ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, Feburary 2016

Sentiment Analysis in Malayalam

Thulasi P K[1]

PG Student, Computer Science & Engineering, NSS College of Engineering, Palakkad, India [1]

Abstract: The most emerging area in NLP now a days is Sentiment Analysis (SA) which is a cognitive process in which the user's feelings and emotions are extracted. It has a variety of applications. It can be used to analyze whether the product review is positive or negative, based on tweets how people respond to adds, bloggers attitude about president changed since election, identifying child suitability of videos based on comments. Although there has been a lot of works published for universal languages like English, works on dialectal languages like Malayalam is comparatively less. But importance of Malayalam is increasing on social medias and shopping sites. This shows the scope of the topic. Another weak point of existing system is that the task done till today is only coarse grained in Malayalam considering only just classification of negative and positive polarity without considering the aspect on which the user is commenting. Such a fine grained task is also considered here most commonly known as Aspect based sentiment analysis. It can contribute to other fields like data mining and web mining.

Keywords: Sentiment Analysis, Aspect-Based Sentiment Analysis, Senti-Wordnet, Polarity, POS tagging.

I. INTRODUCTION

Natural Language Processing is the ability by which a computer program identifies what a human being has said in the exact context spoken by him. It can be also considered as a field of Artificial Intelligence. One of the most emerging field of NLP is Sentiment Analysis. It is in short a cognitive process which can extract user's feelings and emotions. In detail it is defined as subjective information extraction by the use of NLP, text analysis and computational linguistics. It is widely used for a variety of applications which include reviews, social medias for marketing and customer service. It is also otherwise called as opinion mining. It can be used to find out whether a film or product is having positive or negative reviews, to find out child suitability videos based on the comments, analysis of twitter comments etc. Various approaches can be applied for sentiment extraction like machine learning approaches which uses Support Vector Machines, Senti-Wordnet approach. The main problems of existing systems are that the task done today is only coarse grained considering only just creating an inbuilt polarity database and comparing the extracted word's polarity. There are only three main classes that are considered like Negative ,Positive and Neutral. It doesn't take into consideration fine grained tasks like finding the aspect on which the user is commenting. This type of sentiment analysis is often called as Aspect-Based Sentiment Analysis. Even though so many works on Sentiment Analysis have been proposed for Universal Languages like English , the works are very rare for dialectal languages. Through this paper an idea to do Aspect Based Sentiment Analysis for a dialectal language called Malayalam is proposed. Malayalam belongs to Dravidian family of Languages. In 2013 it has been declared as a classical language by Indian Government. The importance of Malayalam is increasing day by day in various shopping sites and in social medias. This reveals the importance of the work.

II. RELATED WORK

So many works have been proposed for Sentiment Analysis in Universal Languages like English, although it is comparatively less for Malayalam. Some of the important background papers that contributed to the proposed idea are been discussed here. Mood extraction [3] is a difficult task to refine moods such as happy, sad, angry etc. There are certain words which point to emotional response of a particular situation/object. Sandosham, dhukam are examples. This feature of a word which helps in analyzing sentiment in a particular context is known as semantic orientation. To do this Reference word sets depicting desirable and undesirable response is created manually.POS tagging of words are done and adjectives and adverbs are extracted based on the assumption that they are polarity words. Then compare the words with a reference word set for moods and do necessary semantic orientation calculation.

In SentiMa [2] rule based approach has been applied. First of all a data base is created which consists of positive and negative polarity words for Malayalam. The extracted words from the sentence is compared with the database to assign the corresponding polarity, if nothing matches then neutral polarity is given. Negation rule is applied here. Final polarity is calculated by summing up the polarity of each word in the sentence.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, Feburary 2016

Another approach [4] increased the number of tagged classes to seven rather than just positive, negative and neutral . The other classes involved inverse negative, intensifier, dialator and special. For each new category the polarity is recalculated based on an algorithm and the final polarity is calculated by summing up the polarity of each word.
This paper [1] involves the important concept to be implemented. Here Aspect-Based sentiment Analysis is done. The efficiency of the system is increased with the introduction of a module called Sent_Comp. This compression module gives output which consists of only sentiment related words .This produces the parse tree correct and it will be easy for finding the aspect on which the user is commenting.

