DOI:

Polarity Classification of Malayalam Document-A Rule Based Approach

Lis Jose
Assistant Professor of Computer Science and Engineering
Amal Jyothi College of Engineering
Koovappally, Kanjirappally
lisjose@amaljyothi.ac.in

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. 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. Malayalam belongs to Dravidian family of Languages. In 2013 it has been declared as a classical language by Indian Government. It discuss about the sentimental analysis of Malayalam language. Malayalam words are classified into positive, negative and ambiguous words. Rule based approach is used and it mainly focused on the opinionated words in each text. Classifying the text based on number of positive and negative words.

Keywords—Rule-based approach, Sentiment Analysis

I. INTRODUCTION

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities. It is one of the most active research areas in natural language processing and is also widely studied in data mining. Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of our behaviors. Sentiment analysis and opinion mining mainly focus on opinions which express or imply positive or negative sentiments [1]. Natural Language Processing is vast area. It is also widely researched in data mining, Web mining, and information retrieval. In general, sentiment analysis has been investigated mainly at three levels.

The task at Document level is to classify whether a whole opinion document expresses a positive or negative sentiment. For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This task is commonly known as. document level sentiment classification. The task at Sentence level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to subjectivity classification which distinguishes sentences (called objective sentences) that express factual information from sentences (called subjective sentences) that express subjective views and opinions. Entity and Aspect level: Both the document-level and sentence level analysis do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Aspect level was earlier called feature level (feature-based opinion mining and summarization).

The section 2 describes the related works in sentiment analysis of tweets. Section 3 briefs the proposed method. Implementation details are specified in section 4. Section 5 shows the result analysis. Finally, the paper ends with conclusion.

II. RELATED WORKS

The rule-based approach has successfully been used in developing many natural language processing systems. Systems that use rule-based transformations are based on a core of solid linguistic knowledge. The linguistic knowledge acquired for one natural language processing system may be reused to build knowledge required for a similar task in another system. The advantage of the rule-based approach over the corpus-based approach is clear for: 1) less-resourced languages, for which large corpora, possibly parallel or bilingual, with representative structures and entities are neither available nor easily affordable, and 2) for morphologically rich languages, which even with the availability of corpora suffer from data sparseness. These have motivated many researchers to fully or partially follow the rule-based approach in developing their Arabic natural processing tools and systems [7].

Sentiment classification concerns the use of automatic methods for predicting the orientation of subjective content on text documents, with applications on a number of areas including recommender and advertising systems, customer

DOI:

intelligence and information retrieval. SentiWordNet is an opinion lexicon derived from the WordNet database where each term is associated with numerical scores indicating positive and negative sentiment information. Our approach comprises counting positive and negative term scores to determine sentiment orientation, and an improvement is presented by building a data set of relevant features using SentiWordNet as source, and applied to a machine learning classifier. We find that results obtained with SentiWordNet are in line with similar approaches using manual lexicons seen in the literature. In addition, our feature set approach yielded improvements over the baseline term counting method. The results indicate SentiWordNet could be used as an important resource for sentiment classification tasks. Additional considerations are made on possible further improvements to the method and its use in conjunction with other techniques [8].

III. PROPOSED SYSTEM

Sentiment analysis has a wide appeal as providing information about the subjective dimension of texts. It can be regarded as a classification technique, either binary (polarity classification into positive/negative) or multi-class categorization (e.g. positive/neutral/negative).Most approaches use a sentiment lexicon as a component (sometimes the only component). Lexicons can either be general purpose, or extracted from a suitable corpus.

A. Sentence Splitting

Before splitting the sentence, should consider a document. Divide each sentence in the document into tokens. Splitting can be done by using some split operations.

B. 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.

C. Aspect Polarity Recognition

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.

D. Tagging Module

Seven classes are used. They are

- Positive
- Negative
- Neutral
- Inverse Negative
- Intensifier
- Dialator
- Special

E. Calculation of polarity

- 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))

F. Recalculation of polarity

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.

IV. IMPLEMENTATION DETAILS

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. There are only three main classes that are considered like Negative ,Positive and Neutral. Even though so many works on Sentiment Analysis have been proposed for Universal Languages like English , the works are very rare for dialectal languages.

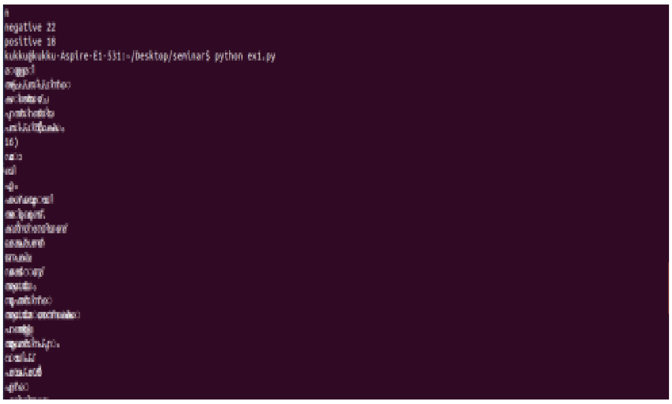
Sentence Splitting: Before splitting the sentence, we should consider a document. Divide each sentence in the document into tokens. Splitting can be done by using some split operations. Tagging Module Prepare two word documents which contain the positive and negative words. By using these words are comparing the documents tokenized words with the word documents that are classified as positive

DOI:

and negative. Calculation of polarity 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 it just extract the polarity only, without taking into consideration the aspect. Summary of Method 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.

V. RESULTS

Display the number of positive words and negative words in the positive, negative document by comparing it with the document that is tokenized.



```
negative 22
positive 18
kukku@kukku-Aspire-E1-531:~/Desktop/seminar$ python ex1.py
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
16)
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
```

VI. CONCLUSION

The methodology will result in splitting the words, POS tagging of the words, calculation of the polarity and display the positive words as positive and negative words are negative.

REFERENCES

[1] Liu, Bing Sentiment analysis and opinion mining,Synthesis lectures on human language technologies 5.1 (2012) : 1-167

[2] Poornima Mehta, Satish Chandra, Parameter Tuning in Updating the Sentiment Polarity of Objective Words in SentiWordNet. IEEE,2015.

[3] Thulasi P K Sentiment Analysis in Malayalam, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Special Issue 1, Feburary 2016.

[4] Shankar R , Shilpa K. M , Sridhar Patil, Suma Swamy 2A Survey on Sentimental Analysis in Different Indian Dialects, International Journal of Advanced Research in Computer and Communication Engineering,Vol. 5, Issue 4, April 2016.

[5] Afraz Z. Syed, Muhammad Aslam,Ana Maria Martinez-EnriquezLexicon Based Sentiment Analysis of Urdu Text Using SentiUnits,Springer ,2010

[6] Thanyalak Rattanasawad, Marut Buranarach, Ye Myat Thein,Thepchai Supnithi, and Kanda Runapongsa Saikaew Design and Implementation of a Rule-based Recommender Application Framework for the Semantic Web Data,2010.

[7] Khaled ShaalanRule-based Approach in Arabic Natural Language Processing, International Journal on Information and Communication Technologies, Vol. 3, No. 3, June 2010.

[8] Bruno Ohana and Brendan Tierney Sentiment classification of reviews using SentiWordNet, 9th. IT T Conference,2009.