Word Sense Disambiguation for Lexicon-based Sentiment Analysis in Hindi

Dhanashree. S. Kulkarni
Gogte Institute of Technology, Belagavi, India.
E-mail: dskulkarni10@gmail.com

Dr. Sunil. F. Rodd
Gogte Institute of Technology, Belagavi, India.
E-mail: sfroddgit@git.edu

Received August 12, 2021; Accepted November 27, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19042

Abstract

The amount of data from users in Hindi language is tremendously increasing on social media, blogs, online forums due to which Sentiment analysis of Indian languages has turned out to be a predominant research area. Lexicon based analysis is one of the techniques that can be used for analyzing sentiments. While using the lexicon-based sentiment analysis, ambiguous words could be an issue. This paper proposes a graph based Lesk approach to handle the word sense disambiguation issue and tries to enhance and improve the lexicon approach. Experiments performed to evaluate the performance of the proposed algorithm shows that the performance of lexicon-based sentiment analysis is increased significantly by using the graph based Lesk approach.

Keywords

Word Sense Disambiguation, Sentiment Analysis, Opinions, Machine Learning, Lexicon, Hindi, Text Analysis.

Introduction

Sentiment Analysis has become a very prominent area of research where sentiments or opinions are analyzed and are categorized based on polarity namely positive, negative, neutral or on emotions such as happy, sad, fear etc. It has been used in a broad range of applications such as reputation management, product analysis, market analysis and recommendation systems. The reviews and comments given by people may be in different languages. Hindi being the national language of India and one amongst the most spoken languages, many people give their comments in Hindi. Therefore, performing sentiment

analysis task for Hindi language is of utter importance and could prove to be very much useful especially for government organizations.

Sentiment Analysis often called as Opinion mining (B. Pang and L. Lee, 2008) uses two well-known approaches for analyzing sentiments namely Lexicon approach and Machine Learning approach. Lexicon approach relies on a dictionary or a sentiment lexicon which is predefined set of words along with their corresponding sentiment value. On the other hand, Machine learning approach generates a system that learns from experience and has the capability to improve itself. Lexicon based approaches are said to be time efficient but are not that accurate (Ł. Augustyniak.et.al., 2016) Machine learning in overall is more accurate but their performance may change when applied to data of different domains (Anthony and M. Gamon, 2005). Lexicon based methods, on the other hand, may show steady and reliable performance when different domains are considered. In case of Hindi sentiment analysis, lexicon-based methods are implemented using a Hindi SentiWordNet. Hence accuracy can be increased in such cases by making efficient use of the Hindi SentiWordNet.

Sentiment analysis is a sub-field of Natural language processing (NLP) and the well-known problem of NLP is the word sense disambiguation issue. Solving the problem of word sense disambiguation is essential for other applications also such as Machine Translation, speech synthesis, information retrieval, text processing, grammatical analysis etc. (Dilip Kumar Sharma, 2015). To build an efficient sentiment analysis system, it is important to find the correct sense of the word depending on the context.

The most important task of opinion mining is to find out the opinion words or sentiment words. This process can be done effectively when the system uses an accurate and large sentiment lexicon. HSWN has been created by IIT Bombay by making use of the Hindi WordNet and English SentiWordNet (SWN) and they followed a process of matching the words which had the same synset IDs (Joshi Aditya et al.,2010). Since Hindi language is ambiguous and words used may change depending on the context, a precise sense for a word in the Hindi SentiWordNet has to be selected. But it has been seen that words from the lexicon are directly selected and the word sense ambiguity issue is not handled. The lexicon does have polysemous words i.e., words which have more than one meaning and to get the right sentiment score, it is necessary to extract the correct sense of the word from the lexicon. There are different approaches that are being used for solving the problem of word sense disambiguation such as first sense of the word (Joshi, Aditya et al.,2010) or taking the average score of the words (Kanika Garg et al.,2020). If first score of the word is taken it may decrease the performance of the opinion mining system. If

averaging is used, it may yield good performance but it does not aim at getting the correct sense and hence does not consider the effects to the domain knowledge (Singh S., et al, 2013).

