Sentiment Analysis: Classification, Approaches, Applications and Challenges

Namisha Mahajan[1], Satinder Pal Ahuja[2], Hardeep Singh Saini[3]
[1]*Asstt. Professor, CSE Department, Indo Global College of Engineering, Mohali, India*
[2]*Professor, Indo Global College of Management and Technology, Mohali, India*
[3]*Professor, Indo Global College of Engineering, Mohali India*

Abstract **-** Sentiment analysis, also called opinion mining, is the field of study that analyses people's opinions, sentiments, evaluations, assessments, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. This field is identified by many different names subject to marginally different tasks performed for example sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. However, all these fields are now represented under the umbrella of sentiment analysis or opinion mining. This paper presents a survey on the Sentiment Analysis applications and challenges with their approaches and techniques.

Keywords **-** *Sentiment analysis, opinion mining, sentiment classification, machine learning approaches, lexicon based approach, polarity*

I. INTRODUCTION

Sentiment refers to feelings, response, reaction, emotions, opinion, outlook or attitude. With the speedy growth of World Wide Web, people time and againshare their knowledge, their thoughts and express their sentiments over internet through social media, forums, tweets, blogs, ratings, reviews and so on. This way of expression has led to radical change in the way in which people communicate and create impact on social, political and economic behaviour of other people. The changing aspects of Web 2.0 allows everyone to boost human collaboration capabilities on global level by enabling every individual to share opinions by means of commenting on published content and user generated content. Due to this increase in the textual data and user generated opinions, there is huge demand by companies, service providers, researchers and politicians to analyse the sentiment of the opinion in order to implement better decision choices to discover new business strategies and advertising campaign. The term sentiment analysis and opinion mining can be used interchangeably.[1][2] Sentiment analysis has a profound impact on linguistics, Natural language processing (NLP), management sciences, political sciences, economics, social sciences and businesses as every domain is affected by people's opinion.

Steps in sentiment analysis process: Sentiment analysis is a complex process. The steps encompassed in sentiment analysis [6][7] can be explained by a flowchart.

Identifying and categorizing data sources
In this step identification and classification of data is done on the basis of data collected from user generated content contained in micro-blogging sites, forums and social networks.

Text pre-processing and preparation
Initially data collected is unorganized and is expressed in different ways by using different vocabularies, slangs, context of writing etc. Therefore, it is required to extract and classify meaningful data using text analytics and NLP. This extracted data is further cleaned before analysis using this step. Any non-textual content and unimportant content is identified and eliminated.

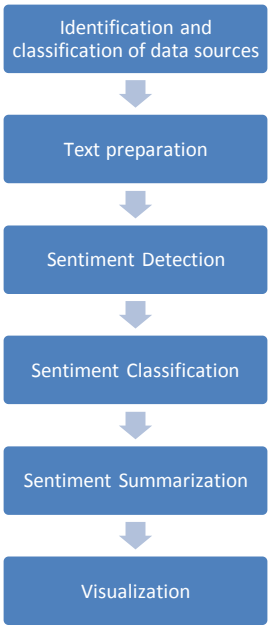

Fig.1: General workflow for the Sentiment Analysis Process

Sentiment Detection
Extracted data is examined. Sentences with subjective expressions (opinions, beliefs, reviews and views) are held. However those with objective communication (facts, factual information) are rejected.

Sentiment Classification
Classification of sentences with subjective expressions is done as positive, negative or neutral using this step.

Sentiment summarization
In this step, the sentiments are summarized into collective scores for positive and negative orientation along with relevant fragments.

Visualization
When the analysis and scoring is finished, the text results are displayed on graphs like pie chart, bar chart and line graphs. Time can also be analysed and displayed realistically with the use of graphs by constructing a sentiment time line with the chosen value (frequency, percentages, and averages) over time.

II. SENTIMENT CLASSIFICATION

Sentiment Classification is the task of classifying a target unit of text in a document to positive, negative or neutral class. On the basis of granularity level it can be performed at three levels.

Document Level
At this level whole document is considered to be a basic information unit. Objective at this level is to classify the target opinion document as expressing positive or negative sentiment about the single entity only.A document with multiple entities leads to a problem of varying sentiments. Due to this issue analysis at this level is not applicable to documents evaluating more than one entity.

Sentence Level
This level focuses on sentences and tries to determine its sentiment. This level of analysis is closely related to subjectivity classification which distinguishes objective sentences (sentences with factual information) from subjective sentences (sentences with subjective views). After distinguishing, sentiment is determined for the subjective sentences. At this level problem may arise as even some objective sentences tend to contain sentiment.

