A Novel Approach for Sentiment Analysis of Punjabi Text using SVM

Amandeep Kaur and Vishal Gupta
Department Computer Science and Engineering, Panjab University, India

Abstract: *Opinion mining or sentiment analysis is to identify and classify the sentiments/opinion/emotions from text. Over the last decade, in addition to english language, many indian languages include interest of research in this field. For this paper, we compared many approaches developed till now and also reviewed previous researches done in case of indian languages like telugu, Hindi and Bengali. We developed a hybrid system for Sentiment analysis of Punjabi text by integrating subjective lexicon, N-gram modelling and support vector machine. Our research includes generation of corpus data, algorithm for Stemming, generation of punjabi subjective lexicon, developing Feature set, Training and testing support vector machine. Our technique proves good in terms of accuracy on the testing data. We also reviewed the results provided by previous approaches to validate the accuracy of our system.*

Keywords: *Sentiment analysis, subjective lexicon, punjabi language, n-gram modeling, support vector machine.*

Received June 17, 2014; accepted December 16, 2014

1. Introduction

Although Today, Punjabi is widely used and well spoken language in the various parts of world. This language has 100+ million speakers and wide coverage area across the web. But the thing which is scarce is the resources and tools to do successive research for this language. In this research, we used some tools of native and other languages such as Hindi subjective Lexicon developed by Piyush [5], Punjabi dictionary which contains 35031 words. We combined the features of Lexicon and N-gram modelling to train support vector machine and developed a classification model which will further used for classifying the text. In our work, we tried to provide following contributions as research work:

1. Punjabi subjective lexicon is developed using Punjabi dictionary and Hindi Subjective Lexicon provided by [22].
2. Developed corpus dataset for training and testing the system.
3. Generation of Algorithm for stemming using the rules given by Dr. Gupta [19].
4. Devised an Algorithm for the use of Subjective lexicon by combining the simple scoring method and unigram method which provides the better efficiency.
5. Generation of feature set for better identification of type of sentiment class.
6. Hybrid System for Punjabi sentiment classification by integration of Subjective lexicon, N-gram modelling and Support vector machine.
7. Training and testing support vector machine to resolve the error.
8. Machine learning is implemented through SVM.

2. Related Work

Through the literature survey [27], we have come up with the following levels at which research of Sentiment analysis has been done:

- Document level [32, 43]
- Sentence level [21, 24, 28, 45, 47]
- the Word level [2, 46]

2.1. Non-Indian Languages

With the research of General Inquirer system [42] in 1966, IBM has begun the research under this area. Selection of features for this kind of research is explained by khan *et al*. [8]. In the area of research for English language, most of research has been done by Sebastiani and Esuli. In 2006, they come up with great lexical resource named Sentiwordnet [6, 18].

For languages having scarcity of resources, a bootstrapping method was given by Banea *et al*. [7]. Rao and Ravichandran [37] developed the semi supervised label propagation. Kim and Hovy [29] did research to analyse opinions or judgement. Kamps *et al*. [26] used adjectives given in wordnet to do research in the field of sentiment analysis.

2.2. Indian Languages

In case of Indian Languages, major issue is limitation of resources. For Bengali language, Das and

Bandhopaday did a lot of research and developed Sentiwordnet [13, 14]. Joshi *et al*. [25] worked for Hindi language. They used English Sentiwordnet and English-Hindi Word net Linking for developing H-SWN (Hindi-Sentiwordnet). Our research work is related to Punjabi Language.

3. Approach Used

In Natural Language Processing, English Language is taken as test language to apply different techniques. Following are the common approaches used in this field:

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

Using the above approaches, we have developed a Hybrid System by combining Subjective Lexicon, N-grams and Support Vector Machine. Implementation is done using Following Framework tools:

- Weka: It is java based toolkit developed at University of Waikato, New Zealand for machine learning [44].
- Naïve Bayes: This classifier uses Bayes Theorem which relates the probabilities of event given and event already occurred.
- Support Vector Machine: SVM classifier separates the dataset into classes by constructing N-dimensional hyper plane. WLSVM and LibSVM are the built-in libraries which we have used.

3.1. Stemming

To get the maximum coverage, we used the concept of stemming. Root words and stemmed words are given Table 1. We have formulated the following Algorithm for stemming:

Table 1. Stemming.