Related Work

Performing Sentiment analysis in Hindi uses different techniques in order to find the sentiment at document level, sentence level as well as aspect level (Vandana Jha et al., 2016). Lexicon based technique is one such approach that makes use of a dictionary to find out the corresponding score of the words (Sharma P, Moh TS, 2016). But there are some problems such as negation handling, discourse analysis that may affect the performance of the lexicon-based sentiment analysis (Mittal N et al, 2013). Word sense Disambiguation (WSD) is also one such problem that needs to be dealt with.

Satyendr Singh et al., (2013) evaluated the WSD problem using the Leacock Chodorow semantic relatedness method. The method only considered Nouns. Chandra Bhal Singh et al., (2016) proposed the Lesk Approach on bigrams and trigram sentences considering verbs words which are ambiguous as the target words.

A supervised method was presented by Sarika et al., (2016) for calculating cosine similarity between query which contained the ambiguous word and the sense knowledge data and selecting the sense which obtained the maximum similarity.

Pamungkas et al., (2017) tried to handle ambiguous issue of lexicon-based sentiment analysis by implementing different methods of path based semantic relatedness and showed that the performance of sentiment analysis is improved. Firdous Hussaini et al., (2018) presented a sentiment mining system which was based on HSWN scores and the accuracy of the system was increased by solving the WSD issue by averaging of the scores if a particular word had multiple senses. Anidhya Athaiya et al., (2018) introduced a genetic algorithm for Hindi word sense disambiguation by analyzing the neighboring words of the target word. The proposed algorithm only dealt with nouns.

Pooja Sharma et al., (2019) explored the knowledge-based technique and used the Lesk approach along with the Hindi wordnet to perform word sense disambiguation in Hindi. Tripathi et al., (2020) proposed a score-based approach for word sense disambiguation problem in Hindi by using Lesk algorithm and calculating the sense score using gloss, hypernym, hyponym and synonym given in the Hindi wordnet. Considering hypernym and hyponym with words which have no available senses may sometimes affect the performance of the algorithm.

There is substantial amount of work done using the Lesk approach for word sense disambiguation but the proposed methodology modifies the Lesk approach to provide more efficiency and uses it along with the Hindi SentiWordNet to improve the performance of the lexicon approach. The proposed method presents a graph-based technique along with the LESK approach which is inspired from the work of (Sinha, R. and Mihalcea, R., 2007).

Proposed Methodology

The proposed method is called as Graph based Modified Lesk Technique and is explained in the Algorithm 1. Preprocessing needs to be done which includes stop word removal and POS tagging. The Hindi Wordnet is organized specifically based on Parts of speech. Hence Pos tagging is required to access the suitable gloss of a particular word from the Wordnet. The input to the algorithm will be words of a particular sentence which will be indicated as candidate vertices of the graph. Next step is to extract the senses from the Wordnet and represent them as nodes of the graph. Once done, similarity between various word senses is calculated using the Adapted Lesk approach. An edge is created between the two-word senses and an appropriate weight is assigned depending on the similarity index and then calculate indegree of all nodes. The node which is having maximum indegree for every word is selected.

Figure 1 shows the graph-based method with an example. Suppose there are three words W1, W2, W3 taken as input. Senses of each word are extracted from the wordnet as shown in the figure. Once the indegree is calculated based on similarity index returned by the Lesk approach, the sense for W1 will be that of $S1^2$ because it has maximum indegree of 6. Similarly, the senses for W2 and W3 will be selected depending on maximum indegree. In cases where indegree match, averaging can be used.

For all the word senses that are chosen, synset ids are matched with the one in the Hindi SentiWordNet and the corresponding scores are obtained.

Algorithm 1: Graph based Modified Lesk technique.

INPUT: Words of a Sentence S (candidate vertices of the graph).
OUTPUT: Positive and negative score of the words.
Step 1: Pick the word senses from Hindi WordNet and represent them as nodes of the graph.
Step 2: Calculate Similarity between different word senses using Modified Lesk approach.

