A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study

Aditya Joshi[1] **Balamurali A R**[2] **Pushpak Bhattacharyya**[1]
[1]Dept. of Computer and Science Engineering, IIT Bombay
[2] IITB-Monash Research Academy, IIT Bombay
Mumbai, India - 400076
{adityaj,balamurali,pb}@cse.iitb.ac.in

Abstract

Sentiment Analysis (SA) research has gained tremendous momentum in recent times. However, there has been little work in this area for an Indian language. We propose in this paper a fall-back strategy to do sentiment analysis for Hindi documents, a problem on which, to the best of our knowledge, no work has been done until now. (A) First of all, we study three approaches to perform SA in Hindi. We have developed a sentiment annotated corpora in the Hindi movie review domain. The first of our approaches involves training a classifier on this annotated Hindi corpus and using it to classify a new Hindi document. (B) In the second approach, we translate the given document into English and use a classifier trained on standard English movie reviews to classify the document. (C) In the third approach, we develop a lexical resource called Hindi-SentiWordNet (H-SWN) and implement a majority score based strategy to classify the given document.

A comparison of performance of these approaches implies that we can adopt a fall-back strategy for doing sentiment analysis for a new language, viz., (1) Train a sentiment classifier on in-language labeled corpus and use this classifier to classify a new document. (2) If in-language training data is not available, apply rough machine translation to translate the new document into a resource-rich language like English and detect the polarity of the translated document using a classifier for English, assuming polarity is not lost in translation. (3) If the translation cannot be done, put in place a SentiWordNet-like resource for the new language and apply a majority strategy to the document to be classified. Two additional contributions of our work are (i) the development of sentiment labeled corpus for Hindi movie reviews and (ii) construction of a lexical resource, Hindi SentiWordNet based on its English counterpart.

1 Introduction

Sentiment Analysis is a natural language processing task that deals with finding orientation of opinion in a piece of text with respect to a topic (Pang and Lee, 2008). The increasing user-generated content on the Internet is the driving force behind the sentiment analysis research. The advances in sentiment analysis have opened up an important channel using which businesses can gauge the online trend about their products and services. Majority of the existing work in this field is in English (Pang and Lee, 2008). Our work is a foray into sentiment analysis for Hindi.

Hindi is an Indian language which enjoys a speaker population of 43 crore. (2001 census of India[1]; rounded to the most significant digit) With Unicode standards in place for Hindi, the usage of Hindi is growing. Today, several websites provide support for Hindi and other Indian languages. In fact, a new word 'चिट्ठा ' (literally translated as a 'bundle of chits/posts') has been coined for a '*blog*'. To the best of our knowledge, no work has been reported for sentiment analysis in Hindi. Das and Bandyopadhyay(2009) also mention the same dearth of research in sentiment analysis for Indian languages. Being the first to attempt sentiment analysis in Hindi, we address the following questions in this work of ours:

[1] $http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement1.htm$

Proceedings of ICON 2010: 8th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2010

- *How well will a classifier modeled on a Hindi corpus perform?*
- *Can machine translation be used for SA in Hindi if there does not exist a system for SA?*
- *How can a resource be created for the purpose of sentiment analysis in Hindi using the existing resources available for English and Hindi?*
- *How effective is a lexical resource for SA in Hindi?*
- *How do these strategies fare against each other?*

Our technique for in-language sentiment analysis addresses the first question. For the second question, we give an approach to perform machine translation-based sentiment analysis. To address the third and the foruth questions, we describe an approach for resource-based sentiment analysis where we create a lexical resource based on SentiWordNet 1.1 (Esuli and Sebastiani, 2006). The last question is addressed by comparing the performance of the approaches. The result is a fall-back strategy that forms the crux of the work.

The contributions of this work are:

- *Creation of a manually annotated corpus for Hindi.*
- *Creation of Hindi SentiWordNet (H-SWN), based on the equivalent for English.*
- *Learning the best parameters for different approaches and comparison of their performance for sentiment analysis in Hindi. The results are comparable to the ones in multilingual sentiment analysis studies involving European languages (Banea et al., 2008b; Denecke, 2008).*

The rest of the paper is organized as follows. Section 2 describes the work related to ours. The resource developed for sentiment analysis in Hindi is described in Section 3. The approaches and experiment setup are discussed in Section 4 and 5 respectively. The results and discussions are described in Section 6. Finally, Section 7 presents the conclusion and the future work.