Word	Root Word	Suffixes
ਵੱਡੀ	ਵੱਡਾ	ਾ
ਸੁੰਦਰਤਾ	ਸੁੰਦਰ	ਤਾ
ਕੌੜੀ	ਕੌੜਾ	ੀ

Algorithm for Stemming:

```
1. Develop Array of Suffixes which we have to be identify and
   remove from the word to get a Root Word. List of suffixes is
   collected from research By Dr. Gupta [19]. Say this Array as
   SuffixArray.
2. Boolean = Check wordtobeStemmed Exists in the
   Dictionary.
3. If Boolean = true then no Stemming required.
Else
RootWord=getRootWord(WordtobeStemmed)

getRootWord (WordtobeStemmed)
{
 Position = 2
 Length = length of WordtobeStemmed
Loop
Position < length
Suffix = Split the word at position
Flag = check if Suffix exits in the SuffixArray developed in
       Step 1.

If Flag = true
Output= ReplaceSuffix (WordtobeStemmed,Suffix)
Return Output
 End if

 If flag = false and position = length-1
 Return WordtobeStemmed // no stemming possible
 End if
   Position = Position + 1
 End loop
 }

 ReplaceSuffix (WordtobeStemmed , Suffix)
{
 Develop two arrays to determine replacement of suffix with
       kanna (ਾ) or BadiE (ੀ):
 KannaArray = { "ੇ","ੋ","ਿਆ","ਿਓ","ਿਓੁਂ"}
 BadiEArray = {"ਿਏ","ਿਓ"}
 If  Suffix lies in KannaArray
 Then Replace the suffix with Kanna (ਾ)
 Else if (Suffix lies in BigEArray)
 Then Replace the suffix with BadiE (ੀ)
 Else
 Remove the suffix from WordtobeStemmed
 End if
 Return Root_word
}
```

3.2. 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 [15].
- machine translation [15].
- use of word net [5].

In Our generation of Lexicon for Punjabi language, Translation process is applied at word level on the Hindi Subjective Lexicon developed by Arora in 2013. Arora [5] The resultant lexicon is refined by adopting various techniques of error reduction. We have also used Punjabi dictionary (contains 35031 words) to add the words which were not covered by the Hindi Subjective Lexicon. In the resultant Subjective lexicon, we have come with total of 7860 words. The developed lexicon is introduced with the concepts of synonyms and antonyms.

Every entry of the lexicon is tagged with one of the following four parts of speech [20]:

- Noun
- Verb

- Adjective
- Adverb

Structure and Statistics of developed lexicon is formulated in Table 2.

Table 2. Structure and Statistics of Subjective Lexicon.

	Lexicon Statistics (Total = 7860)		**Example**		
Part of speech	Positive	Negative	Pos polarity	Neg polarity	Words
noun	2448	2091	0.0	1.0	ਪਰੇਸ਼ਾਨ ਉਦਾਸ
verb	1203	939	1.0	0.0	ਸ਼ਲਾਘਾ ਉਪਮਾ
adjective	398	299	.01	0.99	ਗੰਦਾ ਮੈਲਾ
adverb	251	231	0.40	0.60	ਘੱਟ ਥੋੜ੍ਹਾ

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

- *Step* 1. Input a text paragraph.
- *Step* 2. Divide and conquer:
  1. Divide the text into n- grams on the basis of full stop.
  2. 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)
  3. 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:
  1. Remove stop words
  2. Remove extra symbols
  3. Perform Stemming
- *Step* 4. Feature Extraction phase:
  1. Extracting Keywords: nouns, adjectives, adverb and verb.
- *Step* 5. Use Subjective Lexicon:
  1. Assign polarities to all keywords having the range 0.0 to 1.0
- *Step* 6. Remove the objective information:
  1. 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:
  1. Sum up the positive and negative polarity
  2. Choose the dominating polarity for the respective sub gram.
- *Step* 8. If dominating polarity is positive then follow the rules below:
  1. 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.
  2. 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:
  1. Repeat step 8 for each gram by taking its respective child sub grams.
  2. Again Repeat step 8 for the parent node i.e input text. Polarity of the parent node determines the overall polarity of the text.

3.3. Negation Handling

There are certain words which are categorized as negation words like-"ਨਹੀਂ", "ਨਾਂਹ", "ਨਾ", "ਨਾਮਾਤਰ", "ਨਾਮੁਨਿਕਨ", "ਨਹੀਂ". Polarity of the sentence gets inverted due to the presence of these words. A list of these kinds of words is maintained to face this issue.

3.4. N-grams

In this technique, Following steps are carried out:

1. Dataset is divided into Training set and Testing set.
2. Using the Training set, we developed an N-gram model.
3. For each input from Testing set, create the trigrams
4. Match these trigrams with trigrams of training data using the N-gram model
5. If matches then increment the value of trimatched (variable name) else create the bi- grams and repeat the previous step for bigrams and unigrams and increment the value of respective match.

