang, M. Zhu, B. Tsou and Matthew M, “Aspect-Based Opinion Polling from Customer Reviews", " Ieee Transaction On Affective Computing”, vol. 2, NO. 1, January-March 2011.
6. E. Haddi, X. Liu, and Y. Shi, “The Role of Text Pre-processing in Sentiment Analysis,” Procedia Comput. Sci., vol. 17, pp. 26–32, Jan. 2013.
7. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” vol. 3, pp. 993–1022, 2003.
8. T. Hofmann. P. latent, “ semantic indexing,” In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in inFmnation retrieval, SIGIR '99, pages 50-57, New York, NY, USA, 1999. ACM
9. R. Moraes, J. F. Valiati, and W. P. Gavião Neto, “Document-level sentiment classification: An empirical comparison between SVM and ANN,” Expert Syst. Appl., vol. 40, no. 2, pp. 621–633, Feb. 2013.
10. R. Arora and S. Srinivasa, “A Faceted Characterization of the Opinion Mining Landscape,” pp. 1–6, 2014.
11. [11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," presented at the Proceedings of the ACL-02 conferenceon Empirical methods in natural language processing - Volume 10, 2002.
12. [12] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," Expert Systems with Applications,vol. 36, pp. 6527-6535, 2009.
13. R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," Journal of Informetrics, vol. 3, pp. 143-157, 2009.
14. E. Riloff, S. Patwardhan, and J. Wiebe, "Feature subsumption for opinion analysis," presented at the Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 2006.
15. C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," presented at the Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, 2005.
16. P. D. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” no. July 2002, 2001.
17. A. Harb, M. Planti, M. Roche, A. Harb, M. Planti, M. Roche, N. Cedex, and A. Harb, “Web Opinion Mining: How to extract opinions from blogs? To cite this version: Web Opinion Mining: How to extract opinions from blogs? Categories and Subject Descriptors.”
18. M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Comput. Linguist. vol. 37, pp. 267-307, 2011.
19. B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” Found. Trends® Inf. Retr., vol. 2, no. 1–2, pp. 1–135, 2008.

20. J. Read, “Using emoticons to reduce dependency in machine learning techniques for sentiment classification,” In Proceedings of the ACL Student Research Workshop, 2005.

21. M. Taboada, J. Brooke, M. To_loski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis”. Computational Intelligence, 2010.

22. D. Bespalov, B. Bai, A. Shokoufandeh, and Y. Qi, “Sentiment Classification Based on Supervised Latent n-gram Analysis,” pp. 375–382, 2011.

23. M. Gamon, “Sentiment classification on customer feedback data: noisy data , large feature vectors , and the role of linguistic analysis.”

24. H. Yu and V. Hatzivassiloglou. “Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical methods in natural language processing, EMNLP '03, pages 129-136.

25. W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Eng. J., May 2014.

AUTHOR(S) PROFILE

Shailesh Kumar Yadav, received the MCA degree from Uttar Pradesh Technical University, Lucknow in 2011. Currently pursuing M.Tech in Computer Science and Engineering from Pondicherry University, Puducherry, India. His research interests are sentiment analysis, opinion mining, data mining, big data, machine learning, natural language processing (NLP).