III.PROBLEM DEFINITION

The main problems existing in this domain are there are only coarse grained tasks considering only polarity, but not fine grained tasks considering the aspect on which the user is commenting. Although so many works were proposed for universal languages like English, it is less for dialectal languages like Malayalam. Ambiguities and Language divergences are more . Large amount of data is required for data analysis.

IV.PROPOSED SYSTEM

The proposed system has the following architecture as shown in Fig 1. The detailed explanation of the architecture is as follows. First module is the sentence splitting module. Here we will first divide the sentence into different tokens.Sandhi splitter can be used for the purpose. Example can be given as

Cinema valare nannaytundu -> cinema+valare+nannaytundu

The second module consists of the POS tagging. A Malayalam POS tagger is used for the purpose.A TnT tagger is used for the purpose. There are two phases for this tagger. One is the training phase and the other is the tagging phase.In training phase untagged Malayalam corpus is given, it is tagged and trained to produce two types of files called Lexical file and N-gram file. The lexical file consists of frequency of word tag in training corpus. It is used to find out the lexical probability or word likelihood. The N-gram file is used to find contextual frequencies of unigrams, bigrams, trigrams etc. Third module is the compressor module. Here I will be using a senti-wordnet in Malayalam to find only sentiment related words. It will contain polarity like positive, negative ,neutral. Example is

Padam nannayilla
Padam -> neutral
Nannayilla -> negative

The word Padam is neutral, while the word Nannayilla is negative.Thus when compared to the database the polarity of the words are found and when we calculate the overall polarity then the polarity is negative.

Fourth module will store the aspect words in a separate file. Usually the aspect will be a Subject or a noun. This is an assumption.

Fifth is the tag module. Seven classes are used. They are
Positive -> nalladu
Negative -> cheetah
Neutral -> paatu
Inverse Negative -> alla
Intensifier -> valare
Dialator -> kurachu
Special -> maduthu

Based on these classes the polarity is calculated as follows
If tag is inverse negative then score(previous positive or negative word)=-1*(score (previous positive or negative word))

If tag is intensifier then score(next positive or negative word)=2*(score(next positive or negative word))
If tag is dialator score(next positive or negative word)=1/2*(score(next positive or negative word))
If tag is special then and tag of previous is neutral then score (previous)=(-1*score(previous))

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, Feburary 2016

Final module is calculation of whole polarity of the sentence. This is done by summing up all the polarity and if that is negative then the polarity is negative or else if it is positive then polarity is positive and else it is neutral. Suppose there is no polarity word in a sentence, like "angane alla". Here we will just extract the polarity only, without taking into consideration the aspect.