Step 3: Construct an edge between two-word senses and assign weight to it depending on the similarity index returned by the Lesk approach.
Step 4: Calculate indegree of all the nodes in the graph and choose the node (word sense) for every word which has the maximum indegree
Step 5: For the chosen word senses, match its synset_id with the one in the SentiWordNet and extract the corresponding positive and negative score

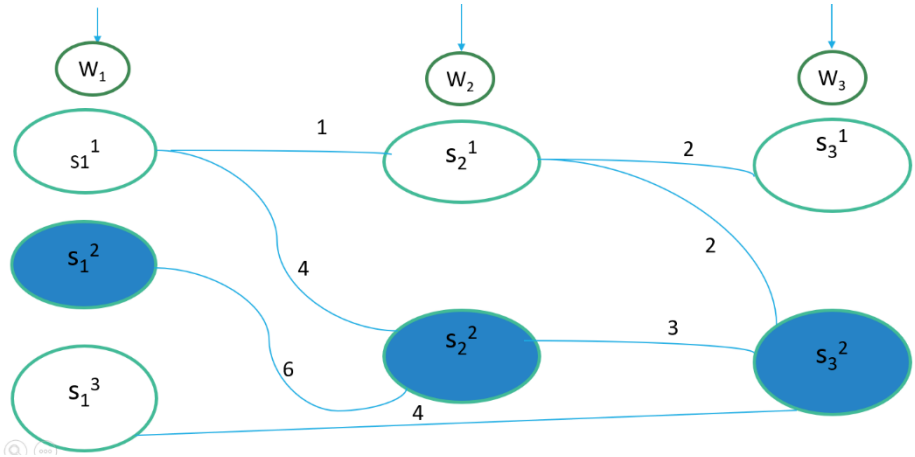

Figure. 1 Graph based method for choosing appropriate wordsense

Results Discussion

To conduct the experiments dataset was developed which contains around 4028 sentences in Hindi that were collected from various sources such as movie reviews, product reviews and travel reviews. Output of the proposed system is evaluated to check its efficiency using four parameters namely Accuracy, Precision, Recall and Fscore It is compared with the simple Lexicon approach as shown in Figure 2. The results show a considerable increase in all the parameters when the proposed approach is used.

The Adapted Lesk approach is based on overlapping and tries to find a match between context bag which is built from the context in which the ambiguous words occur and the sense bag which is built from the gloss of the ambiguous word present in the Hindi Wordnet. The sense that has maximum overlap wins. The adapted Lesk approach is further experimented and modified to include not only glosses to create the sense bag but examples, hypernym, hyponym, meronym and antonym from the wordnet. When the results were compared there was as such no significant change in accuracy, but Recall and F-measure were affected. Incorporating the Hypernym, hyponym of each sense slightly affected the performance of the algorithm When different combinations were tried out it was seen those glosses along with examples could produce a much better output. Hence in the final Modified Optimized Lesk method only glosses and examples were considered.

The improvement in the Lesk technique and the corresponding results are shown in Figure 3.

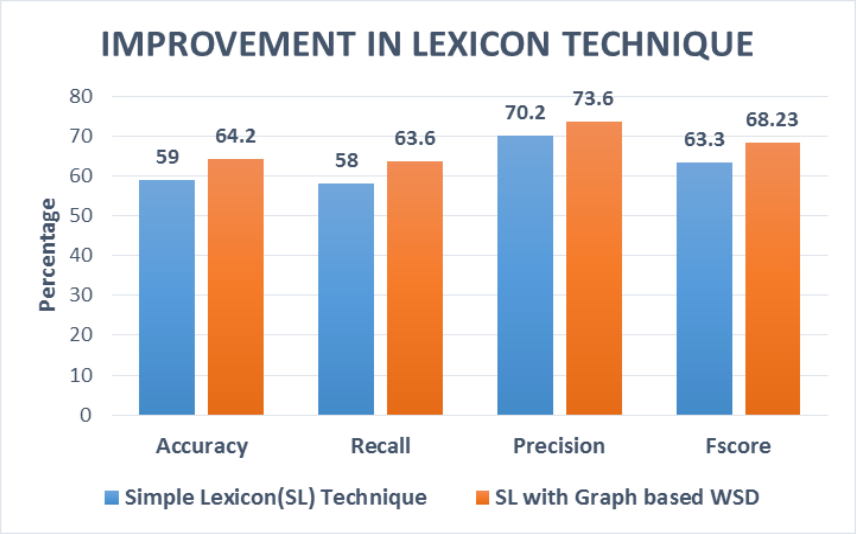

Figure 2 Comparison of simple Lexicon (HSWN) and Graph based WSD lexicon technique

The most commonly used word sense disambiguation techniques with lexicon approach is either taking the first sense or averaging of all senses. The results of HSWN with graph based Lesk approach was compared with using HSWN with first sense and using HSWN with averaging. The comparison results are shown in Figure 4. The HSWN with graph based Lesk approach performs exceedingly well when compared to HSWN with first sense and is equally good as HSWN with averaging technique. The HSWN with graph based Lesk technique thus handles the word sense disambiguation problem in an efficient way and enhances the lexicon approach for performing Hindi sentiment analysis.