2 Related Work

Much of the published works are for English. Pang and Lee (2004) use a supervised approach to classify movie reviews into two classes after performing subjective feature extraction. They achieve an in-language classification accuracy of 86% using the unigram model. In contrast, Dave *et al.* (2003) find that bigram-based features produce better results. Both these systems suggest that the *Term Presence* vector expresses sentiment more precisely than the *Term Frequency*. Whitelaw *et al.* (2005) construct a lexicon that provides appraisal attributes for terms and use them as features for classification. These features along with the bag-of-words model gives 90.2% accuracy. Preprocessing steps like stemming and lemmatization have been found to be detrimental to classification accuracy by Leopold *et al.* (2002).

Chaumartin *et al.* (2007) use SentiWordNet (Esuli and Sebastiani, 2006) for finding the polarity of the newspaper headings. Verma and Bhattacharyya *et al.*(2009) use the same resource for developing both resource-based and machine learning-based methods for classification of movie reviews. Denecke (2008) also experiment with SentiWordNet to study whether such a resource can aid in sentiment classification. In general, these studies observe that resource-based sentiment classifiers are less effective compared to machine learning-based approaches (Verma and Bhattacharyya, 2009; Denecke, 2008). Nevertheless, we believe that lexical resources created for such resource-based systems can additionally be utilized for developing decisive feature vectors. This work of ours involves creation of one such resource for Hindi.

Under the domain of multilingual sentiment analysis, work has been carried out for several European languages. Banea *et al.* (2008b) study the effectiveness of machine translation (MT) for construction of resources and tools in the target language. Their experiments give variants of generating an annotated corpus in Romanian/Spanish. Denecke *et al.*(2008) explore the same for German where they translate a movie review corpus to English and use SentiWordNet to generate classifiers that predict polarity of the documents.

Mihalcea *et al.*(2007) explore the technique of rapidly developing resources in a target language. (Romanian in this case) The study suggests that machine learning-based approaches are better than resource-based approaches. Banea *et al* (2008a) creates similar lexicon for sentiment analysis in Ro-

manian using a bootstrap method.

Ahmad *et al.*(2006) build a local grammar consisting of collocations for a language and annotate each of these collocations with sentiment. The work is based on Chinese and Arabic. The authors report that the accuracy of the extraction is in the range of 60-75 % (Ahmad et al., 2006).

Das and Bandyopadhyay (2009) report the only known work of sentiment analysis in an Indian language, specifically Bengali. The authors use SentiWordNet as well as Subjectivity lexicon for generating a lexical resource that contains sentiment values along with their POS tags mapped directly from the English words. The generated resource contains 35805 Bengali entries. Using the lexicon and features like positional aspect, a supervised classifier is generated. This classifier achieves a precision of 74.6% and a recall of 80.4%.

3 Hindi-SentiWordNet (H-SWN): Lexical Resource for Hindi

A naive approach to predict the sentiment of a document is to use the prior polarity of terms present in it. In order to find the polarity, a lexical resource is required. In this section, we explain the methodology used to develop such a resource for Hindi. We call this resource Hindi-SentiWordNet (H-SWN). H-SWN is created by exploiting two lexical resources namely SentiWordNet (SWN) and English-Hindi WordNet linking (Karthikeyan, 2010).

3.1 SentiWordNet (SWN)

SentiWordNet 1.1[2] is an automatically generated WordNet-based lexical resource with polarity scores attached to the *senses* instead of the *terms*. Each synset is annotated with three scores -positive, negative and objective score. The sum of the scores adds to 1. For example, one of the senses of the word "*influence*" appears as: "*n 148171 0.125 0.625 influence#n#2*" The scores 0.125 and 0.625 are the positive and negative scores respectively. The third score is the objective score which is the remainder probability. (i.e. 1 - 0.125 - 0.625 = 0.25)

3.2 WordNet linking

The WordNet linking[3] used for the purpose provides a mapping between synsets of WordNets of different languages. An example entry in this linking is: *Hindi ID: 10001 (Noun), English ID: 532338 (Noun)*

3.3 Development of H-SWN

The underlying assumption for development of H-SWN is that the sentiment of a synset is retained across languages. Hence, since the sentiment score associated with a synset is available for English in SWN, the sentiment can be projected to the corresponding synset in Hindi.

