International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), Nov 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Opinion Mining and Sentiment Analysis of Punjabi Text- A Review

Gurmeet Kaur

University College of Computer Applications, Guru Kashi University, Bathinda, Punjab, (India)

ABSTRACT

Opinions are central to almost all human beings and are the key factor of our social behavior. The opinions in writing including sentiments, evaluations, emotions, attitudes are the subject of study of Sentiment Analysis and Opinion Mining. Basically, sentiment analysis is an automatic information retrieval from the opinions generated by the users. However, a little research work has been done in this area for the Indian regional languages. Over the years, outstanding growth could be observed in the use of regional languages on the Web in the form of blogs, opinions, tweets, hashtags, news and reviews etc. This paper provides an analysis of the research work carried for mining the opinions expressed in Punjabi language.

Keywords –Opinion Mining, Punjabi Language, Natural Language Processing (NLP), Sentiment Analysis, Emotion Detection.

I. INTRODUCTION

Opinion Mining is a type of Natural Language Processing (NLP) for automatic extraction of positive, negative and neutral opinion or attitude of the writer. The core task in sentiment analysis is determining the polarity of given text at the document level, sentence level or the feature/aspect level and it could be an emotional state also such as 'angry', 'happy' or 'sad'. Even opinions are remarkably of two types i.e. regular opinions e.g. "Toyota Corolla is good car" and comparative opinions e.g. "The Toyota Corolla is not as good as Honda Civic". The whole social media research today revolves around sentiment analysis. Opinion holder, object and opinion are the three basic components of an opinion.

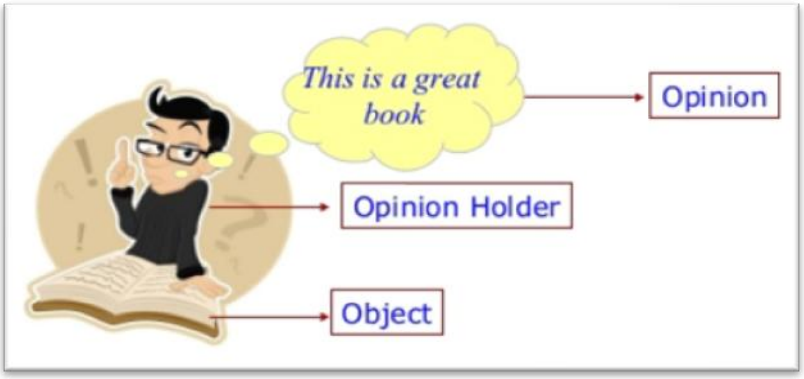

Figure 1 Basic Components of an opinion.

The terms 'opinion mining' and 'sentiment analysis' basically represent the same field of study and are interchangeably used in this paper. Different techniques used for sentiment analysis are supervised learning, unsupervised learning and rule based learning.

Abundance of research work has already been carried out in mining opinions written in English language but the work for Indian languages is at early level of research. Punjabi is an Indo-Aryan language spoken by over 100 million native speakers in the world and is the native language of the Punjab state of India. It holds the position of fifth most spoken language in Canada. There are two major writing systems prevalent for Punjabi i.e. Gurmukhi and Shahmukhi. Punjabi lexicon has the influence of English due to addition of new areas from diverse areas. The grammar of the Punjabi language is the study of the word order, case marking, verb conjugation and other morphological and syntactic structures of the Punjabi language.

II. LITERATURE SURVEY

This paper works for analyzing research work done for mining sentiments written in Punjabi language. There are many challenges and issues faced in mining Punjabi text due to scarcity of adequate resources.

Gupta and Kaur [1] proposed developing the Punjabi Subjective Lexicon using the Hindi Subjective Lexicon already developed by Arora [2] and devising an algorithm combining the unigram method and simple scoring method which provided the better efficiency and accuracy of 54.2%. But they found that the hurdles of lexicon coverage, context dependency and the vocabulary mismatch are still to overcome. They used the concept of synonym and antonyms with similar polarity.

Sharma [3] tried to explore and analyze the Naïve Bayes classification methods based on supervised learning technique and used unigram, bigram and combination of both techniques with a conclusion that combining unigram and bigram techniques gives better accuracy.

Kaur and Saini [4] used various techniques Support Vector Machine, Naïve Bayes, k-nearest neighbor, decision tree and other lexicon based approaches for sentiment classification written in different language families (Indo-Aryan, Dravidian and Tibeto-Burman) and World language (English). Performance accuracy achieved in sentiment classification in Indian languages and English language is shown in Fig.2

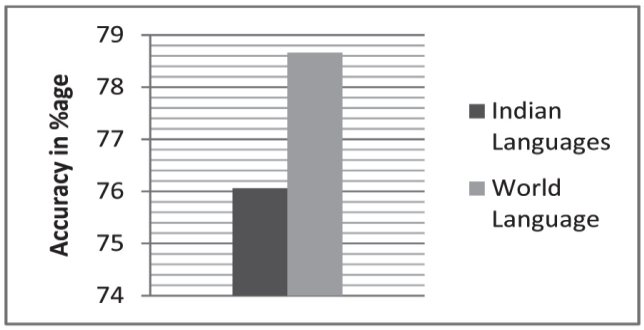

Figure 2 Accuracy in opinion mining task in Indian Languages and World language

Kaur and Gupta [5] developed Punjabi dataset corpus for training and testing the system by collecting the data from Punjabi newspapers and blogs and developed a hybrid system for Punjabi sentiment classification by integrating subjective lexicon, N-gram modelling and Support Vector Machine. They used Java language for implementation of the system.