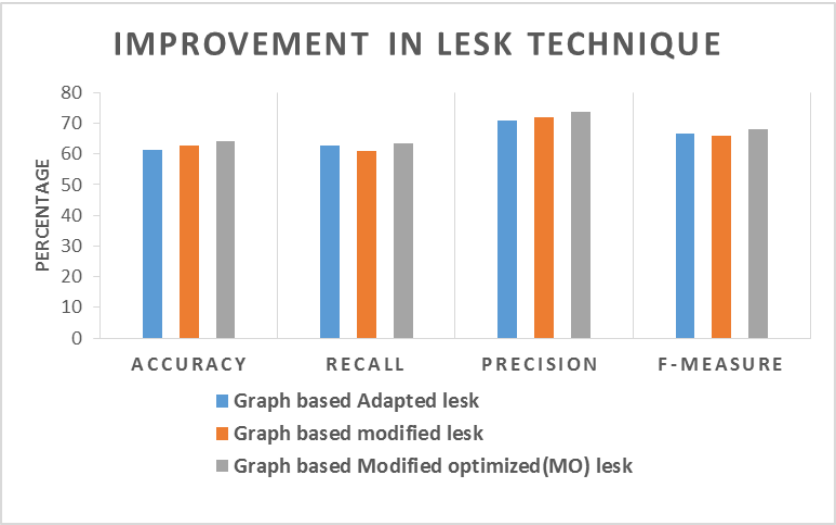

Figure 3 Improvement in LESK technique

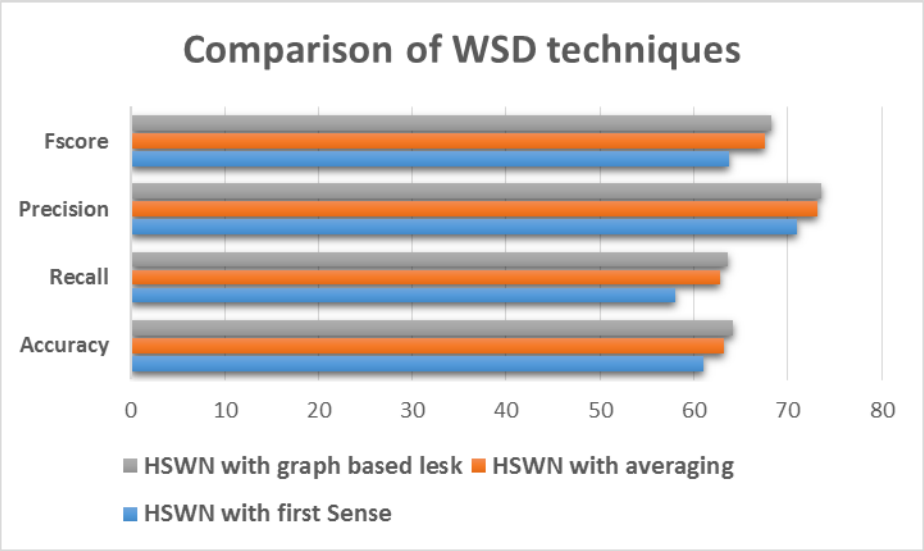

Figure 4 Comparison of word sense disambiguation techniques

Conclusion

Sentiment analysis is an upcoming field of research with very little work done in performing Sentiment analysis in Hindi Language. This paper makes some empirical contributions to Sentiment analysis in Hindi by enhancing the lexicon approach and making efficient use of the HindiSentiWordnet. A new method called HSWN with graph based Lesk approach is introduced that solves the problem of word sense disambiguation and increases the performance of the system. Results show that the proposed approach which uses WSD process significantly enhances the sentiment classification accuracy of simple lexicon approach.

