ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

SURVEY REPORT

Available Online at www.ijarcs.info

Survey on Sentiment Analysis in Hindi Language

Komal Garg
M.Tech Scholar: Dept. of Computer Science
Sant Longowal Institute of Engineering & Technology
Sangrur (Punjab),148106, India

Preetpal Kaur Buttar
Assistant Professor: Dept. of Computer Science
Sant Longowal Institute of Engineering & Technology
Sangrur (Punjab),148106, India

Abstract: The recent improvement in web technologies and increasing user-generated content (UGC) in Hindi on the web is the inspiration behind the sentiment analysis in the language. This UGC can turn out to be extremely helpful for researchers, governments and organization to learn what's on public mind, to make sound decisions. Sentiment Analysis is a natural language processing task that mine information from various content structures such as reviews, news, and blogs and classify them on the basis of their polarity as positive, negative or neutral. In this paper, the authors have tried to investigate the on-going research in the field of sentiment analysis, particularly in the Hindi language.

Keywords: sentiment analysis, Hindi, document level, sentence level, aspect level

I. INTRODUCTION

Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to recognize and summarize the opinions expressed by an individual in source materials. The basic task in sentiment analysis is the classification of the polarity of a given text. This arrangement can be possible at different levels: document, sentence, or feature/aspect level- whether the expressed sentiment in a document, a sentence or an entity feature/aspect is positive, negative or neutral.

With internet reaching to more and more people, the users and contributors to the increasing web space are rising tremendously. Hindi is the fourth largest spoken language and has millions of speakers across the world (approx. 490 million speakers) and majority of them are from India. One in five (21 per cent) prefers to access Internet in Hindi in the country. Hindi content consumption on the web has started to take off. It has grown 94 per cent year-on-year compared to 19 per cent growth for English content. Many established Internet companies have started websites which provide information in Hindi. Sanjeev Beekchandani, CEO of Naukri.com has started the Hindi version of his matrimonial portal Jeevansathi.com. As large amount of information is generating regularly, there is need to mine this information by analyzing people's opinions, their views and take the necessary actions.

For example: "A person wants to buy a new car and he is interested in buying the car of Brand A. To get more knowledge about the features of car A, he looks after different websites. There, he read different types of mixed reviews about the features of car of brand A and cars of other brands also (say, brand B car). The person get confused and is not able to have unbiased decision about whether he should car of brand A or brand B."

Therefore, summarization is necessary to present an at-a-glance presentation of the main points made in different reviews about a product/service.

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.

Blogs-The name related to universe of all the blog destinations is called blogosphere. People write about the topics they want to share to others on a blog. Blogging is an event thing on account of its simplicity and straightforwardness of making blog entries, its free frame and unedited nature. We locate an extensive number of posts on every topic of interest on blogosphere [1].

Review Sites- Sentiments are the choice makes for any client in making a buy. The client created audits for products and services are accessible on web. The sentiment classification user reviewer's information gathered from the sites like www.amazon.com which has a large number of item audits by consumers [2].

Micro-blogging- Millions of messages appear everyday in micro-blogging websites such as Twitter, Facebook, etc. Twitter messages express opinions which are used as data source for classifying sentiment [3].

III. DIFFERENT LEVELS OF SENTIMENT ANALYSIS

There are three different levels of Sentiment Analysis: Document-level, Sentence-level, and Aspect-level SA.

1) **Document-level**: SA expects to order an assessment report as communicating a positive or negative feeling or slant. This level considers the entire archive a fundamental data unit. *For example: a product review*, the system determines whether the review expresses an overall positive or negative opinion about the product. The drawback of document-level SA is that it assumes each document expresses opinions ona single entity(i.e. a single product).

