SCORE-BASED SENTIMENT ANALYSIS OF BOOK REVIEWS IN HINDI LANGUAGE

Firdous Hussaini[1], S. Padmaja[2] and S. Sameen Fatima[3]

[1]Department of CSE, STLW, Osmania University, Hyderabad, India
[2]Associate Professor, Keshav Memorial Institute of Technology, Hyderabad, India
[3]Professor and Head, Department of CSE, UCE Osmania University, Hyderabad, India

ABSTRACT

Sentiment analysis has been performed in different languages and in various domains, such as movie reviews, product reviews and tourism reviews. However, not much work has been done in the area of books considering the high availability of book reviews on Hindi blogs and online forums. In this paper, a score-based sentiment mining system for Hindi language is discussed, which captures the sentiment behind the words of book review sentences. We conducted three experiments using scores from the Hindi-SentiWordNet (H-SWN), first using parts-of-speech tags of opinion words to extract their potential scores. Then, we focused on word-sense disambiguation (WSD) to increase the accuracy of system. Finally, the classification results were improved by handling morphological variations. The results were validated against human annotations achieving an overall accuracy of 86.3%. The work was extended further using Hindi Subjective Lexicon (HSL). We also developed an annotated corpus of book reviews in Hindi.

KEYWORDS

Natural Language Processing, Sentiment Analysis, Lexicon-based, Word-sense Disambiguation

1. INTRODUCTION

We live in the age of social media where interaction with people on digital platforms is a common activity in everyday life. Not only do these digital platforms help in exchanging updates in the lives of people, they also give people an opportunity to be heard. There is no denying that social media has helped people to voice their opinions in areas like politics, business, marketing, public relations, etc. It has enabled users to actively engage in trending conversations and to post their opinions freely on the web. These opinions, written in different languages, are a rich source for subjective information which can be used in data analysis, recommendation systems and various other applications.

Extraction and evaluation of public opinions is an extensive task due to the huge data available in digital form which may or may not be subjective in nature. Therefore, opinion mining (also known as sentiment analysis) techniques are helpful in detecting the presence of an opinion and finding its polarity, i.e. classifying it as either positive or negative.

Sentiment analysis is a natural language processing (NLP) related task which aims to identify the emotion of a writer or a speaker. “Sentiment analysis, also called opinion mining, is a field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and

DOI: 10.5121/ijnlc.2018.7511

their attributes" [1]. It deals with finding the orientation of an opinion in a piece of text by capturing an individual's feelings in relation to a particular subject or issue.

1.1. Motivation

With the advent of Unicode standards, different languages around the globe now have their own space on the Internet. One of these languages is Hindi and its content consumption is growing by 94% year-on-year [2]. Web pages in Hindi language are increasing day by day making user reviews freely available on blogs and online forums. Among these are book reviews which people write to express their feelings and opinions about the books they read. Evidently, reviews affect the sale of books [3] because people mostly rely on them before they buy them. Book reviews usually summarize the content of the book and promulgate the reviewer's opinions on it. A positive opinion about a book or its author results in its recommendation. Also, classifying opinions brings together like-minded readers that enjoy reading similar genre of books.

1.2. Challenges

There are a few challenges that need to be dealt with when mining sentiments in Hindi language:

- **Noisy data:** The content may include slang words, abbreviations, emoticons, hash tags, etc. which may show a certain sentiment but must be refined in order to be identified by the system.

Examples:

-- झिलाउ /jhilāu/ means "intolerable": a slang word expressing emotion but not found in lexicon. Such words can be appended to subjectivity lexicon.

– ☺ ☹ 😐 : Emoticons that symbolize strong sentiments can be replaced with opinion words which convey similar sentiments.

- **Word Sense Disambiguation:** A word can have more than one meaning depending on its usage. For example, the word अलग /alag/ has 5 senses as adjective and 2 senses as adverb in the Hindi WordNet lexicon. It could mean "distinct", "separate" or even "aloof", depending on the context. If the most appropriate sense of the word is not chosen, then the polarity of the whole sentence may change.

