LEXICAL RESOURCE BASED HYBRID APPROACH FOR CROSS DOMAIN SENTIMENT ANALYSIS IN MALAYALAM

Anagha M*[1], Raveena R Kumar*[2], Sreetha K*[2] Rajeev R. R.[3] and P. C. Reghu Raj[4]

[*1] M. Tech, Computational Linguistics, Government Engineering College, Sreekrishnapuram, Palakkad, Kerala, India
Anaghamanoharan3@gmail.com[1]
[*2] M. Tech, Computational Linguistics, Government Engineering College, Sreekrishnapuram, Palakkad, Kerala, India
veenakalathil@gmail.com[2]
[3] M. Tech, Computational Linguistics, Government Engineering College, Sreekrishnapuram, Palakkad, Kerala, India
sreetha227@gmail.com[2]
[3] Research Officer –VRCLC, IIITM-K, Thiruvananthapuram, Kerala, India
rajeev@iiitmk.ac.in [3]
[4] Dept. of Computer Science and Engineering, Government Engineering College, Sreekrishnapuram, Palakkad, Kerala, India
pcreghu@gmail.com [4]

Abstract: In this paper, Cross Domain Sentiment Analysis of Malayalam reviews is done by classifying the polarity of opinions obtained from the user as positive, negative and neutral. Sentiment analysis is an application of Natural Language Processing and text analysis which helps to identify the emotions in a given context. In this work a hybrid approach for Sentiment Analysis is used in which a Hindi Wordnet based lexical resource file is created which is used for finding polarity of review and Machine Learning method is used for tagging certain special cases. Certain other rules are also incorporated to handle special cases. The system yields output that varies in degree between 0 and 1.

INTRODUCTION

Now-a-days people depend more on web and social networking sites for gathering information and opinions on various products and services. Social networking sites have turned communication to interactive dialogues. People share their opinions and various interests and based on the opinions obtained from these sites they take decisions in buying a product or in watching a movie etc. Sentiment analysis extracts the opinions that appear in the web and checks the attitude and judgement of the individual who posted it, about a particular area. The main task in Sentiment Analysis is to gather reviews from different websites and to check the polarity of the reviews. The reviews are classified into *positive, negative, neutral*. Sentiment analysis is a tough task as the sentiments are expressed in natural language.

Natural Language Processing using computers is still a challenging problem. Malayalam language is free ordered and highly agglutinative in nature, which makes extracting the sentiments from a Malayalam sentences much more difficult task (Govindaru et al, 2012). Sentiment analysis can be broadly classified into word level, sentence level, and document level. This work proposes a Hybrid approach for Cross Domain Sentiment Analysis of Malayalam reviews. The lexical resource file created is used to find out the polarity of the subjective words and simultaneously machine learning approach is used for tagging the comments to extract special tags that alters the overall sentiment of the sentence. After alterations the overall score of the sentence is found out, which determines the overall sentiment of the review. The dependency between words in natural language cannot be identified easily using statistical tools. To identify the relationship between words context knowledge is also needed.

The rest of this paper is organized as follows: Section 2 describes various challenges in Sentiment Analysis. Section 3 describes the related works done. Section 4 and 5 presents the proposed methodology, in which the working of rules and the proposed algorithm are included and Section 6 focuses on the experimental results and discussion. Finally, results are summarized and concluded in Section 7. Section 7 also briefs about the future scope of the work and different ways to improve the efficiency of the system.

CHALLENGES IN SENTIMENT ANALYSIS

1. Implicit Sentiment and Sarcasm: Sentences may carry implicit sentiments, which mean the opinion can be expressed without having any sentiment bearing words in it (Subhabrata and Pushpak, 2012). For Example: “Ee book engane vayikkum!” This sentence does not explicitly carry negative sentiment bearing words although it is a negative sentence. Thus identifying semantics is more important in Sentiment Analysis than syntax detection.

