Sentiment Analysis in Marathi using Marathi WordNet

Chitra V. Chaudhari[1], Ashwini V. Khaire[2], Rashmi R. Murtadak[3], Komal S. Sirsulla[4]

[1,2,3,4]Department of Infomation Tehnnology Matoshri College of Engineering & Reasearch Centre, Nashik.

Abstract: As large amount of work in sentiment analysis has been done in English and Hindi language, as English and Hindi are the language mostly used in India, but there is a need to perform sentiment analysis in Marathi language for the state Maharashtra. 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. It deals with analyzing emotions, feelings, and the attitude of a speaker or a writer from a given piece of text. Sentiment Analysis involves capturing of user's behavior, likes and dislikes of an individual from the text. The target of SA is to find opinions, identify the sentiments they express, and then classify their polarity. Our proposed approach for Sentiment Analysis of Marathi documents uses Gate Processor (Natural language processor) to find the overall polarity of the document, the final aggregated polarity can be positive, negative or neutral .The idea is to propose an efficient approach for identifying sentiments and opinions from user generated content in Marathi and focus on investigating the influence of negation handling and discourse relations on the performance of Marathi review.

Key Words: Sentiment Analysis (SA), Polarity, Natural Language Processing (NLP), Marathi WordNet, GATE Processor

1. Introduction

The sentiment found within comments, feedback or critiques provide useful indicators form any different purposes. Sentiment Analysis is a task under natural language processing which finds orientation of a person opinion or feelings over an entity[1]. It deals with analyzing personal emotions, feelings, attitude and opinion of a speaker or a writer over an object. The primary target of SA is to find the sentiments expressed by person over an information or entity [2].

Marathi is an Indo-Aryan Language Spoken predominantly by the Marathi people of Maharashtra. It is official and co-official language in the Maharashtra and Goa states of Western India, respectively. There were 73 million speakers in 2007. [3]

There is no existing system which calculate sentiment associated with Marathi text as sentiment analysis is very difficult for Marathi language due different complexity associated with Marathi text but for other languages like Hindi, English it is easily available.

Marathi language lacks availability of efficient resources like parser and tagger which are essential for extracting sentiment. There are many scenarios where same words may be used in multiple contexts and context dependent word mapping is still a difficult task, error prone and requires manual efforts to find accurate polarity of word.

Proposed system is defined which uses Marathi WordNet to calculate polarity associated with sentiment words Sentiment analysis has lead to determine ones attitude of their speaking or writing through calculating polarity associated with the information. Sentiments can be mined from various sources like texts, tweets, news articles, comments, blogs, social media or from any source of information which is either available online or offline. Sentiment Analysis has been quite popular and has lead to building of better products, understanding user's opinion, executing and managing of business decisions. People rely and make decisions based on reviews and opinions. This research area has provided more importance to the mass opinion instead of word-of-mouth.The discipline that covenant with the automatic handling of natural language in speech or text is called Natural Language Processing (NLP). NLP research areas constitutes either general NLP problems, which need to be decipher in any appliance areas like Co-reference resolution, Word Sense Disambiguation or a specific end applications like Questioning and Answering, Machine Translation, Text Summarization, Information Recovery, Information Mining. Many traditional data mining errands in NLP basically focus on extorting data from documents according to a topic.[4]

General Architecture for Text Engineering or GATE is a Java suite of tools originally developed at the University of Sheffield beginning in 1995 and now used worldwide by a wide community of scientists, companies, teachers and students for many natural language processing tasks, including information extraction in many languages. GATE includes an information extraction system called ANNIE (A Nearly-New Information Extraction System) which is a set of modules comprising a tokenizer, a gazetteer, a sentence splitter, a part of speech tagger, a named entities transducer and a co-reference tagger. ANNIE can be used as-is to provide basic information extraction functionality, or provide a starting point for more specific tasks. JAPE transducers are used within GATE to manipulate annotations on text.

