Proposed Algorithm of Sentiment Analysis for Punjabi Text

Amandeep Kaur
University Institute of Engineering and Technology, Chandigarh, India
Email: amandeepk.cql@gmail.com

Vishal Gupta
University Institute of Engineering and Technology, Chandigarh, India
Email: vishal@pu.ac.in

***Abstract*— Sentiment Analysis is to identify and classify the opinions/emotions/sentiments in written text. Till date, English Language includes most of the research work in this area. In this paper, we discussed the various approaches used to accomplish the sentiment analysis and research work done for Indian Languages like Hindi, Bengali and Telugu. We proposed an algorithm by using subjective lexicon which is created by using Hindi Subjective Lexicon. Our approach proves good performance on the testing data. We compared the results with already existing approaches.**

***Index Terms*—Sentiment Analysis, Punjabi Language, Subjective Lexicon.**

I. INTRODUCTION

There are 100+ million speakers of Punjabi language spread across the world. The coverage area of this language is also increasing across the web. Web pages contain important information relating to corporate and government. For the research on sentiment Analysis, Punjabi language does not contains much work. Sentiment and opinions can be better considered by focusing on the Adjectives and Adverbs. In this paper, we proposed the method for the usage of subjective lexicon to understand the sentiments in Punjabi text. Our Algorithm depends only on one resource, Subjective Lexicon. We combined the already existing approaches; unigram presence method and simple scoring method. We considered the concept of the synonyms and antonyms. We assumed the similar polarity of synonyms and opposite polarity of antonyms. We tested the proposed approach for Punjabi language using testing data.

Main contribution of our research-

- Developing the Punjabi Subjective lexicon Using the Hindi Subjective Lexicon available at[14]
- Devise an Algorithm Combining the unigram method and simple scoring method which provides the better efficiency.

We have also tested the unigram presence method and simple scoring method and compared the performance with that of our algorithm.

II. RELATED WORK

Research in Sentiment Analysis can be categorized in following levels-

- Document level [17,20]
- Sentence level [10,11,15,21,23]
- Word level [1,22]

A. Non- Indian Languages

In the beginning of the research under the area of sentiment analysis, General Inquirer system [19] was developed by the IBM in 1966. This system is related to the behavior science and having collection of 11789 words. Semantic orientation of adjectives is predicted by method developed by [9] in 1998. ha Turney had done research on POS tags in 2002[20].

For English, a lot of research has been done. Esuli and Sebastiani developed Sentiwordnet [2,8] in 2006.

For under resourced languages, a bootstrapping method is proposed by Banea et. al. [4]. Kamps et. al. [13] had done reasearch work for sentiment analysis by considering the adjectives in wordnet. Kim and Hovy [16] tried to analyse judgement opinions. Rao and Ravichandran [18] performed the semi-supervised label propagation.

B. Indian Languages

For Indian Languages, there is not much amount of research done. For Bengali language, Sentiwordnet is developed by Das and Bandhopaday.[5,3] They used English Sentiwordnet and English-Benagli Dictionary to develop the Bengali Sentiwordnet. In [6], the authors proposed four approaches to classify the sentiment into positive and negative. First strategy, an interactive game predicts the sentiment of a word. Second Strategy is for English and Indian languages by the usage of Bi- Lingual dictionary. Third Strategy added the concept of synonyms and antonyms for Word Net expansion. Fourth strategy used the pre-annotated corpora for the machine learning.

Joshi et. al [12] used English-Hindi Word net Linking and English Sentiwordnet for developing H-SWN (Hindi-Sentiwordnet).

doi:10.4304/jetwi.6.2.180-183

Our work is concentrated on the Punjabi Language. In this work, we used the Hindi WordNet[7] to develop the Subjective Lexicon for the Punjabi language.

III. Approach Used

English Language is processed for most of the research in Natural Language Processing. Most popular approaches used in the area of sentiment analysis are-

- Subjective Lexicon,
- N-gram Modeling,
- Machine Learning.

Out of these approaches, we have worked by using the subjective lexicon.

A. Subjective Lexicon

Punjabi Language is very scarce because of the lack of limited resources developed till now. Basically, three popular methods are used for the generation of subjective lexicon-

- Use of Bi-Lingual Dictionary[6],
- Machine Translation[6],
- Use of Word net [7].

