# Aspect Based Opinion Mining: In Context of Hindi Roman Language Using NLP

**Mariya Zafar[1] Masood Ahmad[2] Dr. Shafeeq Ahmad[3]**
[3]Director
[1,2,3]Department of Computer Science & Engineering
[1,2,3]Dr. A.P.J. Abdul Kalam Technical University (Lko) U.P

*Abstract*— The internet revolution has brought about a new way of expressing an opinion. It has become a medium through which people openly express their views on various subjects. These opinions contain useful information which can be utilised in many sectors which require constant customer feedback. Analysis of the opinion and it's classification into different classes is gradually produced as a key factor in decision-making These strategies usually attempt to extract the overall sentiment revealed in a sentence or document, either positive or negative, or somewhere in between. However, a downside of those strategies is that the information is often degraded, particularly in texts wherever a loss of information may occur due to mismatch of grammar in global language. In this paper, we tackle such situation Firstly with the implementation of Autocorrect feature, Secondly by acknowledging those Aspects based words which user writes in their home language. Then we apply our limited datasets to regain such words in its standard form. This method helps us to make our sentiment analysis more efficient.

***Key words:*** Natural Language Processing (NLP), Context-Free Grammar, Opinion Mining, Auto-Correction, Feature Extraction, POS Tagging, Roman Hindi Language

## I. INTRODUCTION

Opinion mining is an evolving space of information retrieval. This area is a combination of data mining and natural language processing techniques. Client opinion mining is a method to search out user's insight regarding different product [1]. Natural language processing (NLP) will give us labels, the correlation of labels, and therefore the ontological knowledge about them, so we are able to automate the acquisition of contextual knowledge. In this paper, we tend to work on auto correction of the source comments. There are 2 major approaches: first, different language databases organize words according to numerous semantic concepts. Using these, we are able to build special purpose databases that can predict the labels involved given a certain context. Here we tend to build a knowledge base for the purpose of describing common daily activities; Second, statistical language tools will give the correlations of various labels. We tend to show a way to find out a language model from massive corpus data that exploits these correlations and propose a general improvement scheme to integrate the language model into the system. Experiments conducted on 3 multi-label everyday recognition tasks support the effectiveness and efficiency of our approach, with significant gains in recognition accuracies when correlation information is used in this paper, we mainly focus on how to improve the performance of comment based analyser by providing a touch of Preprocessing of an NLP technique. Preprocessing the input text is an essential component in a Natural Language Processing (NLP) system. We are taking the preprocessors in the context of building a database which gives us clear output to those comments which are difficult to read because of its Indian language issue. Whenever we come across for translation we encounter with the unique formats in an input text and getting its appropriate translation is difficult. Sometimes they may not have proper grammatical structure and may not be able to handle using a language rule because of its different platform. This paper presents a strategy to identify the special formats in English text like 'acha', 'bkvas' 'bekar', 'shi h', which generally called Roman Hindi. This Roman Hindi is a pattern based system with context-free grammar like structure for English which generates a contrived target for a group of Indian languages. The preprocessor is one of the main modules in this translation System it manipulates the English input text to fabricate an input which is more competent for an engine to generate the felicitous translation. Extensive research is carried out in this area to disambiguate and process the input.

### *A. Motivation*

Most of the research works in Opinion Mining are carried out in English language and almost people comment on movie or product in short word and Hindi language. In this process, the common Hindi word does not consider.

## II. LITERATURE REVIEW

In 2004 Liu and Hu made one of the early attempts on mining and summarising customer reviews. They proposed a system an opinion summarization system which uses NLP (Natural Language Processing) linguistics parser [1]. That parser tags the parts of speech for each word in the sentence. They used association miner algorithm for mining frequent features on noun/noun phrases. For the classification of positive and negative opinions, adjectives words were analysed and WordNet [2] dictionary was used for the orientation of words. They also used the same system and attempt for the mining the opinions which carry some specific features of the interested product [3].

Another system with the name OPINE was proposed by Popescu et al [4]. OPINE is unsupervised information extraction systems which is capable of identifying the features of the products and the opinions related to that, determine the polarity and then rank the opinions according to strengths. Wang et al proposed a web-based product review and customer opinion summarization system with the name of SumView [5]. SumView is using the Feature-based weighted Non-Negative Matrix Factorization (FNMF) algorithm for classifying sentence into relevant features cluster. It provides a summary of

All rights reserved by www.ijsrd.com 

information contained review documents by selecting the most representative reviews sentences for each extracted product feature.