- **Morphological Variations:** Hindi is a synthetic language, i.e. it extensively uses a fusion of morphemes to form a word. When a morpheme is joined with an affix such that it changes the dictionary meaning of the word, the morphological variation is said to be derivational. For example, अनुचित /anuchit/ which means "inappropriate" is a fusion of अन /an/ + उचित /uchit/ i.e., "not" + "appropriate". On the other hand, inflectional variations only change the grammatical features like gender, case, number and tense. For example, when the word अच्छा /achchha/ "good" is joined as अच्छा/achchha/ + ई /ī/ = अच्छी /achchhī/, it inflects feminine gender.

- **Capturing idioms/verb phrases:** There are idioms and verb phrases in Hindi language which represent strong sentiments and are difficult to comprehend. For example, the

sentence below uses the idiom दिल छू जाना /dil chhu jāna/ which means "heart- touching", carrying a positive sentiment:

कहानी आप क` दिल को छू जाय`गी।
/kahāni āp ke dil ko chhu jāyegi/ "The story will touch your heart."

2. RELATED WORK

Research in sentiment analysis has been quite productive in the last few years. Hu and Liu [4] proposed a lexicon-based method for predicting sentiment of customer reviews at aspect-level classification. First, they identified the opinion words (adjectives) in the text. Then, they used a bootstrapping technique to determine the semantic orientation of each opinion word and finally summarized the results. Kim and Hovy [5] implemented the use of "a measure of sentiment strength" (score) to classify opinion words and sentences into positive and negative categories. Sentiment analysis techniques have also been extended to Indian languages. Lexical resources have been developed in Bengali, Hindi and Telugu. Small amount of work has been done using cross-lingual sentiment analysis in other Indian languages including Tamil, Punjabi, Marathi and Manipuri. Das and Bandyopadhyay [6] proposed computational techniques for generating SentiWordNet for Indian languages which include dictionary-based, WordNet-based, corpus-based and also a gaming methodology where sentiment score is calculated based on the inputs given by different players.

Joshi et al., [7] proposed a fall-back strategy for sentiment analysis in Hindi. They created a manually annotated corpus for Hindi (Movie reviews). They also constructed H-SWN, a lexical resource based on the equivalent of the one in English. They also studied three approaches that can be adopted for sentiment analysis in a new language: training a sentiment classifier on in-language labelled corpus and use it to classify a new document; applying machine translation to translate the new document into a resource-rich language like English and then using a classifier to detect its polarity; and using a resource like SentiWordNet for the new language to detect the document's polarity. Chakrabarti et al., [8] documented their experiences in building the Indo WordNet. They built an on-line lexical database for Hindi language, inspired by the design of the English WordNet. For each word, a synonym set was found which represents one lexical concept. The synonym sets are related to other synonym sets through semantic relationships of hypernymy, hyponymy, meronymy, holonymy and antonymy. They included some unique features like the concept of gradation with antonymy and meronymy in the Indo WordNet.

Bakliwal et al., [9] developed a Hindi subjectivity lexicon of adjectives and adverbs using WordNet and Breadth First Graph traversal method [10]. They made an initial seedlist of words along with their polarity and expanded it based on the synonyms and antonyms of the words present in the list, assuming that synonyms possess similar polarity and antonyms the opposite. They used three validation strategies for their proposed work. First, they tested their lexicon on pre-annotated reviews. Then, they validated the lexicon against existing resources. Finally, manual annotators were hired to annotate the generated lexicon. They also developed an annotated corpus for Hindi product reviews. Mittal et al., [11] proposed new rules for handling Negation and Discourse relations in sentences. Their other work includes expansion of H-SWN and development of annotated corpus for Hindi movie reviews.

3. PROPOSED WORK

This section presents an overview of the proposed work. Figure 1 illustrates the four phases of the system: corpora acquisition, data pre-processing, inter-annotator agreement and sentence sentiment classification.

3.1. Corpora Acquisition

In this phase, book review documents were scraped from web pages. There are a few websites which provide book reviews in Hindi language such as patrika.com [12] and hindi.webdunia.com [13]. In our work, reviews were scraped using Python's *urllib* package and BeautifulSoup4. This reduced the manual work required to extract data from the Web. Python's *urllib.request* module was used for fetching URLs while BeautifulSoup4 library assisted in the extraction of Unicode characters from the web pages into a CSV file.