Aspect Level
At this level the targets are discovered on which opinions have been expressed in a sentence. Targets can be an object, its features, components and various attributes. Sentiment classification at this level identifies the sentiment of the opinions about a particular target. For example, the sentence "although junk food is not good for health, but I still love street food" clearly has a positive tone, we cannot say that this sentence is entirely positive. In fact, the sentence is positive about the street food (emphasized), but negative about junk food (not emphasized).

III. CLASSIFICATION OF SENTIMENT ANALYSIS APPROACHES

There are three approaches for sentiment classification [9]. These techniques are:

Machine learning approach: Machine Learning (ML) provides solutions to critical applications which includes data mining, natural language processing, image recognition, and expert systems. It uses several learning algorithms to determine the sentiment by training on a known dataset. It performs the supervised or semi-supervised learning by extracting the features from the text and learns the model[10][11]. The machine learning approach predicts the polarities of sentiments on the basis of trained as well as test data sets. It applies the ML algorithms and uses linguistic features. Supervised approach is used when there is a predictable set of classes (positive and negative). This method needs labelled data to train classifiers[3]. In a machine learning based classification a training set is used by an automatic classifier to learn the different characteristics of documents, and a test set is used to validate the performance of the automatic classifier. The unsupervised methods are used when it is difficult to find labelled training documents. Unsupervised learning can be applied as no prior training is essential in order to mine the data. Unsupervised approaches applied at document-level are based on determining the semantic orientation of explicit phrases within the document. The document is classified as positive if the average semantic orientation of these phrases is above some predefined threshold otherwise it is deemed negative.
The benefit of supervised method is its ability to adjust and create trained models for definite purposes and contexts. This method is rarely used for new data as availability of labelled data is pre-requisite. And providing that can be expensive or even prohibitive. This is the main disadvantage. Commonly used machine learning approaches are:

(i) Bayesian Networks: it is a probabilistic approach that represents a set of random variables and their conditional dependencies by directed acyclic graph. In DAG nodes are variables and arcs represent the dependence between variables. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.
(ii) Naive Bayes Classification: It is an approach particularly suited when the dimensionality of the inputs is high. Despite its simplicity, it can often outperform more sophisticated classification methods. It requires small number of training data to estimate the parameters necessary for classification.
(iii) Maximum Entropy: This method is mostly used as alternatives to Naive Bayes classifiers because it does not assume statistical independence of the random variables (features) that serve as predictors. The principle behind Maximum Entropy is to find the best probability distribution among prior test data.
(iv) Neural Networks: This model is based on a collection of natural/artificial neurons uses for mathematical and computational model analysis
(v) Support Vector Machine: It is a supervised learning model which analyses data and patterns that can be used

for classification and regression analysis. The basic idea behind this is to find a maximum margin hyper plane represented by vector. It finds an optimal solution.[7]

The main advantage of machine learning approaches is the ability to adapt and create trained models for specific purposes and contexts. The limitation is that it is difficult to integrate into the classifier, general knowledge which may not be acquired from training data. Because of the reliability of machine learning approaches on domain specific features from the training data, they have poor adaptability to different text genres and domains.

Lexicon based approach: Involves calculating sentiment polarity of a review on the basis of the semantic orientation of phrases, idioms, words or sentences in the review. The "semantic orientation" is a measure of subjectivity and opinion in text[10].This approach uses a predefined list of words with a specific sentiment attached. They are based on the counting of positive and negative words. These methods vary according to the context in which they were created. This approach doesn't need labelled data. It is difficult to create a unique lexical-based dictionary which can be used in different contexts. For example slang used in social networking sitesdoesn't find place and support in lexical methods.

Among the lexicon-based approaches the most used are:

(i) Dictionary based approach: In this approach every word is translated according to its dictionary meaning without laying emphasis on the correlation between consecutive words.
(ii) Novel Machine Learning Approach: it integrates important linguistic features into automatic learning
(iii) Corpus based approach: This is applied in the following two scenarios: (1) given a seed list of known (often general-purpose) sentiment words, discover other sentiment words and their orientations from a domain corpus, and (2) adapt a general-purpose sentiment lexicon to a new one using a domain corpus for sentiment analysis applications in the domain. However, the issue is more complicated than just building a domain specific sentiment lexicon because in the same domain the same word can be positive in one context but negative in another[8]. It has been widely used to explore both written and spoken texts in order to assign a sentiment factor of words that depend on frequency of their occurrences
(iv) Ensemble Approaches in sentiment classification: it increases classification accuracy by combining arrays of specialized learners.