All the system discussed above perform opinion mining and a sort of sentiment analysis. However, their work revolves around the English language. De-cheng and Tian-fang proposed a framework for topic identification and features extraction from the opinions and reviewed posted in the Chinese language [6]. El-Halees proposed systems which extract the user opinion from the Arabic text [7]. His proposed system uses three methods i.e. lexicon-based method, machine learning method and k-nearest model for better performance. Abbasi et al worked on the sentiment analysis of the opinions posted in English or Arabic. They used Entropy Weighted Genetic Algorithm ( EWGA) for the assessment of key features [8]. Almas and Ahmad used computational linguistics for English, Arabic and Urdu. However, their work is purely based on the financial news data and describe a method of mining some special terms from the data which they called it local grammar [9]. Syed et al proposed an approach of sentiment analysis using adjective phrase [10]. Javed and Afzal proposed lexicon based bi-lingual sentiment analyses system that analyse the tweet of users written in two languages i.e. English and Roman-Urdu. They used two type of lexicon for sentiment analysis. For the English language, they used SentiStrength lexicon [11] and for Roman-Urdu, they manually developed a lexicon, as there is no other such lexicon is available for Roman-Urdu [12].

## III. PROPOSED METHOD

This paper is concerned with the correction of those comments which is unreadable by the software. It creates a big ambiguous data which causes loss of information approx 30 percent of an important data set. For example ‘zada nahi’ , ‘accha', ‘bekar movie h’, etc types of words which are not grammatically correct and not in sentences. To overcome this issue we come with an idea of the auto-correction tool. We take the data set for analysis, which includes auto correction, determining word focus towards the negative side or positive or in between them. It required preprocessing with the help of NLP-Natural Language processing.

Comment analysing structure needs to get its 100% accuracy, but due to lack of understanding of language gives the strong reason to get over from those comments. By correcting it gives us more useful data to evaluate the result. To know the correct result is a most promising task because It is used in decision making, based on the feedback. In the proposed method, we have focused more on how to improve the words extraction from the given reviews or opinions.

This section proposes the methodology and framework used for classification of comments. Diagrammatically it is shown in figure 1.The steps involved are Normalization, standard Feature extraction, additional feature extraction, feature selection and finally classification.

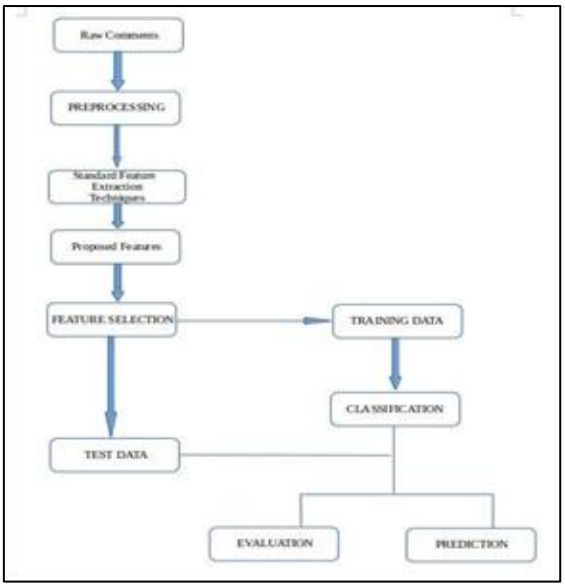


Fig. 1: Proposed Method Framework

Main approaches used in Aspect based Sentiment Analysis are:

1) Auto correction of misspelled words.
2) Common and short Hindi word - It is a collection of words (adjective and adverb) or phrases with a value is assigned +1, -1 and 0 (positive, negative or neutral).
3) Using Natural Language Process Machine Learning - performs the NLP Post tagging of sentences and finding the adverb and adjective.

