e-ISSN: 2349-9745
p-ISSN: 2393-8161

Scientific Journal Impact Factor (SJIF): 1.711

International Journal of Modern Trends in Engineering and Research

www.ijmter.com

Sentiment Analysis in Hindi Language : A Survey

Sumit Kumar Gupta[1], Gunjan Ansari[2]
[1]*Dept. of Computer Science and Engineering, JSSATE Noida*
[2]*Dept.of Information Technology, JSSATE Noida*

Abstract— With recent development in web technologies and mobile technologies, with increasing user-generated content in Hindi on the internet is the motivation behind the sentiment analysis Research that is growing up at a lightning speed. This information can prove to be very useful for researchers, governments and organization to learn what's on public mind, to make sound decisions. Opinion Mining or Sentiment Analysis is a natural language processing task that mine information from various text forms such as reviews, news, and blogs and classify them on the basis of their polarity as positive, negative or neutral. But, from the last few years, enormous increase has been seen in Hindi language on the Web. Research in opinion mining mostly carried out in English language but it is very important to perform the opinion mining in Hindi language also as large amount of information in Hindi is also available on the Web. This paper gives an overview of the work that has been done Hindi language.

Keywords- Natural Language Processing, Opinion Mining, Sentiment Analysis, Indian Language, Machine Learning

I. INTRODUCTION

Text mining is a sub area of the study of natural language processing that relates to understanding and generating the human languages such as English, French, Japanese, and Hindi etc. Most of the work in Opinion mining has been done in English language; Very little attention has been paid in the direction of sentiment analysis for other languages. As the internet is reaching to more and more people within the world, there is tremendous increase in Web content of other languages because people feel comfortable with their native language. Hindi is the 4th largest spoken language and has 490 million speakers across the world majority of whom are from India (www.wikipedia.org).There are many websites which provide information in Hindi, ranging from various news websites such as (] http://dir.hinkhoj.com/, http://bbc.co.uk/hindi) for sites providing information regarding the culture, music, entertainment and other aspects of arts (http://www.webdunia.com/, http://www.virarjun.com/, http://www.raftaar.in/). There are various weblogs which are written in Hindi. Large amount of Hindi content is available on the web, so it is necessary to mine this large amount of information and extracting the opinions from this data which helps the users and the organizations in taking decisions.
Sentiment Analysis is a natural language processing task that deals with finding orientation of Opinion in a piece of text with respect to a topic. It focuses on categorizing the text at the level of subjective and objective nature. Subjectivity indicates that the text contains/bears opinion content whereas Objectivity indicates that the text is without opinion content. Some examples-Subjective- शाहरूख और काजौल की यह फिल्म अच्छी है । (this sentence has an opinion, it talks about the movie and the writer's feelings about "**अच्छी**" and hence it's subjective). The subjective text can be further categorized into 3 broad categories based on the sentiments expressed in the text.

1. **Positive**- यह फिल्म अच्छी है |

2. **Negative**- यह होटल बहुत खराब है।

3. **Neutral**- मुझे दोपहर तक भूख लगने लगती है | (this sentence has user's views, feelings hence it is subjective but as it does not have any positive or negative polarity so it is neutral.)

II. DATA SOURCE

People and companies across disciplines exploit the rich and unique source of data for varied purposes. The major criterion for the improvement of the quality services rendered and enhancement of deliverables are the user opinions. Blogs, review sites and micro blogs provide a good understanding of the reception level of products and services.

A. Blogs

The name associated to universe of all the blog sites is called blogosphere. People write about the topics they want to share with others on a blog. Blogging is a happening thing because of its ease and simplicity of creating blog posts, its free form and unedited nature. We find a large number of posts on virtually every topic of interest on blogosphere. Sources of opinion in many of the studies related to sentiment analysis, blogs are used [1].

B. Review Sites

Opinions are the decision makes for any user in making a purchase. The user generated reviews for products and services are largely available on internet. The sentiment classification uses reviewer's data collected from the websites like www.gsmarena.com (mobile reviews), www.amazon.com (product reviews), www.CNETdownload.com (product reviews), which hosts millions of product reviews by consumers [2].