3.5. Feature Set

Features using subjective lexicon and n-grams are constructively shown in Table 3.

Table 3. Feature set.

Approach	**Feature**
Subjective Lexicon	Positive
	Negative
n-grams	No. of unigrams(Positive)
	No. of bigrams(Positive)
	No. of trigrams(Positive)
	No. of unigrams(Negative)
	No. of bigrams(Negative)
	No. of trigrams(Negative)

Increase in the number of features directly affects the increase in the accuracy of system shown in Figure 1. So, we used eight features for maximum coverage. Positive and negative features are taken in terms of subjective lexicon and n-grams.

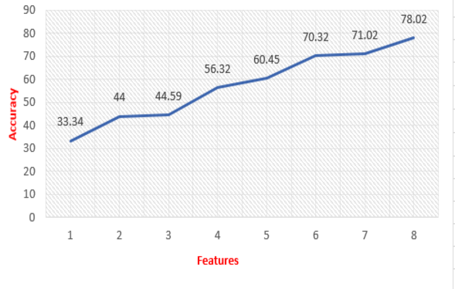


Figure 1. Increasing accuracy with increase in number of features.

3.6. Corpus Generation

Punjabi dataset is generated by collecting data from Punjabi newspapers [3, 4, 10, 11, 12, 16, 17, 23, 30, 31, 33, 34, 35, 38, 39] and Blogs [1, 9, 36, 40, 41]. Statistics of Resultant Dataset is given in Table 4.

Table 4. Statistics: punjabi corpus.

	Training	Testing	Newspapers	Blogs
Number of documents	600	284	721	163
Number of sentences	30000	14200	36050	8150
Number of sentences per document(avg)	50	50	50	50
Number of words(total)	21000	99400	252350	57050
Number of words per document	359	359	350	350

3.7. Human Annotation

For accuracy of polarity assigned in the Subjective Lexicon, we hired 3 human annotators who have provided their agreement/disagreement regarding the polarity.

4. Methodology

Process flow of system is described by following Algorithm given below.

Algorithm:

1. *Compute the positive and negative score using subjective lexicon described in Section B. Compute the no. of positive and negative trigrams, bigrams, and unigrams using n- gram modelling explained in Section E.*
2. *Compute the no. of positive and negative trigrams, bigrams, and unigrams using n- gram modelling explained in Section E.*
3. *Devise the feature vector which will include the information of step1 and 2. The design of Feature Vector is explained in Table 3.*
4. *This Feature Vector is used for Training and Testing Support vector machine by using Training and Testing datasets.*

5. Result Evaluation

Implementation of system is done using Java. Figure 3 shows the user interface, input and output of the system. Input is Punjabi text in Unicode format or File can be browsed for input which should be saved by setting encoding is equal to UTF-8. Our System takes the input and provide the result corresponding to all combinations of Lexicon and N-gram and hybrid system. Output also include the Graph which shows the comparison of processing time taken by all combinations and Hybrid System. Graph concludes that Hybrid System always takes lowest processing time. Results of our technique are evaluated to compare with the following techniques in Figure 2:

- Hindi Subjective Lexicon
- Hindi Sentiwordnet
- Bi-lingual Dictionary
- Translated Dictionary

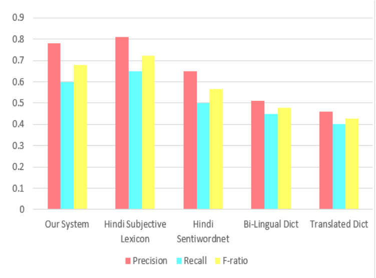
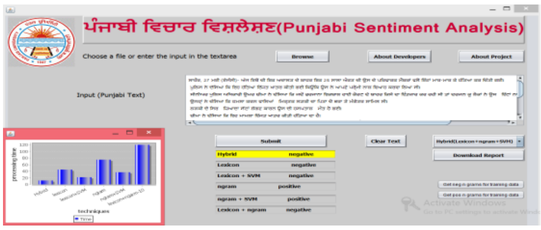


Figure 2. Comparison of our system with the existing resources.


ਪੰਜਾਬੀ ਵਿਚਾਰ ਵਿਸ਼ਲੇਸ਼ਣ(Punjabi Sentiment Analysis)
Choose a file or enter the input in the textarea
Browse
About Developers
About Project
Input (Punjabi Text)
Submit
Clear Text
Hybrid(Lexicon+ngram+SVM)
Download Report
Get neg n grams for training data
Get pos n grams for training data
processing time
120
100
80
60
40
20
0
Hybrid
lexicon
lexicon/SVM
ngram
ngram/SVM
lexicon+ngram-10
techniques
Time

Hybrid negative
Lexicon negative
Lexicon + SVM negative
ngram positive
ngram + SVM positive
Lexicon + ngram negative

Figure 3. Snapshot showing Output of the System.