### *A. ALGORITHM 1*

Finding Adjective and Adverb using NL
Data: Movie Review as M
Result: Extracted adjectives( and most frequent in M
begin
Initialize:
Load nlp model jar file
Adjective_and_adverb list = [];
Perform:
parts-of-speeh tagging(pass sentence to the nlp model) ;
Collection of all extracted adjectives, J from M;
for each j in J do(J is dataset of adjective and adverb)
if(j value exist in J)
find its value
update the total value
end
return the total value
end

### *B. ALGORITHM 2: Negation Phrases Identification*

Words such as adjectives and verbs are able to convey opposite sentiment with the help of negative prefixes. For instance, consider the following sentence that was found in an electronic device’s review: “The built in speaker also has its uses but so far nothing revolutionary." The word, “revolutionary" is a positive word .

1) Data: Set of positive and negative seed words P and N
2) Result: Classification into positive, negative and neutral
3) Require: Tagged Sentences, Negative Prefixes
4) Ensure: NOA Phrases, NOV Phrases
5) for every Tagged Sentences do
6) for i/i + 1 as everyword/tag pair do
7) if i + 1 is a Negative Prefix then

All rights reserved by www.ijsrd.com 

8) if there is an adjective tag or a verb tag in next pair then
9) NOA Phrases←(i, i + 2)
10) NOV Phrases←(i, i + 2)
11) else
12) if there is an adjective tag or a verb tag in the pair after next then
13) NOA Phrases←(i, i + 2, i + 4)
14) NOV Phrases←(i, i + 2, i + 4)
15) end if
16) end if
17) end if
18) end for
19) end for
20) return NOA Phrases, NOV Phrases

*C. AUTO CORRECTION of unidentified roman Hindi language*

Apostrophes have different roles in different aspects. Sometimes we use the apostrophe with contractions like I'm, How're, He's, She'll, etc. These entries will have an expansion. Examples are illustrated below.

a) Input Text : I'm your friend.
Preprocessor Output : I am your friend.
b) Input Text :achha ha.
Preprocessor Output : good.
c) Input Text :bekar movi h
Preprocessor Output :bad movie

The Dataset we have used contains a list of comments and respective labels. These should be converted into a feature vector which is used by our machine-learning algorithms. For this, we use different Natural language processing techniques to obtain an accurate representation of the comments in feature form. We use various techniques based on our observations.

*1) Removing unwanted strings*

For the comments to be used by machine-learning algorithms they should be in standard form. Raw comments present in the dataset which contains many unwanted strings like '\xa0','\\n' and many such encoding parts should be removed. Hence the first step is to preprocess the comments by removing unwanted strings, hyphens and punctuations. The following figure demonstrates an example of this step.

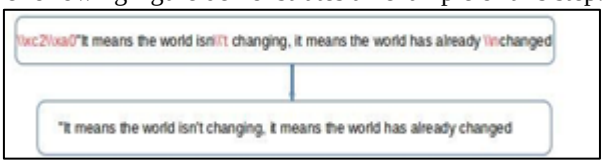


Fig. 2: Removing unwanted strings

*2) Correcting words*

One of the reasons comments are classified as insulting is the presence of profane or abusive words. The total number of bad words present in comments is taken as one of the features. A dictionary of 500 bad words [16] is compiled, which also includes variations of words (@$$,s h I t). This dictionary is used because people using the online forums sometimes use special characters to build an insulting word (!d!ot,@$$ole). When we encounter such words, the dictionary helps to convert them into a natural form. Also, Stemming is applied to capture bad word variations that are not contained in the dictionary. Stemming reduces a word to its core root, for example embarrassing is reduced to embarrass. Here it is noted that stemming is only applied to bad word dictionary, not on the dataset used, as it will lead to information loss. Again a small dictionary and a spell checker are used to convert all variations of "you", "you're" (e.g u, ur etc) which are present in the dataset as participant use them as part of a flexible language. The following figure demonstrates an example of this step.

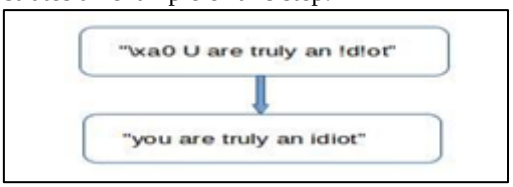


Fig. 3: Correcting words

*3) Parts of Speech Tagging*

As the identification of the opinion polarity relies on the adjective used in the sentence, it's important here to identify the adjectives within the sentences. part of Speech Tagging is employed for that identification purpose. SharpNLP [13] is used for POS tagging. It's a C sharp port of Java OpenNLP tools. OpenNLP is a library of machine learning based toolkit [14]. it has the inbuilt facilities of text processing like tokenization, segmentation of sentences, PoS tagging, parsing chunking and plenty of others. It's a very powerful tool even it can be used to extract triplets from the text [15]. in this case, the SharpNLP perform all the pre-processing of data like tokenization, sentence segmentation, and part-of-speech tagging.