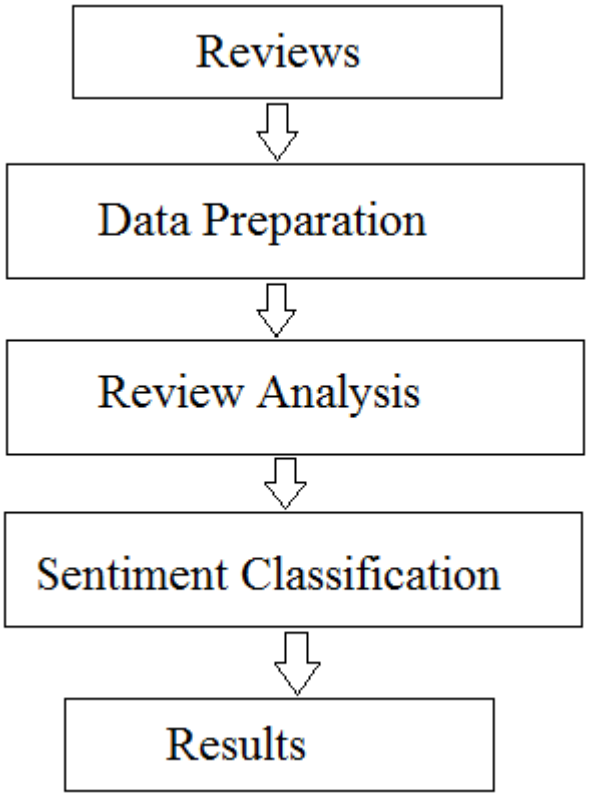
2) **Sentence-level**: SA intends to group conclusion communicated in every sentence. The principal step is to distinguish whether the sentence is subjective or objective. In the event that the sentence is subjective, Sentence-level SA will figure out if the sentence communicates positive or negative assessments. 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.

3) **Aspect-level:** Classifying opinion texts at the document level or the sentence level is often insufficient for applications because they do not identify opinion targets or assign sentiments to such targets. For complete analysis, we need to discover the aspects and determine whether the sentiment is positive or negative on each aspect. Therefore, we go to the aspect level, where the opinion target is decomposed into entity and its aspects. Thus, aspect-based sentiment analysis covers both entities and aspects. *For example: "The songs of this movie are amazing but the story is boring."* This example has two aspects of movie: songs and story. The sentence is positive for "the songs" but it is negative for "the story".

IV. SENTIMENT ANALYSIS MODEL

The typical Sentiment Analysis Model is shown in given figure. The data preparation step performs necessary data preprocessing and cleaning on the dataset for the subsequent analysis. Preprocessing step includes removing information about the reviews that are not required for sentiment analysis, such as review dates and reviewers' names. The review analysis step analyzes the linguistic features of reviews so that interesting information, including opinions and/or product features, can be identified. Two commonly adopted tasks for review analysis are POS tagging [4] and negation tagging. After this phase, sentiment classification is performed to get results.



V. CHALLENGES FOR SENTIMENT ANALYSIS IN HINDI

Unavailability of standard corpus - Unlike English language, there is no training corpus or dataset available for Hindi language.

Word Order- Word arrangement plays an important role in a sentence as it identifies 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 is followed by a verb which is followed by an object. Word order plays an important role in deciding the polarity of a text. Slight variations and changes in the word order can affect the polarity of the text.

Morphological Variations- Hindi language is morphologically rich. Lots of information is fused in the words as compared to the English language where we add another word for the extra information.

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. Also, it's quite complex to handle all the spelling variants while training a model.

Lack of resources- Lack of sufficient resources, tools and annotated corpora also adds to the challenges while addressing the problem of sentiment analysis.

VI. LITERATURE REVIEW

Reitan, Faret, Gamback and Bungum described the first sophisticated negation scope detection system for Twitter sentiment analysis [6]. Proposed system was evaluated on existing corpora from other domains and on a corpus of English Twitter data annotated for negation. . The system consists of two parts: a negation cue detector and a negation scope classifier. The cue detector uses a lexicon lookup that yields high recall, but modest precision. The negation scope classifier produces better results than observed in other domain. The negation cue variation in the Twitter data was quite low. Due to part-of-speech ambiguity it was unclear for some tokens whether they functioned as a negation signal or not.

Dadvar, Hauff and Jong studied the impact of negation detection in SA in movie reviews [7]. The problem of determining the polarity of sentiments in movie reviews when negation words, such as not and hardly occur in the sentences is investigated. Different negation scopes (window sizes) that affect the classification accuracy are examined to investigate how it would affect the polarity identification of the sentences. The results show that there is no significant difference in classification accuracy when different window sizes have been applied.

Singh and Piryani have stated a new kind of domain specific feature-based heuristic for aspect-level sentiment analysis of movie reviews [8]. They have devised an aspect oriented scheme that analyses the textual reviews of a movie and assign it a sentiment label on each aspect. The scores on each aspect from multiple reviews are amassed and a net estimation profile of the motion picture is produced on all parameters. They have utilized a Senti Word Net based plan with two different linguistic feature selections comprising of adjectives, adverbs and verbs and n-gram feature extraction. They have utilized Senti Word Net plan to figure the record level opinion for every motion picture investigated and and compared the results with results obtained using Alchemy API. Toward the end, the notion profile of a motion picture is contrasted and the report level estimation result. The results obtained show that the scheme produces accurate and focused sentiment profile.