Results are compared by calculating parameters (Precision, Recall, F-measure and Accuracy). Table 6 shows the results computed and comparisons which conclude the better performance of our system. Confusion Matrix is created using SVM to conclude the performance of system shown in Table 5.

Table 5. Statistics: confusion matrix.

	Positive	Negative	Neutral
Positive	259	32	9
Negative	69	223	8
Neutral	56	21	207

Table 6. Accuracy results.

Method	News	Blogs	Naïve Bayes	SVM
Hindi Sentiwordnet	67.50	68.04	58.42	61.11
Hindi Subjective Lexicon	81.28	71.4	58.6	61.56
Bi-lingual Dictionary	70.30	67.05	58.21	61.58
Translated Dictionary	62.30	61.05	58.38	61.6
Our System				
Lexicon	54.2	56.1	55.43	60.02
Lexicon + Negation handling	67.50	59.2	62.3	71.34
Lexicon + Negation handling +Stemming	73.45	68.1	68.52	74.56
N-gram	30.2	32.31	32.9	33.53
N-gram+ Negation Handling	34.1	41.21	41.84	42.1
N-gram+ Negation handling +Stemming	42.41	46.54	47.33	48.02
Lexicon+ N-gram+ Negation handling +Stemming	71.23	70.34	72.67	78.02

6. Conclusions and Future Work

The system developed as part of our research, proves better performance but still efficiency is compromised. The factors which contribute towards low efficiency, can be considered for future research-

1. Coverage by lexical resouces: lexicon used by us is deprived of concept of understanding the contextual information. So, Machine learning can be used in the integration with the Subjective Lexicon for the training of the system.
2. Contextual issue-Meaning of word is independent from that of other word in the sentence so Lexicons fail at this particular point. This problem is also categorized as contextual dependency. To solve this problem dynamic prior polarity can be added.
3. Vocabulary mismatch- vocabulary composition differs due to Cultural anthropology which states that the difference in the culture based on location, society, environment, and people. This problem can be solved by using the concept of Morphological Analysis.

We have done our research by the using of subjective lexicon, support vector machine, n -gram modeling and combination of these. This research can be further extended by adding more features like:

- Word length
- Punctuation
- Vocabulary Richness
- Frequency of function words
- Phrase patterns

References

[1] Aarsi, www.punjabiaarsi.blogspot.in, Last Visited 2014.

[2] Agarwal A., Biadsy F., and Mckeown K., "Contextual Phrase-Level Polarity Analysis using Lexical a_ect Scoring and Syntactic n-Grams," *in Proceedings of 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, pp. 24-32, 2009.

[3] Aj A., www.ajdiawaaz.com, Last Visited 2014.

[4] Ajit, www.beta.ajitjalandhar.com, Last Visited 2014.

[5] Arora P., "Sentiment Analysis for Hindi Language," Master's thesis, Hyderabad, 2013.

[6] Baccianella S., Esuli A., and Sebastiani F., "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion Mining," *in Proceeding of 7th conference on International Language Resources and Evaluation (LREC'10)*, Malta, 2010.

[7] Banea C., Mihalcea R., and Wiebe J., "A bootstrapping method for building sub-jectivity lexicons for languages with scarce resources," *in Proceeding of 6th International Language Resources and Evaluation (LREC'08)*, Marrakech, 2008.

[8] Khan K., Baharudin B., Khan A., "Identifying Product Features from Customer Reviews Using Hybrid Dependency Patterns," *The International Arab Journal of Information Technology*, vol. 11, no.3, pp. 281-286, 2014.

[9] Blogger, Kamal K., www.kamalkang.blogspot.in, Last Visited 2014.

[10] Chardhikala, www.chardhikala.com, Last Visited 2014.

[11] Daily J., www.dailyjanjagriti.com Last Visited 2014.

[12] Daily T., www.dailypunjabtimes.com, Last Visited 2014.

[13] Das A., "Opinion Extraction and Summarization from Text Documents in Bengali," Doctoral Thesis, Jadavpur University, 2011.

[14] Das A. and Bandyopadhyay S., *SentiWordNet for Bangla*, the Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary, 2010.

[15] Das A. and Bandyopadhyay S., "SentiWordNet for Indian Languages," *in Proceeding of 8th Workshop on Asian Language Resources*, Beijing , pp. 56-63, 2010.

[16] Desh S., www.deshsewak.in, Last Visited 2014.

[17] Desh T., www.deshvideshtimes.com, Last Visited 2014.