## IV. CONCLUSION

This paper explored the applicability of feature selection methods for sentiment analysis and investigates their performance on roman Hindi language in term of recall, precision and accuracy. Roman Hindi aspect-based sentiment analysis is investigated on movie reviews corpus with a size of 2520 comments. The experimental results show that information gain gives a stable performance for a different number of features. During this study, Gain ratio gave the most effective results for a large range of sentimental features selection more than 5000 features. Moreover, we found that performance of the sentiment Analysis from comments improved because it adds those unreadable comments with the proper meaning that is the root cause of lacking accuracy. We fixed this issue by converting those Roman Hindi aspect based sentiments to the proper English words.

## V. FUTURE WORK

In India, there are 22 official languages and 13 languages have over 10 million speakers. With multiple sources of information available for each language, it's straightforward to collect data and analyse them. In context to Indian Languages, earlier work done for sentiment analysis has been on based on language conversion only. The nature of Indian languages varies a great deal in terms of the script, representation level and linguistic characteristic etc. So, there is a large amount of work that needs to be done to understand the behavior of Indian languages and perform the analysis of same accordingly. We have manually prepared dataset dictionary available for Hindi roman words. In future, we can try and come up with the more focused approach and other big data set for other Indian languages than Hindi.

All rights reserved by *www.ijsrd.com* 

As discussed above there is a lack of annotated datasets and resources for Indian languages, so it needs considerable focus and time to be given. Once we have sufficient data to experiment with, various machine learning techniques can be easily used and applied to learn from the text more effectively.

## REFERENCES

[1] Hu, M. and B. Liu. Mining and summarizing customer reviews. in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004: ACM.
[2] Miller, G.A., WordNet: a lexical database for English. Communications of the ACM, 1995. 38(11): p. 39-41.
[3] Hu, M. and B. Liu. Mining opinion features in customer reviews. in AAAI. 2004.
[4] Popescu, A.-M., B. Nguyen, and O. Etzioni. OPINE: Extracting product features and opinions from reviews. in Proceedings of HLT/EMNLP on interactive demonstrations. 2005: Association for Computational Linguistics.
[5] Wang, D., S. Zhu, and T. Li, SumView: A Web-based engine for summarizing product reviews and customer opinions. Expert Systems with Applications, 2013. 40(1): p. 27-33.
[6] LOU, D.-c. and T.-f. YAO, Semantic polarity analysis and opinion mining on Chinese review sentences [J]. Journal of Computer Applications, 2006. 11: p. 30-45.
[7] El-Halees, A. Arabic opinion mining using combined classification approach. in the proceeding of: 2011 International Arab Conference on Information Technology ACIT.
[8] Abbasi, A., H. Chen, and A. Salem, Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems (TOIS), 2008. 26(3): p. 12.
[9] Almas, Y. and K. Ahmad. A note on extracting sentiments in financial news in English, Arabic & Urdu. in The Second Workshop on Computation, al Approaches to Arabic Script-based Languages. 2007.
[10] Syed, Z., M. Aslam, and A. Martinez-Enriquez, Adjectival Phrases as the Sentiment Carriers in the Urdu Text. Journal of American Science. 7(3): p. 644-652.
[11] Thelwall, M., et al., Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 2010. 61(12): p. 2544-2558.
[12] Javed, I. and H. Afzal. Opinion analysis of Bi-lingual Event Data from Social Networks. in ESSEM@ AI* IA. 2013: Citeseer.
[13] Sharp NLP. [cited 2014 11 04, 2014]; Available from: http://sharpnlp.codeplex.com.
[14] Baldridge, J., The opennlp project. URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012), 2005.
[15] Rusu, D., et al. Triplet extraction from sentences. in Proceedings of the 10th International Multiconference" Information Society-IS. 2007.
[16] B. Zhou, Y. Xiong, and W. Liu, "Efficient web page main text extraction towards online news analysis," in e-Business Engineering, 2009.ICEBE'09. IEEE International Conference on. IEEE, 2009, pp. 37–41..

All rights reserved by *www.ijsrd.com* 