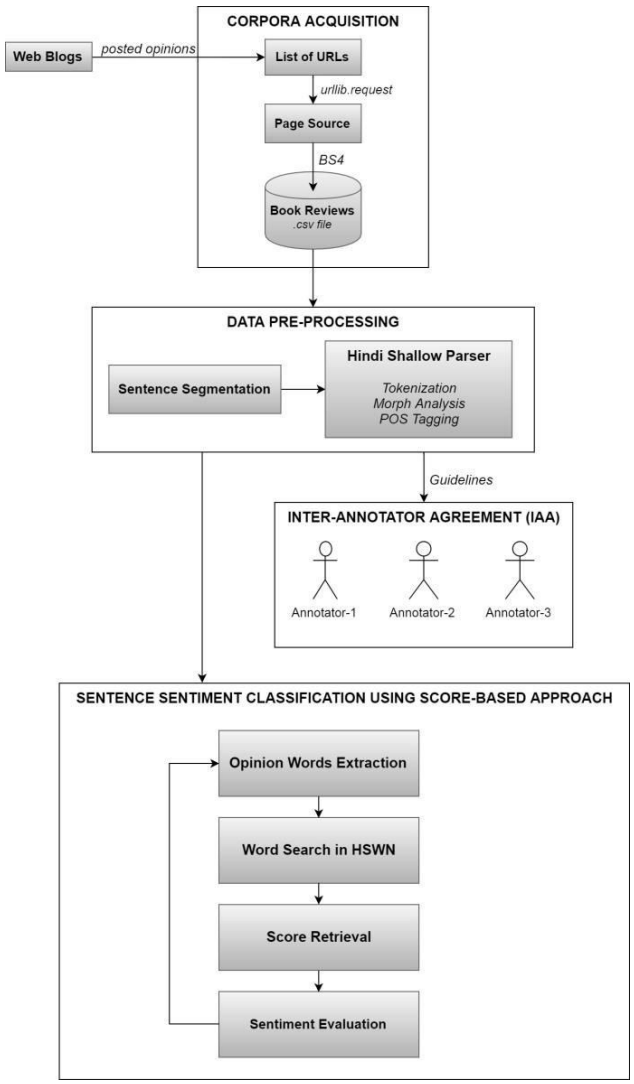

Figure 1. Score-based Sentiment Analysis Architecture

3.2. Data Pre-processing

Data extracted from the web pages is usually in its raw form and must be pre-processed before applying methods for sentiment analysis. In this phase, we performed sentence segmentation and shallow parsing. The document reviews extracted and stored in the CSV file were parsed into sentences. Spelling and grammatical corrections were made to each of these sentences. Sentences that summarized the content of the book were removed manually. Hindi shallow parser developed by the Language Technologies Research Centre IIIT-H [14], was used to perform tokenization, part-of-speech tagging, chunking and morphological analysis on the selected review sentences.

3.3. Inter-annotator Agreement

Three Hindi-speaking annotators were asked to classify the review sentences into positive (POS), negative (NEG) or Neutral (OBJ) classes. This was done based on the following set of guidelines which were given to them:

1. Look for the occurrence of sentiment-bearing words associated with the entity (book) and its aspects. If you feel that the sentiment towards the target is positive, then label the sentence as POS, or if negative, then label it as NEG.

2. Identify idioms and verb-phrases that show positive or negative opinions. Example:

छू जाना,आँख नम होना

/chhu jāna/ /ānkh nam hona/
"touching" "teary-eyed"

3. Some sentiment-bearing words seem to be positive for one aspect and negative for another. In such cases, examine carefully and label according to your preference.

Example: अलग /alag/

i. इस पुस्तक की सबस` खास बात इसका <u>अलग</u> विषय ह"। (POS)

/is pustak ki sabse khās bāt iska <u>alag</u> vishay hai/
"The most special thing about this book is its <u>distinct</u> topic."

ii. ि◌हं◌ पु◌स्तक का शीषक और आि◌रण भ◌ी कह◌ाननयो◌ं स` <u>अलग</u> लगत◌ा ह◌ै (NEG)

/wahin pustak ka shirshak aur āwran bhi kahāniyon se <u>alag</u> lagta hai/
"The book's title and cover also seem <u>irrelevant</u> from the stories."