The algorithm used for creating the H-SWN is as follows:

1. For each synset in the SWN, repeat 2 to 3:
2. Find the corresponding synset in Hindi WordNet
3. Project the scores of a synset in SWN to the corresponding synset in Hindi WordNet

The result is H-SWN with sentiment-related scores associated with synsets. The distribution statistics for H-SWN with respect to POS tags is given in Table 1. The WordNet linking is still under construction and as it expands, the number of synsets will increase. The projection method described here can also be utilized for porting polarity scores for all the languages present in the WordNet linking.

Parts of Speech (POS)	Number of synsets
Adjective	3108
Adverb	313
Noun	8861
Verb	3971
Total	16253

Table 1: Different synset distribution based on the POS

4 Our Approaches

In this section, we describe three approaches for sentiment analysis of the Hindi documents. The final prediction is in the form of polarity labels namely, positive or negative.

4.1 In-language Sentiment Analysis

The approach relies on the availability of resources needed to analyze the sentiment content in Hindi. The approach tells us how well SA would work if the training of the classifier has been done in the same language as the text corpus. Thus, the training and the test documents for this experiment are in Hindi.

[2]The resource is available for download on request to the authors, Please visit the site sentiwordnet.isti.cnr.it

[3]The resource is available on request from http://www.cfilt.iitb.ac.in/

4.2 Machine Translation (MT) - based Sentiment Analysis

The dearth of annotated corpus in Hindi necessitates the study of a machine translation approach to sentiment analysis. In this approach, a translation module is used to translate the documents in Hindi to English. Figure 1 represents the step followed in this process. The assumption here is that **the sentiment of a document is preserved in translation**. One realizes that this assumption is not very unrealistic and forms the basis of many cross-lingual and multi-lingual studies in the past (Banea et al., 2008b; Denecke, 2008).

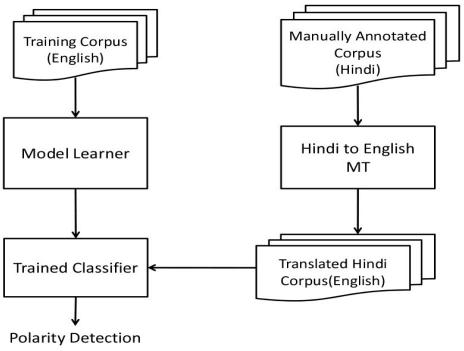

Figure 1: Procedure for MT-based Sentiment Analysis

The classifier is trained on English documents. The Hindi document whose polarity needs to be determined is passed through a translation system to translate it to English. The translated document is then given to the classifier. We assume that poor structure of sentences in the translated document does not hurt as long as the sentiment-bearing words are accurately translated.

4.3 Resource-based sentiment analysis

From previously published works, it is evident that machine learning-based approaches are better suited for sentiment prediction compared to resource-based approaches (Verma and Bhattacharyya, 2009). However, all these approaches need a large amount of training data. A good resource-based classifier can act as a substitute for this large amount of data. In this approach, we aim at evaluating a majority-based sentiment classifier based on H-SWN.

The algorithm for resource-based sentiment analysis is as follows:

- *For each word in the document,*
  - *Apply stop word removal and stemming (depending on the variant of the experiment)*
  - *Look up the sentiment scores for each word in the H-SWN*
  - *Assign a polarity to a word based on the maximum of the scores*
- *Assign to a document the polarity which majority of its words possess*

The experiment is conducted in variants like with/without stop word removal and with/without stemming. In addition, there are two versions of the experiment - one where the scores corresponding to all senses of a word are summed, another in which only the most common sense is considered for polarity determination. For example, for a word with five senses - three positive and two negative, the overall contribution of the word to the sentiment of the sentence is of three positive and two negative words in the first version. In the second version, the word will be assigned the polarity of its most common sense.

The overall polarity of the document is the majority of the polarity of its words. It follows that the problem of thwarted expressions (Pang and Lee, 2008) will derail the approach.

5 Experiment Setup

In this section, we describe the corpus developed for the experiments along with tools and parameters selected.