[18] Esuli A. and Sebastiani F., "Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining," *in Proceedings of 5th Conference on Language Resources and Evaluation*, Genoa, pp. 417-422, 2006.

[19] Gupta V., *Advances in Signal Processing and Intelligent Recognition System*, Springer, 2014.

[20] Hatzivassiloglou V. and McKeown K., "Predicting the Semantic Orientation of Adjectives," *in Proceeding of 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, pp. 174-181, 1997.

[21] Hu M. and Liu B., "Mining and Summarizing customer Reviews," *in Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, pp. 168-177, 2004.

[22] Indian Institute of Technology, http://www.iith.ac.in/, Last Visited 2014.

[23] Indo T., www.indotimes.com.au, Last Visited 2014.

[24] Intelligent T. and Wilson T., "Annotating opinions in world press", *In proceeding of the 4th ACL SIGdial Workshop on Discourse and Diaglogue*, Sapparo, pp. 13-22, 2003.

[25] Joshi A., Balamurali R., and Bhattacharyya P., "A fall-back strategy for sentiment analysis in Hindi: a case study," *in Proceeding of 8th*

International Conference on Natural Language Processing, India, pp.1-6, 2010.

[26] Kamps J., Marx M., Mokken R., and Rijke M., "Using Wordnet to Measure Semantic Orientation of Adjectives," *in Proceeding of 4th International Conference on Language Resources and Evaluation*, Lisbon, pp. 1115-1118, 2004.

[27] Kaur A. and Gupta V., "Proposed Algorithm of Sentiment Analysis for Punjabi Text," *Journal of Emerging Technology in Web Intelligence*, vol. 6, no. 2, pp. 180-183, 2014.

[28] Kim S., "Determining the Sentiment of Opinions" *in Proceedings of 20th International Conference on Computational Linguistics*, Geneva, pp. 1367-1373, 2004.

[29] Kim S. and Hovy E., "Identifying and Analyzing Judgment Opinions," *in Proceedings of Human Language Technology Conference/Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, New York, pp. 200-207, 2006.

[30] Malwa post, www.malwapost.com, Last Visited 2014.

[31] Nawan Z., www.nawanzamana.in, Last Visited 2014.

[32] Pang B., Lee L., and Vaithyanathan S., "Thumbs up? Sentiment Classification using Machine Learning Techniques," *in Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing*, Pennsylvania, pp. 79-86, 2002.

[33] Punjabi Tribune, www.punjabitribuneonline.com , Last Visited 2014.

[34] Punjab post, www.punjabpost.in, Last Visited 2014.

[35] Punjab Info line, www.punjabinfoline.com, Last Visited 2014.

[36] Punjab Screen online, www.punjab-screen.blogspot.in, Last Visited 2014.

[37] Rao D. and Ravichandran D., "Semi-supervised Polarity Lexicon Induction," *in Proceeding of 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, pp. 675-682, 2009.

[38] Rojana S., www.rozanaspokesman.com, Last Visited 2014.

[39] Sea Punjab, www.seapunjab.com, Last Visited 2014.

[40] Shabadan p., www.parchanve.wordpress.com, Last Visited 2014.

[41] Shabad S., www.shabadsanjh.com, Last Visited 2014.

[42] Stone P., Dunphy D., Smith M., and Ogilvie D., *The General Inquirer: A Computer Approach to Content Analysis*, MIT Press, 1966.

[43] Turney P., "Thumbs up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, pp. 417-424, 2002.

[44] University of Waikato, www.cs.waikato.ac.nz/ml/weka/, Last Visited 2014.

[45] Wiebe J. M., Bruce R., and O'Hara T., "Development and use of a Gold- Standard Data Set for Subjectivity Classifications," *in Proceeding of 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, Stroudsburg, pp. 246-253, 1999.

[46] Wilson T., "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," *in Proceedings of HLT-EMNLP*, Vancouver, pp. 347-354, 2005.

[47] Yu H. and Hatzivassiloglou V., "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences," *in Proceedings of 2003 conference on Empirical methods in natural language processing*, Stroudsburg, pp. 129-136, 2003.

Amandeep Kaur has completed her Masters degree in Computer Science & Engineering from Panjab University, Chandigarh, in 2014. Her research interests are Natural Language processing, Image Processing, Machine learning, Computer Vision.

Vishal Gupta is Sr. Assistant Professor in Computer Science and Engineering at University Institute of Engineering and Technology, Panjab University Chandigarh. He has written around 70 research papers in international and national journals and conferences. He has developed a number of research projects in field of NLP and text Mining like keywords extraction system, automatic question answering and text summarization system etc.