IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Survey on adoption of Sentiment Analysis for studying consumer's online buying behavior

Bhumika Pahwa[1*], S. Taruna[2], Neeti Kasliwal[3]

[1]Banasthali Vidyapith, Jaipur, Jaipur, Rajasthan, India
[2]JK Lakshmipat University, Jaipur, Rajasthan, India
[3]IIHMR University, Jaipur, Rajasthan, India

[*]*Corresponding Author: pahwa13bhumika@gmail.com, Tel.: 9999310409*

Available online at: www.ijcseonline.org

Accepted: 16/Nov/2018, Published: 30/Nov/2018

***Abstract*—** Sentiment analysis has emerged as a field that has attracted a significant amount of attention since it has a wide variety of applications that could benefit from its results, such as news analytics, marketing, question answering, knowledge management and so on. This area, however, is still early in its development where urgent improvements are required on many issues, particularly on the performance of sentiment classification. Understanding the thoughts of the people is an essential part of the information-gathering behavior. Opinion-rich resources like online review sites and personal blogs have gained immense popularity as they have become easily accessible and are giving new opportunities and posing new challenges as, now people actively use information technology to search out and understand the opinions of others. The flow of interest in the new systems that directly deals with the opinions as a first-class object has given rise to activities in the area of opinion mining and sentiment analysis that work towards the computational analysis of opinions, sentiments and subjectivity in the text. This paper reviews the sentiment analysis methodology and focuses on the techniques to deal with the challenges of sentiment-aware applications. The purpose of this paper is to describe sentiment analysis in detail and to illustrate the method used for it. This survey consists of approaches that work towards enabling opinion-oriented information seeking systems. The main contribution of this paper includes categorization of a number of articles over the years and the illustrations of the recent trends in research in sentiment analysis and its related areas.

***Keywords*—** Opinion mining; Sentiment analysis; Consumer attitude; Sentiment classification

I. INTRODUCTION

Opinion mining (OM) or Sentiment Analysis (SA) is the computational study of people's opinions, emotions and attitude towards an entity. Entity can correspond to an event, topic or individual.

Sentimental Analysis and opinion mining can be used interchangeable as they are seen to express shared meaning, but some researchers have stated that they have slightly different concepts [1]. Our decision making process consists of various factors but one major driving factor has always been "what others think", and this is not a recent upcoming in the decision making process but its importance goes back a long way. Before the widespread of World Wide Web, people generally used to ask their friends and relatives for recommendations on a buying decision like which camera to buy, or which company has best work environment for a fresher or which college has the best infrastructure etc. Internet has made it possible to learn about the experiences

and opinions of the people who are neither well known professional critic nor are they our personal acquaintances. On the other hand, more and more people are making their opinions available online with the help of blogs, tweets etc which are commonly called as online reviews. Hence, In Online shopping it's essential to analyze the data present on the web and so it becomes important to recognize methods which analyses and classifies the online reviews. Sentiment classification is defined as mining and analyzing of reviews, views, emotions and opinions automatically from the texts by using various methods. Opinion mining (OM) or Sentiment Analysis (SA) is the computational study of people's opinions, emotions and attitude towards an entity. Entity can correspond to an event, topic or individual.

Opinion mining extracts and analyzes people's opinion about an entity whereas sentiment analysis identifies the sentiments expressed in a text and analyzes it. Their mutual objective is to find opinions identifying the sentiments expressed in them and then classify them on the basis of their polarity i.e. Positive/negative/neutral.

Sentimental Analysis and opinion mining can be used interchangeable as they are seen to express shared meaning, but some researchers have stated that they have slightly different concepts [1].

Sentiment Analysis can be done at three different levels namely- Document level, sentence level and aspect level [1].

(a) Document level

In this method, the complete document is considered as a single entity and is analyzed as one. Now conventionally, in opinioned texts the users tend to express both the negative as well as positive opinions about the entity and its various features; and hence document level opinion mining is more of a generalized mining technique. A document shown to be positively opinioned about an entity doesn't imply that the user has only positive reviews about all the features of that entity and similarly a negatively opinioned review doesn't imply that the user has only negative reviews about all the features of that entity. Therefore the outcome of this approach is not precise.