We have chosen the method by using Bi-Lingual Dictionary. In this technique, Translation process is applied at word level on the Hindi Subjective Lexicon which is also called Hindi Sentiwordnet developed by P. Arora in 2013.[7] The resultant lexicon is refined by adopting various techniques of error reduction. The developed lexicon is introduced with the concepts of synonyms and antonyms. Every entry of the lexicon is categorized into four parts of speech-

- Noun
- Verb
- Adjective
- Adverb

Structure of developed lexicon is formulated by examples in Table I.

TABLE I.
STRUCTURE OF SUBJECTIVE LEXICON

Part of speech	Positive	negative	words
n	0.0	1.0	ਪਰੇਸ਼ਾਨ ਉਦਾਸ ਨਾਖੁਸ਼
v	1.0	0.0	ਸ਼ਲਾਘਾ ਉਪਮਾ ਉਸਤਤ ਪ੍ਰਸ਼ੰਸਾ
a	.05	0.95	ਗੰਦਾ ਮੈਲਾ ਖ਼ਰਾਬ ਘਟੀਆ
r	0.40	0.60	ਘੱਟ ਥੋੜ੍ਹਾ ਕੁੱਝ ਮਾਮੂਲੀ

It also possesses the features of word net for better understanding the contextual information.

B. Stemming

The concept of stemming is used to consider the stemmed variant of a word. Table II is given with some words, which need stemming and their root words.

TABLE IV
OVERALL RESULTS HINDI AND PUNJABI

Language	Precision	Recall	F-Ratio
Punjabi	0.78	0.60	0.67
Hindi	0.81	0.65	0.72

C. Language Specific Words

There are various culture specific words of Punjabi language which were not present in the Hindi Sentiwordnet. So, to capture these Language specific words, we have manually developed a seed list of these words and tagged with Punjabi Specific Corpus. Example of seed list is given in table.

D. Negation Handling

There are certain words which are categorized as negation words like- **ਨਹੀਂ**, **ਨਾਂਹ**. These words invert the polarity of the sentence. So, a list of these words is prepared manually to tackle this concept. Example

IV. Methodology

A. Algorithm is given Using Subjective Lexicon

Step 1. Input a text paragraph.

Step 2. Divide and conquer:

- Divide the text into n- grams on the basis of full stop.
- Further Sub divide the n– grams into sub grams on the basis of Separators like comma, semi colon, or other conjunctive words (**ਤਾਂ**, **ਇਸ ਲਈ**, **ਕਿਉਂਕਿ**, **ਜੇ**)
  - Use tree data structure with 3 levels and height=2. Parent node as input text. Internal nodes as grams and leaf nodes as sub grams.

Step 3. Preprocessing phase:

- Remove stop words
- Remove extra symbols
- Perform Stemming

Step 4. Feature Extraction phase:

- Extracting Keywords: nouns, adjectives, adverb and verb.

Step 5. Use Subjective Lexicon:

- Assign polarities to all keywords having the range 0.0 to 1.0

Step 6. Remove the objective information:

- The sub gram which does not contain any polarity is considered as objective information having neutral polarity.

Step 7. Compute the overall polarity of a sub gram:

- Sum up the positive and negative polarity
- Choose the dominating polarity for the respective sub gram.

Step 8. If dominating polarity is positive then follow the rules below:
- Positive polarity must have value at least 0.5 more than the negative. If this is not true then assign the negative polarity to the sub gram.
- If value of negative and positive polarity is equal, then assign negative polarity to the sub gram.

Step 9. Handling Negations: The sub gram that contains negation words (**ਨਹੀਂ**, **ਨਾਂਹ**), invert the polarity of that sub gram.

Step 10. Final Output using iterative Process and Bottom up approach:
- Repeat step 8 for each gram by taking its respective child sub grams.
- Again Repeat step 8 for the parent node i.e input text. Polarity of the parent node determines the overall polarity of the text.

V. Result Evaluation

Results of our approach are compared with the following approaches-

A. Unigram Presence Method-

In this method, count the words of positive and negative polarity in the text and choose the polarity with highest count.

B. Simple Scoring Method-

In this method, we sum up the positive and negative score of each word and choose the polarity with dominant score.

We have tested our approach manually on the testing data by collecting documents written in Punjabi language. We apply the three approaches – our approach, unigram presence method and simple scoring method. We also perform the negation handling and stemming. Table III and Table IV highlights the results computed which shows the better performance of our approach over the already existing approaches and comparison with the analysis of Hindi Language.