Lexicon-based approaches have the advantage that commonly used sentiment lexicons have wider term coverage, but with two main limitations. Firstly, finite words in the lexicon create a problem while extracting sentiment from dynamic environments. Secondly, a fixed sentiment orientation and score to words is assigned, irrespective of the way these words are used in the text.

Hybrid Approach: In this both machine learning and lexicon based approaches are combined together to increase the performance of the sentiment classification.The main advantages of hybrid approaches are the lexicon/learning symbiosis, the detection and measurement of sentiment at the concept level and the lesser sensitivity to changes in topic domain[7]. The main limitation is that noisy reviews are often assigned a neutral score because of the failure to detect any sentiment.

IV. APPLICATIONS

Opinions are central to almost all human activities because they are key influencers of our behaviours. Whenever we need to make a decision, we want to know others' opinions[8].Sentiment analysis is used in various domains for different reasons. Some of the most important applications are:

Online Commerce: Users of different online commerce sites can post their views about their shopping experiences and quality of the products purchased. A brief review is provided and products are rated on the basis of numerous features. Other customers can make up their mind regarding the product by viewing opinions and recommendation information provided. Graphical summary of the overall product and its features is also presented to users. Popular merchant websites like Paytm.com, myntra.com provides review from customers, sellers, business partners, editors with rating information. For example http://careers360.com is a popular website that provides reviews on Universities and Colleges, Courses through current students, alumni. They contain numerous opinions and reviews worldwide. Sentiment analysis helps such websites by converting dissatisfied customers into promoters by analysing this huge volume of opinions.[12]

Marketing: Companies use sentiments to develop their business and marketing strategies, understand customers' views for products or brand, reaction of people to new product launches and reasons for disinterest of customers in some products.

Politics: Sentiment analysis is very important in the field of politics. It is used to track political view, detect consistency and inconsistency between statements and actions carried at the government level.In politics, we can analyse trends, predict election results, identify ideological bias, target advertising/messages accordingly and evaluate public/voter's opinions.[7][12]

Social Sciences: Sentiment analysis is used to monitor and analyse social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

Recommendation Systems: Sentiment analysis can also be helpful for recommendation systems which will not recommend something that receives negative feedback, and for the development of new kinds of search engines.[13]

Text Processing: A large number of text processing applications have already employed techniques for automatic sentiment analysis, for example automatic expressive text-to-speech synthesis, text semantic analysis, tracking sentiment timelines in online forums and news, mining opinions from movie reviews, and question answering.[13]

Future related projects are paedophilia and suicide detection on the Internet (e.g. on discussion boards, social networking websites, chat-rooms, etc.).

In all of these examples, the sentiment input is whether a given consumer opinion has negative, positive or neutral polarity regarding the different target of interest.

V. CHALLENGES

There are several challenges in opinion mining.

- **Co-reference resolution:** This challenge addresses the problem of identifying what a pronoun or noun phrase is referring to. For e.g. "We watched the movie and went to dinner; it was awful". Now this sentence is ambiguous as it is unclear what does "It" refers to in the above sentence?
- **Temporal relations:** Time of reviewing is important for sentiment analysis. The reviewer may think that mobile phones with keypad were good when they were released but now in 2017, he may have negative opinion because of wide range of options available in smart phones with touch screens. Assessment of the opinions that change with time improves the performance of the sentiment analysis system. This helps us to observe if a certain product gets improved with time, or people change their opinion about a product.[12]
- **Review Spam Detection**: Fake reviews corresponding to a particular product on product review sites are called review spams. These are written to promote their products by giving undeserving positive opinions, or defame their competitors' products by giving false negative opinions. The opinion spam identification task has great impacts on industrial communities. If the opinion providing services contain large number of spams, they will affect the user's experience and if the user is cheated by the provided opinion, he will never use the system again.[12]
- Another important challenge is that same word may be positive in one situation and negative in another. Take the word "" for instance. If a customer said a laptop's battery life was long, then that would be a positive opinion. If the customer said that the laptop's start-up time was long, however, that would be a negative opinion.
- A sentence made up of sentiment words but unable to express any sentiment. Such issues arise in case if question is being asked or there is a conditional statement. e.g., "Can you tell me which smart watch is good?" and "If I can find a good watch in the store, I will buy it." Both these sentences contain the sentiment word "good", but neither expresses a positive or negative opinion on any specific watch.
- Sentiment words need lot of attention in sentences with sarcasm, irony and implications. They make it difficult to analyse the polarity of the statement. Such attention is required while analysing the polarity of tweets and blogs related to politics.
- People don't always express opinions the same way. Most traditional text processing relies on the fact that small differences between two pieces of text don't change the meaning very much. In opinion mining, however, "the movie was great" is very different from "the movie was not great".
- Sentences carrying sentiments/an opinion but no sentiment words. These are the sentences that carry factual information e.g. "This smart phone consumes lot of battery", implies negative sentiment about the sentence.
- Finally, people can be contradictory in their statements. They tend to include both positive and negative comments. This can be managed to a certain extent by analysing sentences one at a time. However, the more informal the medium (twitter tweets or blog posts for example), the more likely people are to combine different opinions in the same sentence. For example: "the movie bombed even though the lead actor rocked it" is easy for a human to understand, but more difficult for a computer to parse.