4. Look out for words that show negation in sentiment:

नह◌ीं /nahin/, न /na/, ना /nā/, मत /mat/, ना तो... और ना ह /nā to... aur nā hi/.

5. Sentences which express recommendation are POS. Example:

इस पुस्तक को एक बार पढ़ना तो बनता ह ह"।

/is pustak ko ek bār parhna to banta hi hai/ "This book must be read once."

6. When handling discourse relations, notice that the sentiment following the connectives

ल`ककन, ब◌ावि◌जू◌ि◌ and बल्कक

/lekin/, /bāwjud/ and /balki/
"but", "despite" and "rather"

seems to have more emphasis. Example:

ढ ठ विषय क` साथ बिल` की य` कहानी दिलिस्प तो ह", <u>ल`ककन</u> इसका शशखर थोडा छोटा ककया जा सकता था।

/dhith vishay ke sāth badle ki ye kahāni dilchasp to hai, <u>lekin</u> iska shikhar thora chhota kiya

ja sakta tha/
"This story of revenge with its strong subject is interesting, but its climax could be slightly shorter."

7. Also, when हालाांकि /hālānki/ "although" connects two segments, the sentiment that precedes it has more emphasis.
Example:
पुस्तक की कहानी दुर्जोय की अन्य कहाननयोां स थोडी अलग ह, हालाांकक इसमेां अध्यायोां
को थोडा छोटा ककया जा सकता था।
/pustak ki kahāni durjoy ki anya kahāniyon se thori alag hai, hālānki ismen adhyāyon ko thora chhota kiya ja sakta tha/
"The book's story is slightly different from the other stories of Durjoy, although the chapters could be slightly shorter."

8. When in doubt, label the text as OBJ.

We have used Fleiss's kappa as a statistical measure to evaluate the level of agreement between the annotators. The kappa value was calculated to be 0.794 (79.4%) indicating substantial agreement between the annotators. Finally, the majority vote of annotators was taken for each sentence and human annotation baseline was established. Table 1 summarizes details of the produced dataset.

Table 1. Annotated dataset summary

Total number of sentences	700
Sentences annotated as Positive	300
Sentences annotated as Negative	300
Sentences annotated as Neutral	100
Inter-annotator agreement	79.4%

3.4. Opinion Word Extraction and Score Retrieval

We have considered five parts-of-speech for possible opinion words: adjectives, adverbs, nouns, quantifiers and verbs. Words from the corpus belonging to these parts-of-speech categories were selected along with their part-of-speech tags to perform a stipulated search in the sentiment lexicon for their potential scores. This kind of search that includes the tag eliminated part-of-speech ambiguity.

Once the word as well as its part-of-speech tag was matched in the lexicon, its positive and negative sentiment scores were extracted for evaluation. Words that show negation were also selected from corpus to assist in negation handling.

3.5. Sentiment Evaluation

In this phase, we deal with the scores retrieved for all the opinion words that are present in a sentence. Algorithm 1 explains the process of sentiment evaluation using these scores. We calculated the total positive score and negative score for each sentence. The higher value among the two score determined the polarity of the sentence. The sentences were classified as neutral if the two scores were found to be equal.

In case of word sense ambiguity, the average of the scores of all senses was calculated for evaluation. Algorithm 2 explains sentiment classification using WSD. While working with HSL, objective scores assigned to opinion words were also taken into consideration. HSL [9] already handles word-sense ambiguity problem to some extent, therefore, WSD was not needed.