2. Thwarted Expectations: Sometimes the author deliberately sets up context only to refuse it at the end (Subhabrata and Pushpak, 2012). Consider the following example: “nalla kadhayanu, nannai abhinayichittund, pakshe ee padam oodilla.” Inspite of the presence of words that are *positive* in orientation the overall sentiment is negative because of the crucial last sentence, whereas in traditional text classification this would have been classified as *positive* as term frequency is more important there than term presence .

3. Subjectivity Detection: This is to differentiate between opinionated and non-opinionated text. This is used to

enhance the performance of the system by including a subjectivity detection module to filter out objective facts. But this is often difficult to do (Subhabrata and Pushpak, 2012). Consider the following examples: "Enikk kathakal ishtamalla", "ishtamaayilla enna katha nannayittund." The first example presents an objective fact whereas the second example depicts the opinion about a particular story named ishtamaayilla.

4. Negation: Handling negation is a challenging task in Sentiment analysis. Negation can be expressed in subtle ways even without the explicit use of any negative word (Subhabrata and Pushpak, 2012). Example: "Ee website upayogikkan eluppam alla". The word "alla" reverses the tag of "eluppam".

RELATED WORKS

Sentiment analysis is one of the most active research area in Natural Language Processing. Many works have been done in English and in other languages using machine learning, semantic orientation methods, rule based methods and fuzzy logic.

In the work done by Pushpak Bahattacharyya et al. (2005) the overall polarity of the document was determined by identifying the sentiment words. Machine learning based approaches were used for sentiment categorization. Chaumartin et al. (2007) uses a rule based and linguistic approaches for detecting the polarity in the news headlines, it uses SentiWordNet for assigning polarity to the sentiment bearing words.

Denecke (2008) uses SentiWordNet for determining the polarity of text within a multilingual framework. Aditya Joshi, Balamurali A R, Pushpak Bhattacharyya, (2010) developed a lexical resource for Hindi, called Hindi-SentiWordNet and implemented a majority score based strategy to classify the given document.

In the work done by Govindaru V. et al. (2012) the sentence level mood extraction for Malayalam text was focused. Semantic orientation method was used for mood extraction. In any language there are certain words which point out to emotional response of particular situation, which feature helps in analyzing sentiment.

Shweta Rana. (2014) proposed a method to extract the sentiments in hindi texts and determining the strength of opinion orientation on the product feature using fuzzy logic technique.

Evidently a lexical resource for determinig polarity of sentiment bearing words in Malayalam has not yet developed. So in this paper we have tried to create a Lexical resource for Malayalam, with polarities of Malayalam words which was helpful in multi domain sentiment analysis and special rules are written to handle exceptional cases, which was helpful in improving the accuracy of the system.

SYSTEM DESIGN AND IMPLEMENTATION

In this work, a domain independent sentiment analysis system which extracts the overall mood of Malayalam reviews is proposed. First a large lexical resource file was created with about 20000 subjective words and their positive and negative polarities. Later multi-domain reviews from various web sites were collected and a training corpus was created, that was required as certain categories of words alter the overall polarity of the sentence. Such categories are listed later. The next step involved was to manually tag the training data which was a tough task, because at times the same word may give different moods in different situations. For example, the word "maduthu" gives a negative mood usually. But, when a *positive* word like "chirichu" combines with it, the mood of the sentence "chirichu maduthu" becomes extremely *positive*. The tool used to train the system was TnT.

This work consists of five modules. They are Development of Malayalam Lexical Resource File, tokenizing, extracting polarities of sentiment words, applying rules for special cases and extrcating overall mood.

Development of Malayalam Lexical Resource File

The major task was to build a large lexical resource file with polarities of about 20000 sentiment words. This was prepared by taking Hindi SentiWordNet developed in IIT, Bombay as base and then translating each word from it to Malayalam. Billingual dictionary was used for this process as no efficient Hindi-Malayalam translator is available till date. This was a tedious job as it had to be done completely manually because of no readily available translators. The *positive* and negative polarities of the words were kept same as that of Hindi with the intuition that sentiment of a particular word doesn't change with the change in language. The score ranges between 0 and 1 with increasing polarity.