(b) Sentence Level

Seeing the generalized approach of document level, in this approach the document is broken into sentences and each sentence is treated as a single entity and analyzed separately.

(c) Aspect level

This approach focuses on aspects i.e. features of the products. The idea behind this approach is to find the sentiments w.r.t different features as a user may have different opinions about different features of items. In this method the polarity for each aspect is determined and this carves an entirely new way to analyze the data. The aspect based opinion mining determines the aspects of every entity that are used for representing the sentiments. This makes this approach to be applicable in various fields like movie reviews, travelling, restaurant reviews and product reviews etc.

SA or OM are the fields that have seen a lot of activity since 2000 and the major reasons behind this is the abundance of social media and its tools like twitter, facebook, LinkedIn, Blogs etc, that have made it convenient to access the information about "how people feel about things." Along with this, various companies and organizations have collected a huge amount of data about how their employees or customers feel about the products and services they received from a particular organization. Zheng-Jun et al [2], extracted the aspects and to gain more precise identification of aspects, they proposed to exploit the pros and cons in the reviews as auxiliary knowledge to assist in identifying the aspects in the free text reviews. To achieve this, they first split the text reviews into sentences and then parse each sentence with Stanford parser. Jeevanandam Jotheeswaran [3], focused on the classification of opinion mining as a corpus based approach where he determined the emotional resemblance of the words for learning the probabilistic affective scores from a large corpus. For lexical resources, he used dictionary based approach such as Word Net to obtain the emotion-related words for classification of emotions. Haseena Rehamath [4] has identified various application areas for which sentiment analysis can be used effectively and those factors include product purchase, market research, policy making, quality improvement, recommendation to other customers, opinion spam detection and decision making. Ravendra Ratan singh Jandail [5] applied used different levels of Sentiment Analysis like document level, sentence level, and entity and aspect level to study positive and negative, interrogative and sarcastic, good and bad functionality. Yajun Deng et al [6] used a database of 28 million e-commerce product reviews and identified dimension mapping and sentiment word disambiguation as the major challenges with the help of a dimension based sentiment analysis model. Svetlana et al [7] used sentiment analysis to detect the sentiment in short informal text messages like tweets and SMS and in the phrases.

II. REASONS FOR USING SENTIMENT ANALYSIS

Anyone who wants to buy a product or is searching for a hotel to stay while touring, they usually check the blogs or reviews made by the consumers who have used the product or availed a particular service. There are three levels of opinion mining at document, sentence and phrase levels [9]. The negative, positive, sarcastic, interrogative, good and bad functionalities, conditional sentences and author and reader understanding viewpoints are studied thoroughly in these levels. The opinions posted by people are retrieved from review websites and are processed through phases like opinion retrieval, opinion classification and opinion summarization about a product or topic. The techniques used in opinion mining are supervised machine learning, unsupervised learning, and case based reasoning.

Online product reviews are done for electronic goods, real estate, movie reviews, automobiles and also Durable consumer goods. For example sentiment for a movie review could be taken from tracking sentiment flow from one sentence to the next [10]. Hanshi wang et al (2014)[11] have done a study on evaluation of effectiveness of sentiment analysis in Chinese language related to customer reviews in three domains i.e., electric devices, E-journal and hotel. They have extracted 1375 documents which consist of both positive and negative sentiments [11].

E-commerce websites are being widely used by customers to get the product reviews and their service feedback, which are used to help customers for buying product and services. Many customers visit such websites for review guidance. Online reviews are of great importance to the consumers and sellers to get the feedback on product and services. As per Ghose et al., (2009)[12] ratings or numberings are more important to review the feedback of the product and services.