1.1. Objectives

The proposed system will help to identify the emotions from the text in Marathi language which can either be positive, negative or neutral and also it will help the people conversant to express their views in other languages so that the Maharashtrians can give reviews in Marathi.

2. Literature Survey

In this section we cite the relevant past literature of research work done in the field of sentiment analysis for Hindi language because no work done for Marathi language.

This paper [4] presents system which uses HSWN (HindiSentiWordNet) to extract polarity associated with sentiment words. This paper allows finding sentiment associated with Hindi movie review where overall polarity of the review is classified as positive, negative or neutral using HindiSentiWordNet. To improve the accuracy of finding sentiment synset replacement algorithm is used which helps to find polarity of words which are not available in HSWN.

In this paper [5] an algorithm is proposed for Sentiment analysis of Movie reviews in Hindi with the help of pre-annotated corpus consisting of words/phrases and negation handling of review. Also overall polarity if defined in the review then system recognizes the overall polarity of the system and assign that polarity to the review. Experimentation results indicate that the proposed algorithm is performing well in this domain and achieved the accuracy of 85.42.

Sentiment analysis for the movie reviews is calculated by the authors [6]. They determine the polarity of sentiments of the person in the review and comments when the sentences occur in documentary level and uses SentiWordNet dictionary to determine the scores of each word present in the comment.

They also use rule base and fuzzy logic approach to give the output and achieved the accuracy of 56.

Opinion Mining System [7] is proposed by authors named as Hindi Sentiment Orientation System which is based on Hindi language. Unsupervised approach which based on using dictionary is used to determine the polarity of reviews of user written in Hindi language. Challenges like negation associated in text which reverse the sentiment are also handled. The accuracy of system is evaluated by using 50 sentences of movie reviews and there result showed the accuracy of 65 percent.

3. Proposed System using Marathi WordNet

Sentiment is nothing but an attitude toward something; regard opinion.
Sentiment analysis aims to determine the attitude of the speaker or writer with respect to some topic or the overall contextual polarity of a document.
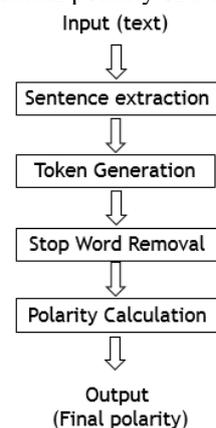

Figure 1. System Architecture.

As given in the system architecture, we give an input to the system. Input is in the form of text or document. That text is given to the sentence extraction which will separate the sentences present in the document.

After separating the sentences tokens or words are separated from the sentence.

The remove stop words block will remove the unnecessary tokens from the remained sentences.

After that the calculate polarity block will check the word in the Gate Processor's Positive and negative words list. And then add the positive and negative polarity of them for a single sentence. Then calculate the polarity for all the sentences present in the document.

Finally aggregate polarity of document is found out where we can classify overall polarity as positive, negative or neutral.

Here each token in the document is matched with the words stored in the Gate Processor's Dictionary that is ANNIE to calculate its relevant polarity. If

there are still words that are not stored in Gate Processor then those words will be saved in separate list so that they can be handled next time for improvement.

4. Methodology

In this paper we collect all the users' reviews which contain the facts and opinions from different resources. These sentences are further processed for extraction to classify as positive, negative or neutral opinions. The final score of each individual sentence is calculated after considering the whole sentences.

4.1. Algorithm

Algorithm is a Step by step procedure designed to perform an operation, and which will lead to the accurate result.[10]
Input: Marathi Sentence
Output: Analysis of Marathi Sentence
Step 1: For each token in the document.
Step 2: Check If (word is present in the list)
Then Retrieve polarity (POL) and go to Step 4
Step 3: If (no polarity assigned to word)
Then fetch next token and go to Step 2
Step 4: word POL= (POL);
End For Loop and go to Step 6 when all tokens are processed.
Step 5: Compute the aggregate polarity of the document (doc POL) by adding the polarities values of the entire tokens.
Step 6: If (doc POL > zero)
Then label the document as positive
Else If (doc POL<zero)
Then label the document as negative
Else classify the document as neutral