Virmani, Malhotra and Tyagi talked about Opinion mining and Sentiment investigation-has developed as a field of study since the World Wide Web and web [9]. The perspective is to extract lines or phrases from crude and immense information. Sentiment analysis on the contrary identifies the polarity of the opinion being extracted. In this paper it is proposed that the sentiment analysis in partnership with opinion extraction, summarization, and tracking the documents of the students. This paper enhances the existing algorithm in order to obtain the collaborated opinion about the students. The outcome viewpoint is presented as very high, high, moderate, low and very low. The paper is based on a case study in which teachers provide their remarks about the students and by applying sentiment analysis algorithm which is proposed, the viewpoint is extracted and represented.

A survey done by Pang B covers techniques and approaches that promise to directly enable opinion-oriented information-seeking systems [10]. Sentiment analysis on online reviews has become increasingly popular. The main focus was on methods that seek to address the new challenges raised by sentiment-aware applications. It Included summarization of evaluative text material. It includes broader issues regarding privacy, manipulation, and economic impact that the development of opinion-oriented information-access services gives rises to.

Arora has examined that Sentiment Analysis recognizes and groups the appraisals/emotions/conclusions of individuals in gathered substance [11]. Till date, English Language incorporates the greatest of the work around there. In this paper, diverse methodologies which are utilized for opinion investigation and exploration work for Indian Languages, like, Hindi, Bengali and Telugu are talked about. An approach is proposed to determine the sentiment positioning that is polarity of the Punjabi language reviews by scoring method. Sentiment analysis is needed to be performed in Punjabi language due to rise in Punjabi data on the web. Separate positive and negative condensed outcomes are created which is useful for the customer or client in decision making.

Akhtar, Ekbal and Bhattacharyya assessed the difficulties of SA in Hindi by providing a benchmark setup [12]; where they created a dataset of high quality and build machine learning models for sentiment analysis keeping in mind the end goal to demonstrate the powerful use of the dataset, and finally make the resource available to the community. The dataset involves Hindi product reviews taken from online resources. For classification, Conditional Random Filed (CRF) and Support Vector Machine (SVM) were used for aspect term extraction and sentiment analysis. Assessment comes about demonstrate the normal F-measure of 41.07% for viewpoint term extraction and precision of 54.05% for notion grouping.

Mittal, Agarwal, Chouhan, Bania and Pareek proposed a method to increase the coverage of the Hindi SentiWordNet for better classification results[13]. Furthermore, effect of the nullification and discourse rules are researched for Hindi SA.

Farooq, Mansoor, ent., investigate the problem of identifying the scope of negation while determining the polarity of a sentence [14]. A negation handling method is proposed which can handle different types of negation while determining the polarity of a sentence. It determine the sequence of words affected by syntactic negation while considering the exceptions such as when the negation even don't have a scope and when the negation inverts the polarity of a clause even though no opinionated word is within the scope. The effect of diminishes is incorporated to reduce the strengths of polarities of those words which are affected.

Arora et al. proposed a graph based method to build a subjective lexicon for Hindi language, using WordNet as a resource [15]. 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.

Mukherjee et al. showed that the incorporation of discourse markers in a bag-of-words model improves the sentiment classification accuracy by 2 - 4% [16]. Bakliwal et al. proposed a method to classify Hindi reviews as positive or negative [17]. 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. proposed a novel approach to detect errors in the tree banks. This approach can significantly reduce the validation time [18]. They tested it on Hindi dependency tree bank data and were able to detect 76.63% of errors at dependency level.

VII. APPLICATIONS

Sentiment Analysis has been broadly utilized for comprehension the subjective nature of content. Areas where Sentiment Analysis can be connected are-

1) **Aid in choice making**- Decision making is an essential piece of our life. This moves from "which articles to purchase", "which eating place to go" to "which bank insurance schemes to strive for", "which schemes to make". Sentiment Analysis can be utilized to decide and select from the accessible alternatives focused around the general sentiments communicated by different clients.