Algorithm 1: Sentiment classification by extraction and summation of scores

Input: Ordered dictionary *corpDict* containing list of extracted word and part-of-speech tag co-ordinates *(w_c, t_c)*
Output: Classification into positive, negative and neutral class

Initialize tp_score, tn_score
for all (w_c, t_c) ∈ corpDict **do**
if w_c == w_s **and** t_c == t_s **then**
//if both word and tag match in sentiment lexicon
tp_score = tp_score + p_score tn_score = tn_score + n_score
end if end for
if tp_score == tn_score **then**
Classify as neutral
end if
if tp_score > tn_score **then**
Classify as positive
end if
if tp_score < tn_score **then**
Classify as negative
end if

Algorithm 2: Sentiment classification using WSD

Input: Ordered dictionary *corpDict* containing list of extracted word and part-of-speech tag coordinates *(w_c, t_c)*
Output: Classification into positive, negative and neutral class
Initialize match, tp_score, tn_score
for all (w_c, t_c) ∈ corpDict **do**
while w_c == w_s **and** t_c == t_s **do** *//while both word and tag match in sentiment lexicon*
match = match+1
end while
if match ==1 **then** *// if word has only one sense*
tp_score = tp_score + p_score tn_score = tn_score + n_score
end if
if match > 1 **then** *// if word has multiple senses*
tp_score = tp_score + avg (p_score) tn_score = tn_score + avg (n_score)
end if end for
if tp_score == tn_score **then**
Classify as neutral

end if
if tp_score > tn_score **then**
Classify as positive
end if
if tp_score < tn_score **then**
Classify as negative
end if

4. EXPERIMENTAL RESULTS WITH ANALYSIS

We conducted three rounds of experiments on the dataset. In the first experiment, the sentiments were evaluated based on parts-of-speech tags collected from the output of shallow parser. Negation tags were also taken into account to determine the sentiment. The results were tested against human annotations which gave an accuracy of 74%. Considering this experiment as a baseline, we made improvisations in the next two experiments.

It was observed that sentences contained opinion words that have more than one sense. Since each sense of a word has a different score assigned to it in the subjectivity lexicon, all the scores were aggregated by the system leading to the misclassification of sentences. Hence, in the second experiment, we focused on such cases where WSD has to be applied. This was done by calculating the average score of all senses first and then adding it to the overall positive and negative score of that sentence. The following is an example where word-sense ambiguity was handled to classify the sentences accurately:

इस पुस्तक की सबस` खास बात इसका अलग विषय ह"।
/is pustak ki sabse khās bāt iska alag vishay hai/
"The most special thing about this book is its distinct topic."

This sentence was annotated as positive. It contains two opinion words खास and अलग. Given below are their positive and negative scores from H-SWN, respectively:

खास 0.5 0.25
अलग 0.125 0.5
अलग 0.0 0.0

Note that अलग has scores for two senses as an adjective. The classification results after the two experiments can be compared in Table 2.

Table 2. Comparison of Sentence Score Evaluation showing total positive score (tp_score) and total negative score (tn_score)

Experiment	tp_score	tn_score	Classification
Baseline	0.625	0.75	Negative
Baseline + WSD	0.5625	0.5	Positive

Some other words that have multiple senses include सहज /sahaj/, सीधा /sīdha/, फीका /phīka/, अच्छा /achchha/, सामान्य /sāmāny/, etc. After conducting this experiment the accuracy increased by 8%. However, it was further observed that some of the opinion words could not be accessed in H-SWN due to their inflected form. As a result, sentences were misclassified. Therefore, a third experiment was conducted by using the root words obtained after morphological analysis to eliminate such cases. Table 3 shows some examples for sentences that were misclassified due to morphological variations. It was noted that accuracy further increased by 4%.

Table 3. Examples of Sentence Sentiment Classification after handling morphological variations

Sentences	**Annotation**	**Prediction before Morph.**	**Prediction after Morph.**	**Inflected Word**	**Root Word**
िहंदी पुस्तक की कहानी भी अनोखी ह"। /wahin pustak ki kahāni bhi anokhi hai/ "The story of the book is also unique."	Positive	Objective	Positive	अनोखी	अनोखा
इस पुस्तक में कथाओं को इतनी अच्छी तरह स बयां ककया गया ह" कक आप उन ककरिारों को महसूस कर सकत हैं। /is pustak men kathāon ko itni achchhi tarah se bayān kiya gaya hai ki āp un kirdāron ko mahsus kar sakte hain/ "The stories in this book have been so well told that you can perceive those characters."	Positive	Negative	Positive	अच्छी	अच्छा
कह-कह कहानी ढ ल पड जाती ह"। /kahin-kahin kahāni dhili par jāti hai/ "In some places the story falls loose."	Negative	Objective	Negative	ढ ल	ढ ला