According to Ghose et al., (2009) ratings or numberings are more important to review the feedback of the product and services. Ghose and Ipeirotis (2011) describe that user perseverance and influence of sale using online comment is found through readability, subjectivity and linguistic

correctness of text comments which make a difference on it [12].

III. RESEARCH METHODOLOGY

The process of sentiment analysis is explained in figure 1, and the detailed step-wise description is given below-

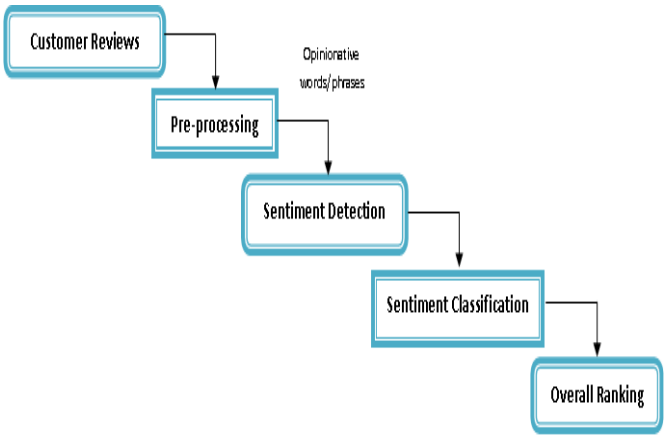

Figure 1: Sentiment Analysis process of online reviews

(a) Customer Reviews: Data Collection

Sentiment analysis takes the benefits of the vast user generated content over the internet. The data source points to queries of user discussions on public forums like blogs, discussion boards and product reviews boards as well as on private logs through social network sites like Twitter and Facebook. Very often, the data log is bulky, disorganized, and disintegrated on multiple portals. Opinions and feelings are expressed in different ways including the amount of details given, type of vocabulary used, context of writing, slangs and lingua variations are just a few examples. This makes manual analysis tedious, and almost impossible. But, with sentiment analysis, innovative text analytics and natural language processing is employed to extract and classify data. Once the data is extracted, it will then be prepared for analysis.

(b) Data Pre-processing or Text Preparation

Text preparation usually involves cleaning the extracted data prior to performing the analysis. Text preparation usually involves identifying and eliminating non textual content from the textual dataset, and any information that can reveal the identities of reviewers including: reviewer name, reviewer location, review date. In addition, any other content that is not deemed relevant to the area of study is also removed from the textual dataset such as includes stop words or words that are not relevant to the course of analysis.

(c) Sentiment detection

Sentiment detection is the third stage of the analysis. This phase requires extracting reviews and opinions from the textual dataset through the use of computational tasks where each sentence is examined for subjectivity. Only the sentences with subjective expressions are kept in the dataset. Sentences that convey facts and objective communication are discarded from further analysis. Detection of sentiments is done at different levels either single term, phrases, complete sentences or complete document with commonly used techniques such as:

- Unigrams: In this approach, each element can be represented as a feature vector on the basis of frequency of the occurrence of a single word. is a classic approach where each element is represented as a feature vector based on frequency of a single word. It is often described as a bag of words approach.
- N-Grams: In this approach the features of a document are represented by multiple words in sequence (e.g.: words in pairs, triplets) which captures more context.
- Lemmas: This involves the use of synonyms rather than the literal word. This method reportedly makes the classification task easier as well as facilitates generalization.
- Negation: This is basically an extension to the n-gram method where the phrases "I like this book" and "I do not like this book" would otherwise be considered similar under majority of classification techniques, but by using negation, both terms are forced into opposite groupings. On the other hand, negation is not always easy to model. For instance, Pang and Lee (2008) [8] explained the difficulty to identify negation when sarcasms and ironies are used in a sentence. Additionally, the negation term does not always reverse the polarity. For example, it will be considered incorrect to attach the word NOT to BEST in the sentence "No wonder this is considered to be the best book".
- Opinion words: These are the words that help in describing people's thoughts and opinions (nouns, verbs, adjectives, adverbs).These words are incorporated into a feature vector where they represent the presence or absence of a word. They help in indicating the subjectivity of a document.