Future work will deal with reducing the time complexity of the proposed approach and improving the lexicon approach by handling problems such as negations, sarcasm, thwarting etc.

References

Pang, B., & Lee, L. (2009). Opinion mining and sentiment analysis. *Computational Linguistics, 35*(2), 311-312.

Augustyniak, Ł., Szymański, P., Kajdanowicz, T., & Tuligłowicz, W. (2016). Comprehensive study on lexicon-based ensemble classification sentiment analysis. *Entropy, 18*(1).

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. *In Proceedings of recent advances in natural language processing (RANLP), 1*(3.1), 2-1.

Sharma, D.K. (2015). A comparative analysis of Hindi word sense disambiguation and its approaches. *In International Conference on Computing, Communication & Automation,* 314-321.

Joshi, A., Balamurali, A.R., & Bhattacharyya, P. (2010). A fall-back strategy for sentiment analysis in Hindi: a case study. *Proceedings of the 8th ICON.*

Garg, K. (2020). Sentiment analysis of Indian PM's "Mann Ki Baat". *International Journal of Information Technology, 12*(1), 37-48.

Singh, S., Singh, V.K., & Siddiqui, T.J. (2013). Hindi word sense disambiguation using semantic relatedness measure. *In International Workshop on Multi-disciplinary Trends in Artificial Intelligence, Springer, Berlin, Heidelberg, 8271,* 247-256.

Jha, V., Manjunath, N., Shenoy, P.D., & Venugopal, K.R. (2016). Sentiment analysis in a resource scarce language: Hindi. *International Journal of Scientific and Engineering Research, 7*(9), 968-980.

Akhtar, M.S., Ekbal, A., & Bhattacharyya, P. (2016). Aspect based sentiment analysis in Hindi: resource creation and evaluation. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 2703-2709.

Bakliwal, A., Arora, P., & Varma, V. (2012). Hindi subjective lexicon: A lexical resource for hindi polarity classification. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC),* 1189-1196.

Jha, V., Manjunath, N., Shenoy, P.D., Venugopal, K.R., & Patnaik, L.M. (2015). Homs: Hindi opinion mining system. *In IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS),* 366-371.

Sharma, P., & Moh, T.S. (2016). Prediction of Indian election using sentiment analysis on Hindi Twitter. *In IEEE international conference on big data (big data),* 1966-1971.

Mittal, N., Agarwal, B., Chouhan, G., Bania, N., & Pareek, P. (2013). Sentiment analysis of hindi reviews based on negation and discourse relation. *In Proceedings of the 11th Workshop on Asian Language Resources,* 45-50.

Gautam, C.B.S., & Sharma, D.K. (2016). Hindi word sense disambiguation using Lesk approach on bigram and trigram words. *In Proceedings of the International Conference on Advances in Information Communication Technology & Computing,* 1-5.

Sharma, D.K. (2016). Hindi word sense disambiguation using cosine similarity. *In Proceedings of International Conference on ICT for Sustainable Development, Springer, Singapore,* 801-808.

Pamungkas, E.W., & Putri, D.G.P. (2017). Word sense disambiguation for lexicon-based sentiment analysis. *In Proceedings of the 9th International Conference on Machine Learning and Computing,* 442-446.

Hussaini, F., Padmaja, S., & Sameen, S. (2018). Score-based sentiment analysis of book reviews in Hindi language. *International Journal on Natural Language Computing, 7*(5), 115-127.

Athaiya, A., Modi, D., & Pareek, G. (2018). A genetic algorithm based approach for Hindi word sense disambiguation. *In 3rd international conference on communication and electronics systems (ICCES),* 11-14.

Sharma, P., & Joshi, N. (2019). Design and Development of a Knowledge-Based Approach for Word Sense Disambiguation by using WordNet for Hindi. *International Journal of Innovative Technology and Exploring Engineering, 8*(3).

Tripathi, P., Mukherjee, P., Hendre, M., Godse, M., & Chakraborty, B. (2020). Word Sense Disambiguation in Hindi Language Using Score Based Modified Lesk Algorithm. *International Journal of Computing and Digital Systems, 10,* 2-20.

Sinha, R., & Mihalcea, R. (2007). Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. *International Conference on Semantic Computing*, 363–369.