Handling Exceptions

After preparing the multi-domain sentiment lexical resource file, the system was also trained using appropriate tag set. This was done to extract the words of certain categories as listed below, that alters the overall sentiment of the sentence. In the training data, the data set was classified into seven classes. The classes are *'positive', 'negative', 'neutral', 'inversenegative', 'intensifier', 'dilator'* and *'special'*. Positive tag was given to the words that contribute happy mood to the comment. For example, "nallathaanu", "kandirikkaam" etc. Negative tag was given to the words that contribute sad mood to the comment. For example," mosham", "cheetha" etc. Neutral tag was given to the words that don't convey any particular mood to the system. For example, "cinema", "paattu" etc. 'Inversenegative' tag was given to such words that inverse the mood of the non-neutral word preceeding it. For example, the words "alla", "illa" etc. inverses the sense of the preceeding word from positive to negative and negative to positive. Another tag was

'Intensifier', it was given to words that intensifies the non-neutral word that follows it. For example, "valare", "orupaadu" etc. Similar to *`Intensifier'* tag, `Dialator0 tag was used which dialates the mood of the non-neutral word that follows it. For example, "kurachu", "lesham" etc. And the last tag given was `Special0 tag.

Such words inverses the mood of word preceeding it, like *'inversenegative'* tag. But unlike *'inversenegative'* tag, it gives a negative sense to the comment if the preceeding word is neutral. For example, the words "maduthu", "vayya" etc. Once the system is trained, it can accept the input text.

WORKING OF THE SYSTEM

I The tokenizing module takes the input text and divides the sentence into tokens. The next module finds the positive and negative polarities of each words from the lexical resource file. The tagging module gives appropriate tags for the tokenized words. After tagging the input sentence, certain rules are given that alter the specific polarities according to the tags preceeding or following it. After doing the necessary alterations the overall score of the sentence is calculated by finding the sum of individual word scores. The overall score being a positive value implies that the sentence is positive and the overall score being a negative value implies that the sentence is negative. Further, the amount of positivity and negativity in the sentence is calculated. Ranging from 0 to 1, the value implies the mood to be neutral to highly saturated. Later the percentage of positivity and negativity in the sentence is found out. Incorporating large lexical resource file with machine learning, thereby making this a hybrid approach, is what makes this work different from the existing systems.

Working of Rules

Rule 1: The content is iterated from the beginning and whenever an *'inversenegative'* tag is seen, the loop iterates backwards and finds out the first tag other than neutral and alters the sign of the score, thereby changing the polarity of the word.

Rule 2: In the case of an intensifier whenever an *'intensifier'* tag is found, the loop iterates forwards and finds the first non-neutral word that follows the "*intensifier*" tag. Then, the score of that tag is doubled.

Rule 3: In the case of a dilator whenever a 'dilator' tag is found, the loop iterates forwards and finds the first non-neutral word that follows the "*dilator*" tag. Then, the score of that tag is halved.

Rule 4: In the case of some special words like "maduthu", "vayya" etc. the loop iterates backwards and if the first seen word is a neutral word the sign of the score of special word is altered. If the first seen word is non-neutral the score of the special word is unchanged.

Algorithm

- Input: Malayalam review collected from web.
- Output: Percentage of positivity and negativity of the review.
- Steps:
  1. Tokenize the input
  2. Extract the polarity of each tokens from lexical resource file.
  3. Give the tokenized file as input to tagger.
  4. Tag the input using TnT tagger.
  5. Extract the tags into a new list and apply the rules.

  (a) If tag(i)==`inversenegative': Score(previous positive or negative word)=(-1*score(previous positive or negative word)).

  (b) If tag(i)=='intensifier': Score(next positive or negative word)=2*score(next positive or neagative word)

  (c) If tag(i)=='dialator':score(next positive or negative word)=1/2* score(next positive or negative word)

  (d) If tag(i)=='special' and tag(i-1)=='neutral': score(i-1)==(-1*score(i-1)).