It is not uncommon to find textual sentences making reference to several objects, features and attributes. Through mathematical algorithms, sentiment analysis can be used to extract these objects, features and attributes and categorize them. This assists in the analysis stages and enhances precision in classification and data summarization.

(d) Sentiment classification

Polarity classification is the fourth stage that classifies every subject sentence in the textual dataset into different groups. These groups generally represent the two extreme points on a scale. Although, classification may also involve

multiple points that are similar to start ratings used by hotels, restaurants and retailers.

A wide variety of machine learning techniques are used in binary and polar classification. Machine learning is linked to the field of artificial intelligence and aims at building computational models from past experiences and observation. It fundamentally promotes the use of computer programming to learn and understand fundamentals a particular data set and then use that knowledge acquired to predict or optimize some future criterion.

The three basic algorithms available for classification includes: Naive Bayes (NB), Support Vector Machines (SVM) and Maximum- Entropy (ME).

A Naive Bayes classifier is a probabilistic classifier based on applying Bayes' theorem assuming that features are independent given the class label. This classifier is constructed based on the frequency of occurrence of each feature per class in the training data set.

Support vector machines are based on the statistical learning theory [9]. Binary classifiers show high generalization capability by looking for a hyper plane that maximizes the separation margin between observations from different classes. The use of kernels allows their use for nonlinear problems.

Under Maximum Entropy a number of models are constructed where each feature correspond to a constraint on the model. The model with the maximum entropy over all models is selected for classification.

Although all three classifiers are validated in the literature [8],[10], they require pre-tagged training data or a data corpus which is not always available, or will take a considerable amount of resources both in terms of time and human resources to build. In addition, the language of the data cannot be ignored. Most literature, tools and techniques available on sentiment analysis are written in English language.

(e) Presentation of output

Analysis is generally intended to convert unstructured and fragmented text into information having implied meaning. After completion of the analysis, various options can be used to display the results of text analysis. Graphical displays like pie charts, line graphs and bar charts are the ones that are primarily used with polarity segmentation on color, frequencies, percentage and size.

IV. RESEARCH GAPS

Over the years, sentiment analysis has become very popular and has gained importance in the field of analytics that deals with understanding what people/ users feel about a product or service. A lot of work has been done in this field and this paper describes with the help of table, the different kind of approaches and algorithm used for different purposes.

The table 1 below lists a number of authors and the work done by them, describing the approaches and algorithms used on different types of data for various purposes.

Table 1: Research Gaps

Author	Approach	Algorithm Used	Dataset Used
Pang *et.al.* Pang et al. (2002)	Classify the dataset using different machine learning algorithms and n-gram model	Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM)	Internet Movie Database (IMDb)
Salvetti *et.al.* Salvetti et al. (2004)	Accessed overall opinion polarity(OvOp) concept using machine learning algorithms	Naive Bayes (NB) and Markov Model (MM)	Internet Movie Database (IMDb)
Beineke *et.al.* (2004)	Linearly combinable paired feature are used to predict the sentiment	Naive Bayes (NB)	Internet Movie Database (IMDb)
Mullen and Collier Mullen and Collier (2004)	Values assigned to selected words then combined to form a model for classification	Support Vector Machine (SVM)	Internet Movie Database (IMDb)
Dave *et.al.* Dave et al. (2003)	Information retrieval techniques used for feature retrieval and result of various metrics are tested	*SVM* ^{lite} , Machine learning using Rainbow, Naive Bayes	Dataset from Cnet and Amazon site
Matsumoto *et.al.* Matsumoto et al. (2005)	Syntactic relationship among words used as a basis of document level sentiment analysis	Support Vector Machine (SVM)	Internet Movie Database (IMDb), Polarity dataset
Zhang *et.al.* Zhang et al. (2015)	Use word2vec to capture similar features then classify reviews using *SVM perf*	*SVM* ^{perf}	Chinese comments on clothing products
Liu and Chen Liu and Chen (2015)	Used multi-label classification using eleven state-of-art multi-label, two micro-blog dataset, and eight different evaluation matrices on three different sentiment dictionaries.	Eight different evaluation matrices	Dalian University of Tech. Sentiment Dictionary,DUTSD, National Taiwan Unive Sentiment Dictionary (NTUSD), Howset Dictionary (HD)
Luo *et.al.* Luo et al. (2016)	Ekman Paul's research approach is used to convert the text into low dimensional emotional space (ESM), then classify them using machine learning techniques Ekman and Friesen (1971)	Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT)	Stock message text data(The Lion forum)
Niu *et.al.* Niu et al. (2016)	Transformed the data into required format with the help of Lexicon analysis and then classified the reviewers using statistical learning methods.	BOW feature with TF and TF-IDF approach	Manually annotated Twitter data