TABLE III
ACCURACY RESULTS

Method	Our Algorithm	Unigram presence	Simple Scoring method
Subjective lexicon	54.2%	47.90%	48.30%
+ Negation handling	67.50%	55.25%	53.98%
+Stemming	78.02%	61%	57.45%

VI. Conclusion and Future Work

The approach proposed by us achieved better accuracy but still performance is low. In this research, we found a lot of hurdles which contribute towards the low performance and these can be the part of future research-

A. Lexicon Coverage-

Our base lexicon developed for Hindi language has limited coverage. So, Subjective Lexicon can be combined with the Machine learning so that train the system which can automatically classify the documents into respective polarity.

B. Context Dependency-

Lexicons fail to relate the meaning of word with the other words in the sentence, so lead to problem of contextual dependency. The solution for this problem can be adding the contextual information with the dynamic prior polarity.

C. Vocabulary Mismatch-

Different people belong to different culture, so there is diversity in the vocabulary composition. Morphological Analysis of the words can be done to get the root word.
We focused our research on the use of subjective lexicon which can be further extended up to use of n – gram modeling, machine learning and combination of these.

References

[1] A. Agarwal, F. Biadsy, and K. R. Mckeown, "Contextual phrase-level polarity analysis using lexical a_ect scoring and syntactic n-grams", 2009.

[2] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining" *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010.

[3] A. Das, "Opinion Extraction and Summarization from Text Documents in Bengali", Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., Jadavpur University, 2011.

[4] R. M. Carmen Banea and J. Wiebe, "A bootstrapping method for building sub-jectivity lexicons for languages with scarce resources" In B. M. J. M. J. O. S. P.D. T. Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Mar- rakech, Morocco, may 2008.

[5] A. Das and S. Bandyopadhyay, "*SentiWordNet for Bangla*. 2010".

[6] A. Das and S. Bandyopadhyay, ' *SentiWordNet for Indian Languages*" 2010.

[7] P. Arora, "Sentiment Amalysis for Hindi Language" Masters thesis, IIT, Hyderabad, 2013.

[8] A. Esuli and F. Sebastiani. Sentiwordnet, "A publicly available lexical resource for opinion mining", *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06*, pp. 417-422, 2006.

[9] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives", *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pp. 174-181, Stroudsburg, PA, USA, 1997.

[10] M. Hu and B. Liu, "Mining and summarizing customer reviews" In *KDD*, pp. 168-177, 2004.

[11] T. W. Intelligent and T. Wilson, "Annotating opinions in the world press", *SIGdial-03*, pp. 13-22, 2003.
[12] A. Joshi, B. A. R, and P. Bhattacharyya, "A fall-back strategy for sentiment analysis in hindi: a case study", 2010.
[13] J. Kamps, M. Marx, R. J. Mokken, and M. D. Rijke, "Using wordnet to measure semantic orientation of adjectives", *National Institute for*, pp. 1115-1118, 2004.
[14] Indian Institute of Technology, Hyderabad http://www.iith.ac.in/.
[15] S. min Kim, "Determining the sentiment of opinions" *Proceedings of COLING*, pp. 1367-1373, 2004.
[16] S. min Kim and E. Hovy, "Identifying and analyzing judgment opinions" *Proceedings of HLT/NAACL-2006*, pp. 200-207, 2006.
[17] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classi_cation using machine learning techniques" *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79-86,2002.
[18] D. Rao and D. Ravichandran, " Semi-supervised polarity lexicon induction" *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pp. 675-682, Stroudsburg, PA, USA, 2009.
[19] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie, "The General Inquirer: A Computer Approach to Content Analysis*" MIT Press*, Cambridge, MA, 1966.
[20] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classi_cation of reviews", 2002.
[21] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara, "Development and use of a gold- standard data set for subjectivity classi_cations" *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Lin-guistics*, ACL99, pp. 246-253, Stroudsburg, PA, USA, 1999.
[22] T. Wilson, "Recognizing contextual polarity in phrase-level sentiment analysis" *Proceedings of HLT-EMNLP*, pp. 347-354, 2005.
[23] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences" *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pp. 129-136, Stroudsburg, PA, USA, 2003.