VI. CONCLUSION

The data is growing at such a pace that it is almost infeasible to retrieve all the important information. But this can be achieved through sentiment analysis. Applying sentiment analysis to mine the huge amount of data has become an important research problem. This paper starts by describing the concept of sentiment analysis. It provides an insight into the steps in the sentiment analysis process, lists the various levels at which sentiment classification is done on the basis of granularity, deeply explains the sentiment classification approaches namely (i) machine learning (ii) lexicon based and (iii) hybrid approach. Although, some of the algorithms used in sentiment analysis gives good results, but still no algorithm is able to address all the challenges. Most of the researchers agree that Support Vector Machines (SVM) has high accuracy than other algorithms, but with some limitations. Sentiment classification is domain dependent. In order to enhance the performance of sentiment classification different algorithms that can work in collaboration with each other should be applied. This

paper also focuses on various application areas and challenges that researchers have to face while mining out the sentiments. Challenge to overcome the ambiguity in a particular problem so that it is easy to use co-reference information. The analysed posts contain irony and sarcasm, which are particularly difficult to detect. So an evolution of much better approaches and tools is required to overcome such kind of limitations.

VII. REFERENCES

[1]. T. Nasukawa and J. Yi, "Sentiment Analysis : Capturing Favorability Using Natural Language Processing," Proc. 2nd Int. Conf. Knowl. capture, pp. 70–77, 2003.
[2]. K. Dave, K. Dave, S. Lawrence, S. Lawrence, D. M. Pennock, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," Proc. 12th Int. Conf. World Wide Web, pp. 519–528, 2003.
[3]. B. Pang, L. Lee, H. Rd, and S. Jose, "Sentiment Classification using Machine Learning Techniques," Conf. Empir. Methods Nat. Lang. Process., pp. 79–86, 2002.
[4]. J. M. Wiebe, "Learning subjective adjectives from corpora," Proc. Natl. Conf. Artif. Intell., no. 1, pp. 735–741, 2000.
[5]. P. D. Turney, "Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews," Proc. 40th Annu. Meet. Assoc. Comput. Linguist., no. July, pp. 417–424, 2002.
[6]. M. Godsay, "The Process of Sentiment Analysis: A Study," Int. J. Comput. Appl., vol. 126, no. 7, pp. 26–30, 2015.
[7]. A. D. Andrea, F. Ferri, and P. Grifoni, "Approaches , Tools and Applications for Sentiment Analysis Implementation," vol. 125, no. 3, pp. 26–33, 2015.
[8]. B. Liu, "Sentiment Analysis and Opinion Mining," Sentim. Anal. Opin. Min., 2012.
[9]. D. Maynard and A. Funk, "Automatic detection of political opinions in tweets," CEUR Workshop Proc., vol. 718, pp. 81–92, 2011.
[10]. C. Costea, D. Joyeux, O. Hasan, and L. Brunie, "A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation," 2012.
[11]. A. Kaur and V. Gupta, "A survey on sentiment analysis and opinion mining techniques," J. Emerg. Technol. Web Intell., vol. 5, no. 4, pp. 367–371, 2013.
[12]. U. Bhattacharjee, R. Hills, and A. Pradesh, "Applications and Challenges for Sentiment Analysis : A Survey," vol. 2, no. 1, pp. 1–6, 2013.
[13]. S. Schrauwen, "Machine Learning Approaches to Sentiment Analysis Using the Dutch Netlog Corpus," Comput. Linguist. Psycholinguist. …, 2010.