2) **Designing and Building Innovative Products**- With presented to extreme rivalry and open to faultfinders through open audits and suppositions, estimation examination prompts better investigation of the items as far as the convenience and human- accommodating nature. It makes an environment for better and more imaginative items.

3) **Recommendation Systems**- Most of the sites we visit have a proposal framework in-assembled to help us, extending from locales identified with books, online-media, amusement, music, film industry to different types of workmanship. These frameworks utilizes our individual data, past history, likes and despises and our companions data to make recommendations.

4) **Products Analysis**- With the assistance of assumption investigation it has gotten to be simpler to investigate different items and settle on the decisions appropriately. This sort of examination additionally serves to choose an item focused around its peculiarity particulars. The examination between two items has additionally been made very simpler.

5) **Business Strategies**- Much of the business methods are been guided regarding the response from the individuals. Organizations intends to fulfill the needs and requests of the clients, accordingly vital moves of organizations are determined through general conclusions and perspectives. With the world connected through innovation occasions have a worldwide effect; the issue/disappointment on one piece of the world has an effect on the other corner of the globe. So it gets to be truly imperative to drive products/administrations as indicated by people in general perspective.

VIII. CONCLUSION AND FUTURE SCOPE

Sentiment Analysis is an emerging field and is important as human beings are largely dependent on it nowadays. Presently, the use of sentiment analysis has spread to services and making applications and developments came into existence in this area. Its main target is to make computer able to identify and create emotions like human being. This paper aims to create a data bank to facilitate the referencing needs of researchers and practitioners in this area. To this end, this paper presents the literature review pertaining to this topic. The literature review is based on the data collected from various research papers, tools and web sources that will strongly assist in easy referencing. It presents an overview on the recent updates in sentiment analysis algorithm and applications. Various articles are categorized and summarized. These articles give contributions to many sentiment analysis related fields that use sentiment analysis techniques for various real-world applications. After analyzing these articles, it is clear that the enhancements of sentiment analysis algorithms are still an open field for research. We assume that in the future the application areas of sentiment analysis will still increase and that the adaption with sentiment analysis techniques will become standardized part of many services and products.

IX. REFERENCES

[1] Singh Vivek Kumar et al, "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". https://pdfs.semanticscholar.org/ad8a/7f620a57478ff70045f97abc7aec9687ccbd.pdf

[4] Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams". Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March.

[5] Pang B., Lee, L., And Vaithyanathain. S. 2002."Thumbs up? Sentiment classification using machine learning techniques", In Proceedings of the Conference on Empirical Methods in Natural Language Processing , 79- 86.

[6] Johan Reitan, Jørgen Faret, Björn Gamback, and Lars Bungum,"Negation Scope Detection for Twitter Sentiment Analysis". http://www.aclweb.org/anthology/W15-2914

[7] Maral Dadvar, Claudia Hauff, and Franciska de Jong,"Scope of Negation Detection in Sentiment Analysis". https://core.ac.uk/download/pdf/11479307.pdf

[8] Singh, V., & Piryani, R. (2013). Sentiment Analysis of Movie Reviews A new feature–based heuristic for aspect-level sentiment classification, Department of Computer Science,South asian University,New Delhi.

[9] Deepali Virmani, Vikrant Malhotra and Ridhi Tyagi, "Sentiment Analysis Using Collaborated Opinion Mining", Department of Information Technology, Bhagwan Parshuram Institute of Technology PSP- 4.

[10] Bo Pang and Lillian Lee,"Opinion Mining ans Sentiment Analysis". http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

[11] Parul Arora and Brahmaleen Kaur(2015), "Sentiment Analysis of Political Reviews in Punjabi Language", International Journal of Computer Applications (0975 – 8887) Volume 126 – No.14.

[12] Joshi, A., A R, B., & Bhattacharyya, P. (2010). A fall strategy for sentiment analysis in Hindi : A case study. International Conference on Natural Language Processing.

[13] Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania, Prateek Pareek,"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.

[14] Umar Farooq, Hasan Mansoor, Antoine Nongaillard, Yacine Ouzrout, M.A.Qadir,"Negation Handling in Sentiment Analysis at Sentence Level", in Jounal of Computers,Volume 12, Number 5, September 2017,pages 470-478.

[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] Subhabrata Mukherjee, Pushpak Bhattacharyya, "Sentiment Analysis in Twitter with Lightweight Discourse Analysis", In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012).

[17] 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.

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