C. Micro-blogging

A very popular communication tool among Internet users is micro-blogging. Millions of messages appear daily in popular web-sites for micro-blogging such as Twitter, Tumblr, Facebook. Twitter messages sometimes express opinions which are used as data source for classifying sentiment. [3].

III. SENTIMENT CLASSIFICATION

Sentiment classification or Polarity classification is the binary classification task of labeling an opinionated document as expressing either an overall positive or an overall negative opinion. A technique for analyzing subjective information in a large number of texts, and many studies is sentiment classification. A typical approach for sentiment classification is to use machine learning algorithms.

A. Machine Learning

A system capable of acquiring and integrating the knowledge automatically is referred as machine learning. The systems that learn from analytical observation, training, experience, and other means, results in a system that can exhibit self-improvement, effectiveness and efficiency. Knowledge and a corresponding knowledge organization are usually used by a machine learning system to test the

knowledge acquired, interpret and analyze. One of the machine learning algorithms is taxonomy based depending on outcome of the algorithm or type of input available.

- Supervised learning generates a function which maps inputs to desired outputs also called as labels because they are training examples labeled by human experts. Since it is a text classification problem, any supervised learning method can be applied, e.g., Naïve Bayes classification, and support vector machines.
- Unsupervised learning models a set of inputs, like clustering, labels are not known during training. Classification is performed using some fixed syntactic patterns which are used to express opinions. The part-of-speech (POS) tags are used to compose syntactic patterns.
- Semi-supervised learning generate an appropriate function or classifier in which both labeled and unlabelled examples are combined [4].

B. Sentiment Analysis Task

Sentiment analysis tasks mainly consists of classifying the polarity of a given text at the document, sentence or feature/aspect level expressing the opinion as positive, negative or neutral. The sentiment analysis can be performed at one of the three levels: the document level, sentence level, feature level

- **Document Level Sentiment Classification**: In document level sentiment analysis main challenge is to extract informative text for inferring sentiment of the whole document. The learning methods can be confused because of objective statements are rendered by subjective statements and complicate further for document categorization task with conflicting sentiment [5].
- **Sentence Level Sentiment Classification:** The sentiment classification is a fine-grained level than document level sentiment classification in which polarity of the sentence can be given by three categories as positive, negative and neutral. The challenge faced by sentence level sentiment classification is the identification features indicating whether sentences are on-topic which is kind of co-reference problem [5]
- **Feature Level Sentiment Classification:** Product features are defined as product attributes or components. Analysis of such features for identifying sentiment of the document is called as feature based sentiment analysis. In this approach positive or negative opinion is identified from the already extracted features. It is a fine grained analysis model among all other models [2]

C. Component of Opinion Mining

There are mainly three Component of opinion Mining, these are

- **Opinion Holder:** Opinion holder is the holder of a particular opinion; it may be a person or an organization that holds the opinion. In the case of blogs and reviews, opinion holders are those persons who write these reviews or blogs.
- **Opinion Object:** Opinion object is an object on which the opinion holder is expressing the opinion.
- **Opinion Orientation:** Opinion orientation of an opinion on an object determines whether the opinion of an opinion holder about an object is positive, negative or neutral.

For example"इस मोबाइल का कैमरा अच्छा है |". In this review, the person who has written this review is the Opinion Holder. Opinion object here is the कैमरा of the mobile and the opinion word is "अच्छा" which is positively orientated. Semantic orientation is a task of determining whether a sentence has either positive, negative orientation or neutral orientation.

IV. EXISTING RESEARCH WORK

There is a growing interest in visualizing sentiments from Web posts and related content.