ककताब पूर तरह स` शाह क व्यल्ततत ि पर <u>खर</u> उतरती ह'। /kitāb puri tarah se shāh ke wyaktitw par <u>khari</u> utarti hai/ "The book is <u>truthful</u> to Shah's personality in every way."	Positive	Negative	Positive	खर	खरा

In comparison, a fourth experiment was conducted using HSL as subjectivity lexicon. Figure 2 is a generated bar graph for all experiments, which indicates an overall improvement in accuracy. Table 4 summarizes the performance evaluation of all experiments. Figure 3 depicts the performance results of working with H-SWN.

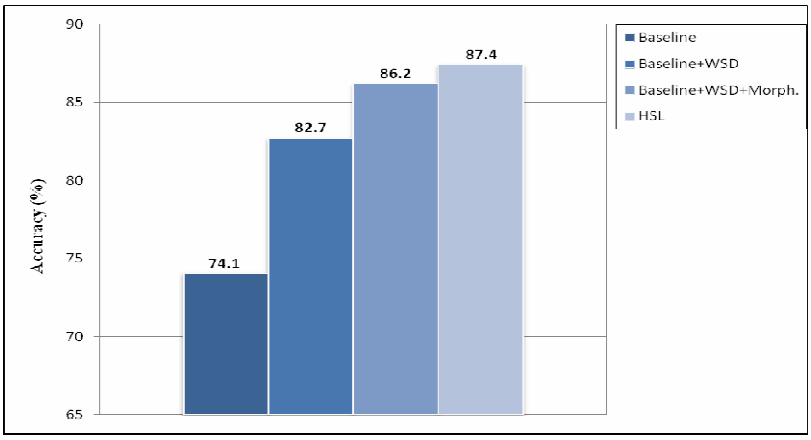

Figure 2. Accuracy Results

Table 4. Performance Evaluation for experiments using score-based sentiment classification, showing accuracy, precision and recall for three classes - positive, objective and negative

Experiment	**Accuracy**	**Positive**		**Objective**		**Negative**	
		Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Baseline	0.74	0.83	0.65	0.46	0.87	0.85	0.78
Baseline + WSD	0.82	0.91	0.76	0.54	0.91	0.91	0.86
Baseline + WSD + Morph.	0.86	0.92	0.83	0.62	0.91	0.92	0.87
HSL	0.87	0.87	0.89	0.68	0.88	0.95	0.85

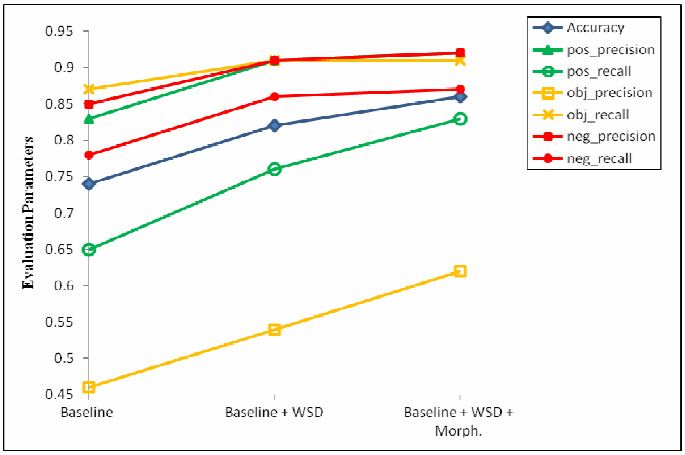

Figure 3: Performance Results

5. CONCLUSION

In this work, score-based approach for sentiment classification of book reviews in Hindi language has been applied. Opinion words were extracted from individual sentences using parts- of-speech tagger, incorporated within the Hindi shallow parser. Unlike the previous works done on sentiment analysis, our work considers not only adjectives but also adverbs, nouns and verbs for extraction. This approach uses subjectivity lexicons for retrieving polarity scores of the extracted words. The overall positive and negative scores were calculated for each sentence, the higher value between the two determining the polarity of the sentence.