IV. OPEN ISSUES

The study illustrated above shows a deeper look in the recent trends of research and while studying the recent articles, some issues were discovered that maybe considered as the open issues in the sentiment analysis research.
The major areas of concern are: Language and Data.

5.1: The Language Issue: It was observed that the eastern languages like Chinese has been used more often recently and accordingly, many sources of data are built for these languages. The researchers are now in the phase of building resources for European language, while there are still not enough resources for middle eastern languages; making it a new trend for research.

5.2: The Data Issue: It has been observed that there is a lack of benchmark datasets in the field of sentiment analysis. The table above illustrates that some major data sources and datasets that were used to carry out different tasks in sentiment analysis. The IMDB and Amazon.com are a few famous data sources of review data. IMDB provides opinioned data on movies while Amazon provides the various products and services reviews.

V. CONCLUSION

This survey paper presents an overview on the recent updates in the sentiment analysis algorithms and applications. Various recently published and cited articles were reviewed and summarized. Articles discussed here contribute to multiple SA related fields that use SA in real world applications. The analysis of the articles clarify that naive bayes and support vector machines are the most commonly used machine learning algorithms for sentiment analysis and they are deemed as a benchmark where proposed algorithms are compared to.

The data and reviews from blogs and forums etc is widely used in sentiment analysis as this media information form a major part in understanding people's opinions about a particular product or services. In various applications, it is important to consider the context of the text and user preferences and that is the reason we need more research on context- based sentiment analysis.

REFERENCES

[1].S. Kajal, P. Vandana, "Opinion Mining: Aspect level sentiment analysis using sentiwordNet and Amazon Web Services" , International Journal of Computer Applications, Volume 158-No 6, January 2017.

[2].D. Yajun, Z. Lizhou , J. Peiquan , "Dimension based Sentiment polarity detection for E-Commerce Reviews", Advanced Science and Technology Letters Vol.45 (CCA 2014), pp.55-59,http://dx.doi.org/10.14257/astl.2014.45.11

[3].J. Jeevanandam, Dr. S. Koteeswaran, "Sentiment Analysis: A Survey of Current Research and Techniques", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 5, May 2015

[4].R. Haseena, "Opinion mining and sentiment analysis- Challenges and application", International Journal of Application on Innovation in Engineering and Management, 401-403 (2014)

[5].S.J. Ravendra Ratan Singh, "A proposed Novel approach for Sentiment Analysis and opinion Mining", International Journal of UbiComp (IJU), 5, (.1/2)April 2014 Pp:1-10

[6].D. Yajun, Z. Lizhou , J. Peiquan , "Dimension based Sentiment polarity detection for E-Commerce Reviews", Advanced Science and Technology Letters Vol.45 (CCA 2014). pp.55-59, http://dx.doi.org/10.14257/ astl.2014.45.11

[7].K. Svetlana, Z. Xiaodan and M. Saif, "Sentiment analysis of short informal texts", Journal of Artificial Intelligence Research,50:723-762, 2014(b)

[8].B. Pang, L. Lee, "Opinion Mining and Sentiment Analysis", Foundation and Trends in Information Retrieval. 2(1-2): 1-135.