In the field of opinion mining a small amount of work has been done in Hindi language. The very first research has been done in Hindi, Bengali and Marathi language. Amitava Das and Bandopadhya [7], developed sentiwordnet for Bengali language. Word level lexical-transfer technique have been applied to each entry in English SentiWordNet using an English-Bengali Dictionary to obtain a Bengali SentiWordNet. 35,805 Bengali entries has been returned by their experiment. Using the lexicon and features like positional aspect, a supervised classifier is generated. This classifier achieves a precision of 74.6% and a recall of 80.4%.

Nishantha et al. [13] proposed approach for sentiment Classification in non-English language like Hindi, Russian and Chinese. Comparisons were conducted within and among the languages. The authors claimed that the wordnet inability to perform the word sense disambiguation is a measure limitation of the proposed algorithm.

Das and Bandopadhya [8], proposed four strategies to predict the sentiment of a word. In the first approach, an interactive game is proposed by them which turn annotated the words with their polarity and validate SentiWordNet for Indian Languages. In Second approach, they used corpus Based approach to increase the coverage of the developed SentiWordNet(s) and to capture the language/culture specific words. In third Approach to overcome the limitation and increase the coverage of the SentiWordNet(s) they present automatic antonym generation technique SentiWordNet(s) and the forth A word-level translation process followed by error reduction technique has been adopted for generating the Indian languages SentiWordNet(s) from the English sentiment Lexicon.

Richa Sharma et. al. [14] proposed a unsupervised dictionary approach that determine the polarity of user Review in Hindi language. Hindi dictionary created by them contains the most frequently used Hindi words and its synonym and antonyms. Their approach is well in this domain and achieved the accuracy of 65%.

Joshi et al. [9] proposed a fallback strategy for Hindi language. This strategy follows three approaches: In-language Sentiment Analysis, Machine Translation and Resource Based Sentiment Analysis. They developed a lexical resource, Hindi SentiWordNet (HSWN) based on its English format. By using two lexical resources (English SentiWordNet and English-Hindi WordNet Linking) H-SWN (Hindi-SentiWordNet) was created by them. By using Wordnet linking, words in English SentiWordNet were replaced by equivalent Hindi words to get HSWN. The final accuracy achieved by them is 78.14.

By using a graph based method Bakliwal et al.[10]created lexicon .They determine that by using simple graph traversal how the synonym and antonym relations can be used to generate the subjectivity lexicon. Their proposed algorithm achieved approximately 79% accuracy on classification of reviews and 70.4% agreement with human annotated.

Namita mittal et al.[11]developed an efficient approach based on negation and discourse relation to identifying the sentiments from Hindi content .They developed an annotated corpus for Hindi language and improve the existing Hindi SentiWordNet (HSWN) by incorporating more opinion words into it. Then they devised the rules for handling negation and discourse that affect the sentiments expressed in the review. Their proposed algorithm achieved approximately 80% accuracy on classification of reviews.

Piyush Arora et al. [15] proposed a graph based method to build a subjective lexicon for Hindi language, using WordNet as a resource. They build a subjective lexicon for Hindi language with dependency on WordNet. They initially build small seed list of opinion words and by using WordNet, synonyms and antonyms of the opinion words were determined and added to the seed list .They traverse Wordnet like a graph where every word in a Wordnet considered as a node, which is

connected to their synonyms and antonyms. They achieved 74% accuracy on classification of reviews and 69% accuracy is achieved in agreement with human annotators for Hindi.

Rao and Ravichandran [16] presented an extensive study on the problem of detecting polarity of words. They considered bi-polar classification of words i.e. a word can be either positive or negative. They performed semi-supervised label propagation in graph for polarity detection of words. Each of these words represent a node in the graph whose polarity is to be determined. They focused on three languages mainly English, French and Hindi but claim that their work can
be extended to any other language for which WordNet is available

Mukherjee et al. [17] showed that the incorporation of discourse markers in a bag-of-words model improves the sentiment classification accuracy by 2 - 4%. Bakliwal et al. [5] proposed a method to classify Hindi reviews as positive or negative. They devised a new scoring function and test on two different approaches. They also used a combination of simple N-gram and POS Tagged N-gram approaches.