A dataset of 700 sentences pertaining to book reviews was considered for this work. These sentences were first annotated by three Hindi-speaking annotators. The mutual agreement between them was calculated and the kappa value was found to be 79.4%. The results obtained from the system were tested against these human annotations. An accuracy of 86.3% was achieved working with H-SWN, after applying WSD and handling morphological variations. An accuracy of 87.4% was achieved working with HSL.

6. FUTURE WORK

This system is highly dependent on the subjectivity lexicon used, i.e. presence of the opinion word and the appropriateness of sentiment score with respect to the domain in question. The system can be enhanced by expanding the subjectivity lexicons to provide better coverage of opinion words and corpus-relevant scores.

Further, issues like discourse relations and idioms that cannot be found in the subjectivity lexicons need to be handled. This will give a better accuracy to the system.

REFERENCES

[1] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

[2] http://trak.in/tags/business/2015/08/19/hindi-content-content-consumption-growth-india-google/

[3] Chevalier, J. A. & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. Journal of Marketing Research: August 2006, Vol. 43, No. 3, pp. 345-354.

[4] Hu, M. & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168-177

[5] Kim, S. M. & Hovy, E. (2004). Determining the sentiment of opinions. In Proceedings of the 20th International Conference on Computational Linguistics, p. 1367. Association for Computational Linguistics.

[6] Das, A. & Bandyopadhyay, S. (2010). SentiWordNet for Indian languages. Asian Federation for Natural Language Processing, China: 56-63.

[7] Joshi, A., Balamurali, A. R. & Bhattacharyya, P. (2010). A Fall-Back Strategy for Sentiment Analysis in Hindi: a Case Study. In Proceedings of the 8th International Conference on Natural Language Processing.

[8] Chakrabarti, D., Narayan, D. K., Pandey, P. & Bhattacharyya, P. (2002). An Experience in Building the Indo Wordnet - A Wordnet for Hindi. In Proceedings of First International Conference on Global WordNet.

[9] Bakliwal, A., Arora, P. & Varma, V. (2012). Hindi subjective lexicon: A Lexical Resource for Hindi Polarity Classification. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC).

[10] Arora, P., Bakliwal, A. & Varma, V. (2012). Hindi Subjective Lexicon Generation using WordNet Graph Traversal. International Journal of Computational Linguistics and Applications 3.1: 25-39.

[11] Mittal, N., Agarwal, B., Chouhan, G., Bania, N. & Pareek, P. (2013). Sentiment Analysis of Hindi Review Based on Negation and Discourse Relation. In Proceedings of International Joint Conference on Natural Language Processing.

[12] http://www.patrika.com/news/books/

[13] http://hindi.webdunia.com/hindi-books-review

[14] http://ltrc.iiit.ac.in/analyzer/hindi/

AUTHORS

Firdous Hussaini did her Masters in Software Engineering from Stanley College of Engineering and Technology for Women, Osmania University and B.E in Information Technology from ISL Engineering College, Osmania University. Her research areas are Natural Language Processing and Machine Learning.

Dr. Padmaja S received her PhD in computer science from Osmania University. She is an Associate Professor at KMIT, Hyderabad. She regularly contributes to scholarly journals, conferences and is also reviewer. She is a resource person for various courses offered at research and academic institutions of repute. Natural Language Processing, Machine Learning, Big data Analytics are few of her areas of research interest. She can be reached at *bandupadmaja@gmail.com* for research collaboration.

Prof. Syeda Sameen Fatima has over 33 years of experience in teaching, research and administration in India, USA and UAE. She took over as Principal in July 2016, and holds the distinction of being the first lady Principal, in the history of the College of Engineering, Osmania University. Currently she is a Professor at the Department of Computer Science and also the Director, Centre for Women's Studies at Osmania University. She has published several papers in national and international journals and conferences. Her areas of interest include Machine Learning, Text Mining and Information Retrieval Systems. She received the "Best Teacher Award" by The Government of Telangana, India in the year 2017. She can be reached at *sameenf@gmail.com*.