5.1 Corpus

The corpus of Hindi documents created for the task consists of reviews from blogs reached through blog directories[4]. The corpus consists of 250 Hindi movie reviews manually annotated with polarity tags resulting in equal amount of positive and negative data. We standardize our corpus to UTF-8 format.

Apart from this corpus, we use the standard movie review corpus in English by Pang and Lee(2004) The corpus consists of thousand review documents each of positive and negative polarity downloaded from www.rottentomatoes.com

5.2 In-language Sentiment Analysis

We use RapidMiner 5.0 for the classification of documents (Mierswa et al., 2006). The learner used for classification is LibSVM[5] with SVM type as C-

[4] www.hindiblogs.org,dir.hinkhoj.com,hindigear.com
[5] http://www.csie.ntu.edu.tw/ cjlin/libsvm

SVC. A linear kernel is used with the vanilla settings for other parameters[6].We create two types of models as a part of the experiment: unigram and bigram based models. We also vary the feature representation (*Term Frequency, TF-IDF, Term Presence*) to see the effect on in-language classification of Hindi review documents. All the results reported for in-language sentiment analysis are based on five-fold cross validation. The best classification accuracy among the various runs is reported.

5.3 MT-based Sentiment Analysis

For MT-based sentiment analysis, we use Google translate[7]to translate the corpus into English. A classifier is then modeled based on the standard movie review corpus. The model is then used to classify the translated documents. The setting of the classifier is same as in-language sentiment analysis. The result reported is only for unigram based *TF-IDF* representation.

5.4 Resource-based Sentiment Analysis

For resource-based sentiment analysis, a classifier is implemented to use scores in H-SWN. The parameters of the experiment are usage of a stemmer, stop word removal and multiple sense consideration as explained above. We use the stemmer and stop word list mentioned in (Harshada Gune and Bhattacharyya, 2010).

6 Results and Discussion

In this section, we discuss the results obtained from the various runs of the experiments conducted.

6.1 In-Language Sentiment Analysis

Table 2 presents the values of classification accuracy for the experiment.

Table 2 shows that the accuracies for *TF-IDF* representation are the highest as compared to the other representations in each set of the experiment. The TP representation, in all cases, performs the worst. Like observations made by earlier work in SA, stemming is detrimental to the accuracy of SA. However, we also make an interesting observation. For a corresponding setting in English, the best reported readings so far are for the term presence representation of feature vector where the features consist of unstemmed bigrams (Pang and Lee, 2004; Dave et al., 2003). On the contrary, for Hindi, Table 2 shows that the highest accuracy for in-language sentiment analysis is for unigram non-stemmed features with TF-IDF representation.

Experiment Setup	Representation	Accuracy
Unigram + Stemmed	TF	67.83
Unigram + Stemmed	TP	66.23
Unigram + Stemmed	TF-IDF	68.65
Unigram + not stemmed	TF	74.57
Unigram + not stemmed	TP	72.57
Unigram + not stemmed	TF-IDF	**78.14**
Bi-gram + Stemmed	TF	61.2
Bi-gram + Stemmed	TP	57.2
Bi-gram + Stemmed	TF-IDF	62.4
Bi-gram + not stemmed	TF	70.02
Bi-gram + not stemmed	TP	59.88
Bi-gram + not stemmed	TF-IDF	71.72

Table 2: Results for in-language sentiment analysis (in %) for Hindi using different features and representations

6.2 Resource-based Sentiment Analysis

In Table 3, we present the classification accuracy values of resource-based sentiment analysis.

Sense Consideration	Stemming	Stop Word Removal	Accuracy
Most Common Sense	No	No	56.35
	Yes	Yes	53.96
All Senses	No	No	**60.31**
	No	Yes	57.53
	Yes	Yes	55.95

Table 3: Results for resource-based sentiment analysis (in %) for Hindi

Table 3 shows that the best results using H-SWN are obtained when used without stemming and stop word removal and when all the senses of a word are considered to arrive at the polarity prediction. The scores are generally low as compared to the scores for in-language sentiment analysis.

These low scores can be attributed to two possible sources of error. The first source is the absence of sense annotation for words present in the documents. Since the exact senses of the words are not available, the experiments use scores corresponding to all synsets. The second reason is the coverage of the current WordNet linking. The WordNet linking used for generating H-SWN is under construction and hence, not all synsets have been assigned a score.