Kaur and Gupta [6] used Weka module for implementing the classifier named Naïve Bayes technique is used as machine learning technique and is integrated with N-gram model which is used for the extraction of features provided for the training of system. The trained model is then validated using the tested data. Stemming algorithm was also generated.

Grover and Verma [7] presented the hybrid design of keywords based approach and machine learning algorithm using rule based engine to detect whether the emotion is present in an input dataset. Learning based Support Vector Machine and Naive Bayes are considered as the keyword classifiers to detect the Ekman's six types of basic emotions (happy, fear, anger, sadness, disgust and surprise). There are three phases of emotion detection as pre-processing (segmentation, tokenization, stemming and stop words removal), feature extraction (scoring of pre-processed dataset) and emotion detection & classification. They used standard Punjabi language dataset of "HC Corpora".

Jain and Sandhu [8] also explored emotion detection from Punjabi text data using Support Vector Machine and Maximum Entropy Algorithm. They used the parameters of precision, recall and f-measure but could not come up with any enhancement in accuracy.

Arora and Kaur [9] developed an application for Punjabi sentiments to be categorized into positive and negative. They used ASP.Net language with SQL Server 2008 to develop an offline application which also keeps a directory of words with positive and negative score.

Deepali and Garg [10] used N-gram and Naïve Bayes technique mining the users opinion about the movies reviews. They could not find it easy to deal with complex sentences where the first part of sentence is positive but the second part is negative. Chopra and Bhatia [11] implemented dictionary based approach and divided the opinions in three sentiments i.e. positive, negative and neutral.

III. COMAPARATIVE STUDY

Sentiwordnet is an opinion lexicon of any language derived from wordnet database where each term is linked with polarity of the word. Sentiwordnet is available for total 57 languages in the world except Punjabi language [12]. Using two different domain corpora news and blogs, Sentiwordnet for Bengali language was developed with a good coverage of 33805 entries [13].

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), Nov 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Author	Language	Techniques Used	Dataset	Features Extracted
Jain U., Sandhu A. [8]	Punjabi	*Supervised:* SVM, Maximum Entropy Algorithm	Punjabi websites, newspapers and Punjabi blogs	*Linguistics:* Joy, Sadness, Fear, Surprise, Disgust and Anger
Sharma A. [3]	Punjabi	*Supervised:* Naïve Bayes Classification	Punjabi websites, newspapers and Punjabi blogs	*Linguistics:* Positive/ Negative Polarity
Kaur A., Gupta V. [1]	Punjabi	Lexicon based, N-grams modelling, ML	Manually developed a seed list of Punjabi words	*Linguistics:* Positive/ Negative Polarity
Kaur A., Gupta V. [5]	Punjabi	*Supervised:* SVM	Punjabi websites, newspapers and Punjabi blogs	*Linguistics:* Positive/ Negative Polarity
Grover S., Verma A. [7]	Punjabi	*Naïve Bayes, SVM, Rule Based Engine*	Standard Punjabi dataset "HC Corpora"	*Linguistics:* Emotion Detection.

Table 1 Sentiment Analysis on Punjabi language text

IV. DISCUSSION

Cambria E., Poria S. [14] states that there are 15 Natural Language Processing problems that need to be solved to achieve human like performance in sentiment analysis. Such problems are organized into three layers: syntactics, symantics and pragmatics as shown in Fig.3. Sentiment analysis is not only about polarity detection.