Ambati et al. [18] proposed a novel approach to detect errors in the treebanks. This approach can significantly reduce the validation time. They tested it on Hindi dependency treebank data and were able to detect 76.63% of errors at dependency level.

Harshada Gune et al. [19] perform shallow parsing on Marathi language. They build a Marathi shallow parser which consists of Marathi POS tagger and Chunker. In their proposed system, morphological analyzer provides ambiguity and suffix information for generating a rich set of features. Generated features are then applied to the CRF based engine which couples them with other elementary features for training a sequence labeller. Verb Group Identifier (VGI) used by the POS tagger for correcting the output of the CRF based sequence labeller. 50% accuracy is achieved by their system.

V. CHALLENGES

Challenges while dealing/working with Hindi language are as follows [12]

I. **Word Order-** Word arrangement in a sentence plays an important role in identifying the subjective nature of the text. Hindi is a free order language i.e. the subject, object and verb can come in any order whereas English is a fixed order language i.e. subject followed by a verb and followed by an object. Word order plays a vital role in deciding the polarity of a text, in the text same set of words with slight variations and changes in the word order affect the polarity aspect.

II. **Morphological Variations-** Handling the morphological variations is also a big challenge for Hindi language. Hindi language is morphologically rich which means that lots of information is fused in the words as compared to the English language where we add another word for the extra information.

III. **Handling Spelling Variations-** In the Hindi language, the same word with same meaning can occur with different spellings, so it's quite complex to have all the occurrences of such words in a lexicon and even while training a model it's quite complex to handle all the spelling variants.

IV. **Lack of resources-** the lack of sufficient resources, tools and annotated corpora also adds to the challenges while addressing the problem of sentiment analysis especially when we are dealing with Non-English languages.

V. **Co reference resolution-** It is a problem of determining multiple expressions that refer to the same thing. For example "राम ने खाना खाया और वह सोने चला गया" "वह" in the second sentence refers राम that is an entity. It is important to recognize these coreference relationships for aspect-based sentiment analysis.

VI. CONCLUSION

Sentiment Analysis/ Opinion Mining is an emerging research field and is very important because human beings are largely dependent on the web nowadays. These days web browser are also provide information in Hindi and several other language one can see write and search the information in Hindi on these search engines so Opinion Mining in Hindi language is required. The rise in user-generated content in Hindi language across various genres- news, culture, arts, sports etc. has open the data to be explored and mined effectively, to provide better services and facilities to the consumers. Opinion Mining has large application areas like Shopping, where websites like Snapdeal.com, flipkart.com, amazon.com etc. Allow customers to express their opinions on their websites which helps other customers to decide whether to buy the product or not. Entertainment, where the people can easily see the critics and viewer's reviews of their favourite movies and shows online. Marketing, now every organization and private firms allow their customers to write the reviews related to their products on their websites which eliminate the needs to conduct surveys.

Large amount of work in opinion mining has been done in English language, as English is a global language, but there is a need to perform opinion mining in other languages also. Large amount of other languages contents are available on the Web which needs to be mined to determine the opinion. Hindi is a national language of India, large amount of Hindi content is available on the Web. From the last few years researchers has performed opinion mining in Hindi language. In this paper an overview of the Hindi based Opinion Mining has given, based on existing researches that has been performed in Hindi language. Techniques and several challenges of Hindi based Opinion Mining are also discussed. But performing opinion mining in Hindi language is not an easy task, because the nature of Indian languages varies a great deal in terms of the script, representation level and linguistic characteristics, etc. To understand the behaviour of Indian languages, large amount of work needs to be done in the field of opinion mining for Hindi language.

Presently very less dataset for performing sentiment Analysis is available. Same work may be replicate for the other Indian language also. most of the work is published in the absence of such a corpus. The task of opinion mining and sentiment analysis is complex and further research is needed into the development of efficient algorithms and into their application to various languages.

