Lexical approach for sentiment analysis in hindi essay sample

This paper presents a study on sentiment analysis and opinion mining in Hindi on product reviews. We experimented with several methods, mainly focusing on lexical based approaches. Different lexicons were used on same data set to analyse the significance of lexical based approaches.

2. 1 Lexicon

Two different lexicons were used in order to test the efficiency of the lexical based approach for sentiment analysis. Each lexicon contains Adjectives and Adverbs and their corresponding positive and negative scores. HSL lexicon has positive, negative and objective score, where as HSWN lexicon has only positive and negative scores. The scores are the probability values of a word being used in a positive, negative or objective (neutral) sense. For any given word in the lexicon, the sum of all the scores is 1. The total score of a word w is given by, total score(w) = P (p) + P (n) + P (o) (1)

General Terms

Languages, Unsupervised

Keywords

Opinion Mining, Sentiment Analysis

1. INTRODUCTION

In view of the growing content on web in various Indian languages, there is a need for an analysis of the data from various sources like blogs, product reviews and other social networking websites. This classification can be useful in product analysis, marketing strategies, advertisements and other user specific recommendation systems. Sentiment analysis has been done in English and other languages. But it is fairly new in Hindi and other Indian languages. In this paper we propose a method to classify the reviews in to either positive or negative using a lexicon. Two different lexicons, HSL (Hindi Subjective Lexicon)1 [1] and HSWN (Hindi Sentence WordNet)2 were used and each lexicon contains Adjectives, Adverbs and their corresponding scores.

Where, P(p), P(n) and P(o) is the probability of word w being used in a positive, negative and objective (neutral) sense. The size of the lexicons is given in the below table.

Table 1: Size of Lexicons

3. LEXICAL BASED APPROACH

A lexical based approach is followed, in which the data set is tested against two different lexicons[2]. Each review in the data set is classified based on the calculated score for adjective and adverb presence. Two types of approaches were followed using the Lexicon. Both the approaches are tested on two lexicons. • Using Hindi Parts-of-speech (PoS) tagger 3 , where only words that are tagged as JJ or RB are scored based on the lexcicon. • Without PoS tagger, where every word in the review is searched against the adjectives and adverbs in the lexicon and score in computed. There is a chance that the scores for the adjectives and adverbs are biased or domain dependent, so the reviews are ranked on based on the presence (occurrence) of them. For each of the above two approaches, the following four methods are followed. 3 http://ltrc. iiit. ac. in/showfile. php? filename= downloads/shallow_parser. php

2. DATA SET

The data set is product reviews in English, translated to Hindi and is validated manually. The data set contains 700 product reviews, out of which 350 are classified as positive and 350 as negative. The length of each review varies from 2 to 30 words.

HSL (Developed at IIIT, Hyderabad) HSWN (Developed at IIT, Bombay)

• Adjective presence in the lexicon. • Adjective and Adverb presence in the lexicon. • Adjective score in the lexicon. • Adjective and Adverb score in the lexicon. Since the Hindi PoS tagger is not an ideal PoS tagger, the above 4 steps are repeated without applying the PoS tagger on the reviews. Type Adj Adj + Adv Adj + Neg Adj + Adv + Neg

With PoS Tag Presence 56. 01 57. 87 57. 44 59. 45 Score 60. 31 61. 03 60. 88 62. 75

Without PoS Tag Presence 58. 73 57. 86 58. 16 59. 88 Score 69. 05 66. 76 66. 61 68. 05

Table 4: After merging both the lexicons PoS tag approach lead to significant decrease in performance (6 to 9%). The usage of current state-of-the-art Hindi PoS tagger for sentiment analysis is not much of a use as there is no imporvement in performance.

Negation Handling

Negative words tend to change the sense of the entire sentence, so to handle this a method was proposed using PoS tagger. The Hindi PoS tagger tags certain words like (nahi, lekin, paranthu) as ’NEG’, (negatives). A window length of 2 is considered to the left and to the right for every occurrence of a negative word . Then the adjectives and adverbs in the window with positive polarity will be converted to negative and vice-versa. Negation handling is applied for all the above four cases.

4. 2 Analysis on different lexicons

From Table 2 and 3, it can be seen the HSL performs better than HSWN. An analysis was made to study the agreement between the two lexicons. The number of common words in both the lexicons and the polarity shift (a word in one lexicon is tagged as positive and the same word is tagged as negative in another lexicon) for the common words is presented in Table 5.

Merging Lexicons

Since, both the lexicons are developed at different research centres following different approaches, there might be a disagreement for certain words and corresponding scores. So, the lexicons were merged i. e., the mean of the scores were taken for words that are common in both the lexicons. The analysis and results are presented in the next section.

4. 3 Analysis on negation handling

As negation handling in based on the PoS tag ’NEG’, it can seen from the above results that there is a small improvement in performance (2 to 4%). Type HSL HSWL Common Words 2493 156 Total Unique Words 10476 1027 Polarity Shift 1069 60

4. RESULTS AND ANALYSIS

The results for lexical based approach are given in Table 2, 3 and 4. Lexicon Type Adj Adj + Adv Adj + Neg Adj+ Adv + Neg HSL Presence 58. 73 57. 73 56. 30 59. 02 Score 66. 33 64. 60 64. 61 65. 47 HSWN Presence 39. 97 42. 40 39. 39 41. 40 Score 42. 83 44. 84 45. 98 44. 13

Table 5: Comparison of Lexicons It can be observed from the above tables that, the results vary a lot when the lexicon is changed. Approximately 40% (Table 5) of the common words in both the lexicons have different polarity. It can be inferred that lexicons are domain dependent and hence, same lexicon cannot be used for analysing data from different sources.

Table 2: Without PoS tagging

Lexicon Type Adj Adj + Adv Adj + Neg Adj+ Adv + Neg

HSL Presence 55. 07 56. 65 56. 59 58. 70 Score 58. 22 58. 94 59. 16 61. 17

HSWN Presence 39. 68 39. 82 39. 97 39. 68 Score 42. 69 42. 83 42. 83 42. 97

5. CONCLUSIONS

A lexical based approach can be used to get some idea on the sentiments of the reviews. As these techniques show some kind of analysis, they can be extened to other languages once the lexicon is made for them. The use of domain specific lexicon can be analysed by extending the dataset to large reviews as seen in blogs, news.

REFERENCES

[1] P. Arora, A. Bakliwal, and V. Varma. Hindi subjective lexicon generation using wordnet graph traversal. In CICLing, 2012. [2] A. Bakliwal, P. Arora, and V. Varma. Hindi subjective lexicon : A lexical resource for hindi polarity classification. In LREC, 2012.

Analysis on the usage of PoS tagger

It can be observed from Table 2 and 3 that the use of Hindi PoS tagger lead to decrease in performance by 3 to 5% for HSL lexicon and no significant change in performance for HSWN lexcicon. In case of the merged lexicon (Table 4), the