4.2. Test Cases

Input to the system is in the form of single sentence or document. Following example demonstrate working of our proposed system:
Sample input document:
दंगल हा खूप सुंदर चित्रपट आहे.यात आमीर खानचे काम प्रशंसनीय आहे.यात त्यांनी एका कठोर मन असलेल्या वडिलांचा अभिनय केला आहे.महिला आणि राष्ट्रीय खेळाडूना प्रेरणा देणारा आहे.एकत्रित हा चित्रपट तुमचे मनोरंजन करेल.
Sentence 1: दंगल हा खूप सुंदर चित्रपट आहे.
Sentence 2: त्यांनी एका कठोर मन असलेल्या वडिलांचा अभिनय केला आहे
Sentence 3: यात अमीर खानचे काम प्रशंसनीय आहे.

Sentence 4: महिला आणि राष्ट्रीय खेळाडूना प्रेरणा देणारा आहे
Sentence 5: एकत्रित हा चित्रपट तुमचे मनोरंजन करेल.
Stop word removal:
Sentence 1: दंगल खूप सुंदर चित्रपट.
Sentence 2: एका कठोर मन वडिलांचा अभिनय.
Sentence 3: अमीर खानचे काम प्रशंसनीय.
Sentence 4: महिला राष्ट्रीय खेळाडूना प्रेरणा देणारा.
Sentence 5: एकत्रित चित्रपट तुमचे मनोरंजन.
Calculating Polarity:
Positive Words: सुंदर, प्रशंसनीय, प्रेरणा, मनोरंजन
Negative Words: कठोर
Overall polarity: positive.

5. Conclusion & Future Scope

Sentiment analysis has lead to determine the attitude or inclination of a communicator through the contextual polarity of their speaking or writing. Sentiment Analysis has been quite popular and has lead to building of better products, understanding user's opinion, executing and managing of business decisions. People rely and make decisions based on reviews and opinions. This Existing systems research area has provided more importance to the mass opinion instead of word-of-mouth, with the system in their daily spoken natural language. One drawback of an NLP based approach is that it would likely perform very poorly when used on grammatically incorrect text.The system can be enhanced by adding more Marathi words into it. To improve the accuracy of finding sentiment synset replacement algorithm is used which helps to find words which are not available in the Marathi WordNet. This Proposed system can be used for different languages like Manipuri, Gujrati,etc.

6. References

[1] Aditya Joshi, Balamurali A, Pushpak Bhattacharyya,"A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study",Proceedings of ICON 2010: 8th International Conference on Natural Language Processing, year 2010
[2] Piyush Arora, "Sentiment Analysis for Hindi Language",year 2013.
[3]http://en.m.wikipedia.org/wiki/marathi-language
[4] Pooja Pandey, Sharvari Govilkar, in the year June 2015 published "A Framework for Sentiment Analysis in Hindi using HSWN", June 2015
[5] Alok Gupta, DipeekaSonavane, Kajal Attarde, Neelam Shelar, Pramila Mate, "Sentiment analysis of movie reviews in hindi", March 2016,
[6] Pranali Tumsare, Ashish .S. Sambare and Sachin .R. Jain, "Opinion Mining In Natural Language Processing Using Sentiwordnet and Fuzzy ". June 2014

[7] Richa Sharma, Shweta Nigam and Rekha Jain,"Polarity Detection Of Movie Reviews In Hindi Language" International Journal on Computational Sciences and Applications (IJCSA) August 2014.

[8] Balamurali, A. R., Aditya Joshi, and Pushpak Bhattacharyya, in the published "Robust sense-based sentiment classification," Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Association for Computational Linguistics, year 2011

[9]https://en.wikipedia.org/wiki/General_Architecture_for_Text_Engineering

[10]http://www.businessdictionary.com/definition/algorithm.html