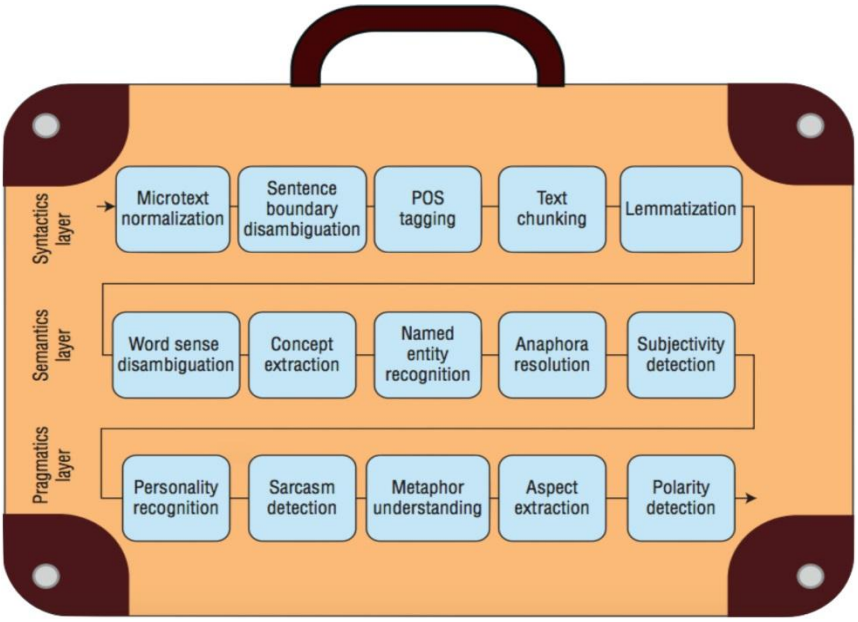

Figure 3 Sentiment Analysis's Big Suitcase of NLP Problems

There is an urgent need to develop multilingual automatic sentiment analysis systems due to following reasons so that the researchers from different countries may build sentiment analysis systems in their own local languages also. As most of the research has been done in English, and too less resources tools in Indian languages are available which can be used to build good sentiment classifiers in these languages. Secondly, in many applications, companies too wish to identify consumer opinions about their products and services in different countries which is only possible through sentiment analysis systems in other languages.

V. CONCLUSION

Sentiment classification task is very challenging in computational linguistic point of view, especially if text is written in other than English language. The availability of linguistic resources for Punjabi language is very scarce such as automatic tools for tokenization, feature selection and stemming etc. With a morphological and grammatical difference of Indian languages from English also place a number of challenges on the way of using current algorithms with accuracy. Enormous amount of research work has to be done in future for sentiment analysis of opinions in Punjabi language and foremost a wordnet of Punjabi needs to be formulated. Problems like Word Sense Disambiguation, word order, co-reference of words, morphological variations etc. need to be worked out.

REFERENCES

[1] Kaur, Amandeep, Gupta, Vishal, *Proposed Algorithm of Sentiment Analysis for Punjabi Text*, Journal of Emerging Technologies in Web Intelligence, *6* (2), 2014, 180-183.

[2] Arora, Piyush, *Sentiment Analysis for Hindi Language*, Masters Thesis, IIT, Hyderabad, 2013.

[3] Sharma, Anu, *Sentiment Analyzer using Punjabi Language*, Int. Journal of Innovative Research in Computer and Communication Engineering, *2* (9), 2014, 5904-5909.

[4] Sharma, Anu, *A Study and analysis of Opinion Mining Research in Indo-Aryan, Dravidian and Tibeto-Burman Language Families*, Int. Journal of Data Mining and Emerging Technologies, *4* (@), 2014, 53-60.

[5] Kaur, Amandeep, Gupta, Vishal, *A Novel Approach for Sentiment Analysis of Punjabi Text using SVM*, The International Arab Journal of Information Technology, *14* (5), 2017, 707-712.

[6] Kaur, Amandeep, Gupta, Vishal, *N-gram Based Approach for Opinion Mining of Punjabi Text*, Proceedings of International Workshp on Multi-disciplinary Trends in AI © Springer, 2014, 81-88.

[7] Grover, Sheeba, Verma, Dr. Amandeep, *Design for Emotion Detection of Punjabi Text using Hybrid Approach*, IEEE International Conference on Inventive Computation Technologies, 2016.

[8] Jain, Er. Ubeeka, Sandhu, Amandeep, *Emotion Detection from Punjabi Text using Hybrid Support Vector Machine and Maximum Entropy Algorithm*, 4(11), 2015.

[9] Arora, P., Kaur, B., *Sentiment Analysis of Political Review in Punjabi Language*, Int. Journal of Computer Applications, 126(14), 2015, 20-23.

[9] Deepali, Garg, N., *Movie Review Mining in Punjabi*, Int. Journal of Application or Innovation in Eng. And Management, 2(12), 2013, 372-375.

[10] Chopra, F., Bhatis, R., *Sentiment Analyzing by Dictionary Based Approach*, International Journal of Computer Applications, 152(5), 2016, 32-34.

[12] Das A, Bandyopadhyay S., *Sentiwordnet for Bangla*, Knowledge Sharing Event-4, Task, 2010

[13] Kaur, Amandeep, Gupta, Vishal, *A Survey on Sentiment Analysis and Opinion Mining Techniques*, Journal of Emerging Technologies in Web Intelligence, *5* (4), 2013.

[14] Cambria, Eric, Poria S., *Sentiment Analysis is a Big Suitcase*, IEEE Affective Computing and Sentiment Analysis, 2017.

[15] Ganeshbhai S.Y., Shah, B.K. *Feature Based Opinion Mining A Survey*, IEEE International Advance Computing Conference, 2015, 919-923.