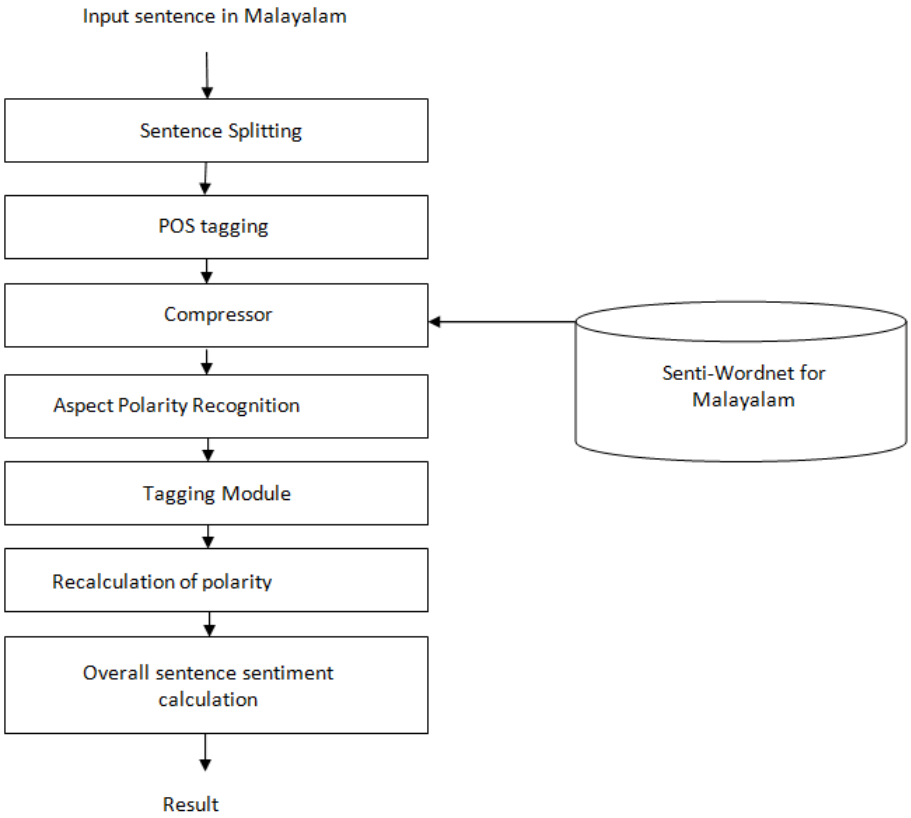

Fig.1.Proposed Method Architecture

V. COMPARISON OF RESULTS

As already mentioned this is a theoretical concept not yet evaluated. A comparison is done with regard to the existing system and the expected result is as shown below.

Table 1.Comparison of Results

Sentence Compression for Aspect-Based Sentiment Analysis	**SentiMa**	**Proposed Method**
Aspect-Based	Not Aspect-Based	
Syntactic Parse Tree	Polarity Checking (3)	Polarity Checking(7)
Sent_Comp	No Compression Module	Senti-Wordnet
Chinese	Malayalam	Malayalam
Result-60%	85%	86%

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, Feburary 2016

VI. CONCLUSION

Sentiment Analysis is an important area of Natural Language Processing. It has a lot of applications in product and film reviews, social networking sites etc.It plays a major role in data mining and web mining. The works are very less in dialectal languages like Malayalam even though so many are there for universal languages like English.Even though some of the works exist it doesn't take into consideration fine grained details like the aspect on which the user is commenting otherwise called as Aspect based sentiment Analysis. This work was an attempt to do aspect based sentiment analysis in Malayalam.

VII. FUTURE WORK

The current system can only analyse some of the emotion smileys. This can be extended by analysing more emotions. Aspect based sentiment analysis helps to improve the quality of the feature on which the author is commenting.

ACKNOWLEDGEMENT

I first of all thank my guide for her valuable suggestions and remarks. I then thank all my friends and colleagues who helped me, my parents and almighty.

REFERENCES

[1] Wanxiang Che,Yanyan Zhao,Honglei Guo,Zhong Su and Ting Liu,"Sentence Compression for Aspect- Based Sentiment Analysis",IEEE/ACM Tansactions on Audio,Speech and Language Processing,VOL.23,NO. 12,December 2015.
[2] Deepu S.Nair,Jisha P.Jayan,Rajeev R.R,Elizabeth Sherly,"SentiMa-Sentiment Extraction for Malayalam",2014.
[3] Neethu Mohandas,Janardhanan PS Nair,"Domain Specific Sentence Level Mood Extraction from Malayalam Text", International Conferene on Advances in Computing and Communications,2012.
[4] Anagha M, Raveena R Kumar, Sreetha K, Rajeev R.R,P.C.Raghu Raj,"Lexical Resource based Hybrid Approach for Cross Domain Sentiment Analysis in Malayalam ,International Journal of Engineering Sciences,2014.