  6. Calculate the overall score of the sentence by finding the difference of total positive and negative score.

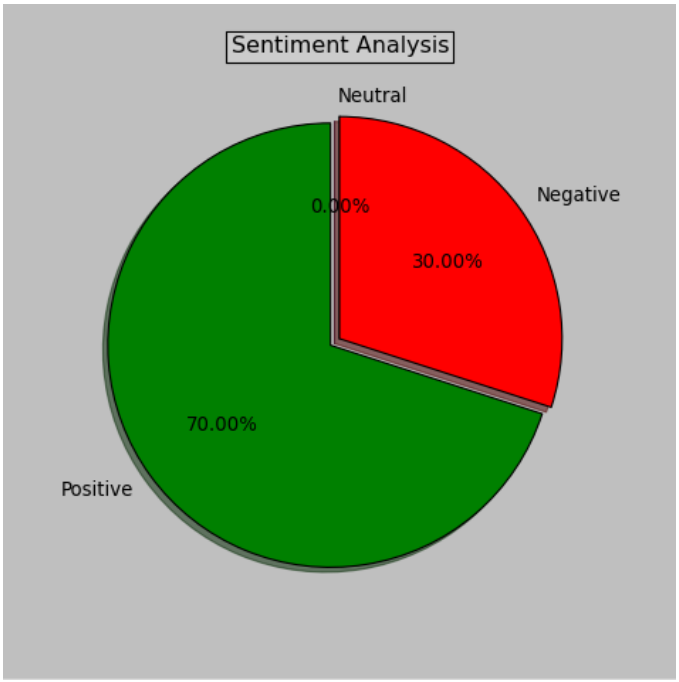

Figure 1. Pie-chart of the positive and negative percentage of the given review.

EXPERIMENTAL RESULTS

User reviews were collected from various online web sites. User reviews were given as input and the percentage of positivity and negativity in the review was obtained as output. The system generated output was compared with manually tagged output since no other work has been done in this particular area till the date. The system gave a performance rate of 93.6% which was the average of

judgement done by ten human judges, who were made to compare the manual and system generated outputs.

CONCLUSION AND FUTURE WORK

This paper proposes a Lexical Resource based hybrid approach to extract sentiments from domain independent Malayalam reviews. The proposed system finds out the polarity of subjective words from the input using the lexical resource file created. Simultaneously it tags the input file using TnT tagger and using the rules specified, exceptions are handled. After alterations the overall sentiment of the input is calculated. The experimental results show that the proposed method helps in decision making about a review effectively. The proposed system requires manual development of sentiment lexicons, to develop lexical resource file. Thus generation of SentiWordNet in Malayalam in order to reduce human intervention will be our priority. This will improve precision and recall for new domains. Use of sandhi-splitter to find the root form of words can be attempted which would make the system more efficient.

REFERENCES

[1] Aditya Joshi, Balamurali A R and Pushpak Bhattacharyya, "A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study", Proceedings Of 8th International Cconference on Natural Language Processing ,2010

[2] Chaumartin, Franc ois-Regis. "A knowledge based system for headline sentiment tagging" In Proceedings of International Workshop on Semantic Evaluations, pages 422-425, 2007

[3] Denecke, K, "Using sentiwordnet for multilingual sentiment analysis." In Proceedings of ICDE-8, volume 2, 2008

[4] Govindaru V. Neethu Mohandas, Janardhanan PS Nair, "Domain specific sentence level mood extractio from malayalam text", volume 1. International Conference on Advances in Computing and Communications, pages 78–81, 2012.

[5] Pushpak Bahattacharyya Alekh Agarwal, "Sentiment analysis; a new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified", Proceedings of the International Conference on Natural Language Processing (ICON), October 2005

[6] Shweta Rana, "Sentiment analysis of hindi text using fuzzy logic." Indian Journal of Applied Research, 2014.

[7] Subhabrata Mukherjee and Pushpak Bhattacharya "Sentiment Analysis, A Literature Survey", Indian Institute of Technology, Bombay Department of Computer Science and Engineering, 2012.