REFERENCES

[1] Singh and Vivek Kumar, **"A clustering and opinion mining approach to socio-political analysis of the blogosphere",** Computational Intelligence and Computing Research (ICCIC), IEEE International Conference, 2010.

[2] G.Vinodhini and RM.Chandrasekaran, **"Sentiment Analysis and Opinion Mining: A Survey",** International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.

[3] Alexander Pak and Patrick Paroubek, **"Twitter as a Corpus for Sentiment Analysis and Opinion Mining**".

[4] Bing Liu. **"Sentiment Analysis and Opinion Mining",** Morgan & Claypool Publishers, May 2012.

[5] V.S.Jagtap and Karishma Pawar, **"Analysis of different approaches to Sentence-Level Sentiment Classification",** International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume 2 Issue 3, PP : 164-170 1 April 2013.

[6] B. Pang and L. Lee, **"Opinion Mining and Sentiment Analysis"** Foundations and Trends in Information Retrieval, vol. 2, nos. 1–2, pp. 1–135, 2008.

[7] A.Das and S.Bandyopadhyay **"SentiWordNet for Bangla"** Knowledge Sharing Event-4: Task, Volume 2, 2010.

[8] A.Das and S.Bandyopadhyay, **"SentiWordNet for Indian Languages",** Asian Federation for Natural Language Processing (COLING), China, Pages 56-63, 2010.

[9] A. Joshi, B.A.R, and P.Bhattacharyya, **"A fall-back strategy for sentiment analysis in Hindi: a case study"** In International Conference On Natural Language Processing (ICON), 2010.

[10] Akshat Bakliwal, Piyush Arora, Vasudeva Varma, **"Hindi Subjective Lexicon : A Lexical Resource For Hindi Polarity Classification".**In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC) , 2012.

[11] Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania,Prateek **"Sentiment Analysis of Hindi Review based on Negation and Discourse Relation"** in proceedings of International Joint Conference on Natural Language Processing, pages 45–50,Nagoya, Japan, 14-18 , 2013.
[12] Bharat R. Ambati, Samar Husain, Sambhav Jain, Dipti M.Sharma, Rajeev Sangal, **"Two Methods to Incorporate Local Morph Syntactic Features in Hindi Dependency Parsing"** In Proceedings of the NAACL HLT 1st Workshop on Statistical Parsing of Morphologically-Rich Languages, pages 22–30, 2010.
[13] Nishantha Medagoda, Subana Shanmuganathan, Jacqueline Whalley **"A Comparative Analysis of Opinion Mining and Sentiment Classification in non-English Language"** International Conference on Advances in ICT for Emerging Regions (ICTer 2013) Colombo Sri Lanka.
[14] Richa Sharma, Shweta Nigam, Rekha Jain **"Polarity detection of Movie Review in Hindi Language"** In International Journal on Computational Science & Application (IJCSA) Vol.4,No.4 August 2014.
[15] Piyush Arora, Akshat Bakliwal and Vasudeva Varma, **"Hindi Subjective Lexicon Generation using WordNet Graph Traversal"** In the proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), 2012 New Delhi, India
[16] D. Rao and D. Ravichandran **"Semi-supervised polarity lexicon induction",** In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09, pages 675–682, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
[17] Subhabrata Mukherjee, Pushpak Bhattacharyya, **"Sentiment Analysis in Twitter with Lightweight Discourse Analysis",** In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012).
[18] Bharat R.Ambati, Samar Husain, Sambhav Jain, DiptiM.Sharma, Rajeev Sangal**, "Two Methods to Incorporate Local Morph Syntactic Features in Hindi Dependency Parsing"** In Proceedings of the NAACL HLT 1st Workshop on Statistical Parsing of Morphologically-Rich Languages, pages 22–30, 2010.
[19] Harshada Gune, Mugdha Bapat, Mitesh Khapra and Pushpak Bhattacharyya, **"Verbs are where all the action lies: Experiences of shallow parsing of a morphologically rich language"**, In Proceedings of COLING ,2010 Beijing, China.