[9].C. Corinna & V. Vladimir, "Support Vector networks", Kluwer Academic publisher, Boston, 1995

[10].S.R. Bhanu , J. Uma Pricilda, "A Review on the concept of sentiment analysis and its role in marketing strategies for E-commerce' ,IIOAB Journal' 216-224, 2014

[11].B.Pang,L. Lee, and S. Vaithyanathan , "Thumbs up?: sentiment classification using machine learning techniques.",In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 79–86,2002.

[12].F.Salvetti, S. Lewis & C. Reichenbach, "Automatic opinion polarity classification of movie", Colorado research in linguistics, 2004.

[13].P. Beinik, T. Hastic, & S.Vaithyananthan, " The sentimental factor: Improving review classification via human-provided information", In proceedings of the 12th annual meeting on association for computational linguistics, page 263, association for computational linguistics, 2004.

[14].T.Mullen & N. Collier, "Sentiment analysis using support vector machines with diverse information sources", in EMNLP, Vol 4, pg 412-41,2004

[15].K. Dave, S. Lawrence , & D. Pennock , "Mining the oeanut gallery: Opinion extraction & semantic classification of product reviews", In proceedings of the 12th international conference on world wide web, pg 519-528, ACM,2003

[16].S. Matsumoto, H.Takamura, & M. Okumara, "Sentiment Classification using word sub-sequences & dependency sub-trees", In advances in knowledge discovery & data mining, pages 301-311, Springer,2005

[17].B. Luo, J. Zeng & J. Duan, "Emotion space model for classifying opinions in stock message board", Expert Systems with Applications, 44 , 138–146 , 2016.

[18].S. M Liu, & J.H. Chen, "A multi-label classification based approach for sentiment classification", Expert Systems with Applications, 42 (3), 1083–1093, 2015.

[19].T. Niu, S. Zhu, L. Pang & A. El Saddik, "Sentiment analysis on multi-view social data" In Multimedia modelling (pp. 15–27). Springer, 2016.

[20].P. Haiyun, C. Erik, H. Amir, "A Review of sentiment analysis research in Chinese language", Cognitive Computing (2017) 9: 423. https://doi.org/10.1007/S125590-17-9470-8.

Author Profiles

Bhumika Pahwa, PhD Scholar from Banasthali Vidyapith, Jaipur is working in the field of Sentiment analysis. She is working as assistant professor at DAV Institute of Management affiliated to MDU Rohtak. She has three years of teaching experience. She completed her M.Tech in Information technology from Banasthali Vidyapith in year 2014. She has several publications to her credit and has presented research papers at National and International conferences and journals.

Dr S.Taruna, the Associate Professor of Institute of Engineering and Technology (IET) at JK Lakshmipat University has 20 years of teaching, research, and administrative experience. Prior to joining JKLU she has worked with Banasthali University for 12 years and CIStems Software Ltd for 8 years. She has also been associated with other colleges like Subodh PG College, Jaipur as a guest faculty. She did her PhD from Banasthali University and currently supervising PhD candidates working in the domain of Communication Network, Data Mining and Cloud Network. Under her supervision one of the students has been awarded her PhD degree in Computer Science & Engineering. Besides teaching & research, she coordinated M.Tech research projects and project placements & training of M.Tech , MCA students at Banasthali University. She has several publications to her credit and has presented research papers at National and

International conferences and journals . She is Reviewer and Committee Member of various International Journals and Conferences.

Dr. Neeti Kasliwal, is an Associate Professor at School of Pharmaceutical Management, Indian Institute of Health Management Research University Jaipur. Her areas of interest are Pharmaceutical Marketing, Consumer Behavior, Advertising Management, Green Marketing and Online Buying Behavior. I have 17 years of teaching and administrative experience. She is also a PhD Supervisor and believes in quality of good work. Currently engaged in research areas of Green Buying Behavior, E-CRM practices of Banks and CSR advertising and its impact of Consumer Behavior.