6.3 Comparison

In Table 4, we summarize the different approaches used for sentiment analysis in Hindi.

It is important to note that the error of the machine translation system affects the performance of MT-based SA. On manual observation, some errors

[6]C=0.0,ϵ=0.0010

[7]http://translate.google.com

Experiment	Accuracy %
In-language sentiment analysis	78.14
MT-based sentiment analysis	65.96
Resource-based sentiment analysis	60.31

Table 4: Comparison of approaches

which may lead to deterioration of sentiment analysis are noticed. Consider the example of a review in the Hindi corpus which is translated as: '*It's infinite, like the last hero Imran not look sturdy*'. Here, the named entity '*Anant*'in the original Hindi corpus was translated to its corresponding English word '*infinite*'.

As expected, in-language sentiment analysis gives the best result for the documents in Hindi. Among the other approaches, the closest that gets to in-language sentiment analysis is MT-based sentiment analysis.

7 Conclusions and Future Work

The work is the first known for sentiment analysis in Hindi and an early one for an Indian language. For the task, we consider three approaches. The first approach is to construct a classifier model for Hindi using a training corpus in Hindi. The second approach is to train a model on annotated English corpus and translate a Hindi document to English in order to use this model. The third approach involves using a majority-based classifier for Hindi SentiWordNet. The results show that the first approach outperforms the others. This implies that the best results can be achieved with an annotated corpus in the same language of analysis. Our results for Hindi sentiment analysis support the published work that in the absence of such a corpus, MT-based systems give superior classification performance as compared to majority-based systems based on lexical resources. The three, in that order, constitute the fall-back strategy we suggest for SA in Hindi.

A possible task for the future with respect to the resource-based sentiment analysis would be to incorporate word sense disambiguation so that a specific sense of word can be looked up in the H-SWN. Another task would be to construct a new version of H-SWN after the linking of Hindi and English WordNet is complete. We expect that the new version will have better coverage.

References

Ahmad, K., D. Cheng, and Y. Almas. 2006. Multi-lingual sentiment analysis of financial news streams. In *Proceedings of ICG*.

Banea, C., R. Mihalcea, and J. Wiebe. 2008a. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. *Proceedings of LREC-08*.

Banea, C., R. Mihalcea, J. Wiebe, and S. Hassan. 2008b. Multilingual subjectivity analysis using machine translation. In *Proceedings of EMNLP-08*, pages 127–135.

Chaumartin, François-Régis. 2007. Upar7: a knowledge-based system for headline sentiment tagging. In *Proceedings of International Workshop on Semantic Evaluations*, pages 422–425.

Das, A. and S. Bandyopadhyay. 2009. Subjectivity Detection in English and Bengali: A CRF-based Approach. *ICON*.

Dave, Kushal, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW-03*, Budapest, Hungary.

Denecke, K. 2008. Using sentiwordnet for multilingual sentiment analysis. In *Proceedings of ICDE-8)*, volume 2.

Esuli, A. and F. Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC-06*, pages 417–422.

Harshada Gune, Mugdha Bapat, Mitesh Khapra and Pushpak Bhattacharyya. 2010. Verbs are where all the action lies: Experiences of shallow parsing of a morphologically rich language. In *Proceedings of COLING 2010)*, Beijing, China, December.

Karthikeyan, Arun. 2010. Hindi english wordnet linkage. Dual degree thesis, CSE Dept. IIT Bombay, May.

Leopold, Edda and Jörg Kindermann. 2002. Text categorization with support vector machines. how to represent texts in input space? *Mach. Learn.*, 46(1-3):423–444.

Mierswa, Ingo, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. 2006. Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of KDD-06*, pages 935–940, August.

Mihalcea, R., C. Banea, and J. Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of ACL-07*, page 976.

Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL-04*, pages 271–278, Barcelona, ES.

Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, Vol. 2(1-2):pp. 1–135.

Verma, Shitanshu and Pushpak Bhattacharyya. 2009. Incorporating semantic knowledge for sentiment analysis. In *Proceedings of ICON-09*, Hyderabad,India, December.

Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of CIKM-05*, pages 625–631.