Research Paper

Volume : 4 | Issue : 6 | Special Issue June 2015 • ISSN No 2277 - 8179

A Novel Hybrid Approach Based on Maximum Entropy Classifier for Sentiment Analysis of Malayalam Movie Reviews

Engineering

KEYWORDS: S entiment Analysis, Maximum Entropy, Malayalam Movie Reviews.

Anagha M	Dept. of Computer Science and Engineering Government Engineering College, Sreekrishnapuram Palakkad, Kerala, India
Raveena R Kuma	Dept. of Computer Science and Engineering Government Engineering College, Sreekrishnapuram Palakkad, Kerala, India
Sreetha K	Dept. of Computer Science and Engineering Government Engineering College, Sreekrishnapuram Palakkad, Kerala, India
P C Reghu Raj	Dept. of Computer Science and Engineering Government Engineering College, Sreekrishnapuram Palakkad, Kerala, India

ABSTRACT

S entiment analysis is an application of Computational Linguistics and Text Mining, in which the hidden emotions in a given text are extracted. In this paper, S entiment Analysis of Malayalam movie reviews is done by classifying the review obtained from user as positive, negative and neutral. A hybrid approach for S entiment Analysis is proposed in this work in which Maximum Entropy Model is used for tagging and certain rules are also incorporated to handle certain special cases. The Maximum Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. It is based on the Principle of Maximum Entropy. It selects the one which has the largest entropy and from all the models that fit our training data. Maximum Entropy Classification finds out in which class the review must belong, given a context so that it maximizes the entropy of the classification system. The rules included ensure that special cases are handled which include negation, intensifiers, dilators etc. The system performed well gi ving a considerable precision rate.

I. INT RODUCT ION

Sentiment classification is one of the most challenging problems in Natural Language Processing. Today Internet has large amount of reviews and feedbacks on almost everything. These include product reviews, feedbacks etc.

Sentiment Analysis focuses on identifying whether a given piece of text is subjective or objective and if it is subjective, then whether it is negative or positive. In business intelligence sentiment analysis play crucial role, it helps the manufacturer of products in analysing product reviews written by users and thus it helps to enhance product quality and services.

In politics, it helps public to identify their right politician by analysing comments. In daily life user review about particular product gives more information while shopping particular product. This will give us better decision while buying the product [6]. Sentiment analysis is a tough task as the sentiments are expressed in natural language.

This paper addresses the problem of sentiment analysis of Malayalam Movie reviews. Malayalam language is free ordered and highly agglutinative in nature, which makes extracting the sentiments from a Malayalam sentences much more difficult task [4]. A hybrid approach for analyzing the sentiment of Malayalam movie reviews is created.

Maximum entropy method is used for tagging the sentences. In Maximum Entropy classification, the probability that a document belongs to a particular class given a context must maximize the entropy of the classification system. By maximizing entropy, it is ensured that no biases are introduced into the system [5]. After tagging the sentence rules are applied for handling special cases, which helps to improve the performance of the system.

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

II. RELAT ED WORK

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, and rule based methods and fuzzy logic.

There are different domain interestingly studied by many researchers such as movie review, product review, travel destination review, social networking, e-learning and many more

Pang and Lee [2] compares different machine learning method for sentiment analysis, and categorises the review to positive or negative.Denecke [3] uses SentiWordNet for determining the polarity of text within a multilingual framework. Aditya Joshi, Balamurali A R, Pushpak Bhattacharyya, [1] 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. [4] sentence level mood extraction for Malayalam text was focused. Semantic orientation method was used for mood extraction.

III. SYST EM IMPLEMENT AT ION

In this work, a domain dependent sentiment analysis system which extracts the overall mood of Malayalam reviews is proposed in which Maximum Entropy Model is used for tagging and certain rules are also incorporated to handle certain special cases. The Maximum Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. It is based on the Principle of Maximum Entropy. It selects the one which has the largest entropy and from all the models that fit our training data. Maximum Entropy Classification finds out in which class the review must belong, given a context so that it maximizes the entropy of the classification system.

The training phase consists of five modules. They are, 1. Collection of data corpora, 2. Tokenizing the data, 3. Preparation of tagged dataset, 4. Training the system, 5. Validating the system. If the system did not function as per the expectations during the validation, the size of the dataset is increased and the process is repeated until a notable result is obtained.

In this work, different movie reviews from various web sites were collected and a training corpus of about 10000 reviews was developed. The first 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. 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 t hat inverse the mood of the non-neutral word preceding it. For example, the words "alla", "illa" etc. inverses the sense of the preceding 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, 'Dialator' 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 'Special' tag. Such words inverses the mood of word preceding it, like 'inversenegative' tag. But unlike 'inversenegative' tag, it gives a negative sense to the comment if the preceding word is neutral. For example, the words "maduthu", "vayya" etc.

Once the system is trained, it can accept the input text. When an input text is given, it also passes through the same modules. The tokenizing module takes the input text and divides the sentence into tokens. The tagging module gives appropriate tags for the tokenized words, using maximum entropy classifier. After tagging the input sentence, certain rules are given that alter the specific tags according to the tags preceding or following it [7]. After making necessary corrections, the last module counts the number of positive and negative tags and calculates the overall positive and negative percentage of the given review.

Working of Rules

Rule 1: When an "inversenegative" tag is identified previous positive or negative tag is flipped.

Rule 2: When an "intensifier" tag is found, the tag of next positive or negative word is found and count of that tag is increased by a half.

Rule 3: When a "dilator" tag is found, the tag of next positive or negative word is found and count of that tag is reduced by a half.

Rule 4: In the case of some special words like "maduthu", "vayya" etc., if the previous tag is neutral then the special tag is flipped to negative. If the first seen word is non-neutral the tags are unchanged.

IV. EXPERIMENTA L 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.

V. CONCLUSION AND FUT URE SCOPE

This proposed system helps us to extract the sentiment in the reviews with maximum entropy classification. The maximum entropy method is used for the tagging purpose. The sy stem shows a considerable performance. In future we would like to add sandhi-splitter to find the root form of words can be attempted which would make the system more efficient. Also some of the challenges like Implicit Sentiment and Sarcasm, Thwarted Expectations etc.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Elizabath Sherly, Rajeev R R and Jisha P Jayan of IIITM-K for providing guidance, technical support and resources.

REFERENCE

[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] B. Pang, L.Lee and Vaithyanathan, S. 2002. "Thumbs up? Sentiment classification using machine learning techniques ". In Proceedings of the conference on empirical methods in natural language processing (EMNLP 2002) Philadelphia, PA, USA, (2002), pp:79– 86. | [3] Denecke, K, "Using sentiwordnet for multilingual sentiment analysis." In Proceedings of ICDE-8, volume 2, 2008 | [4] Govindaru V. Neethu M ohandas, 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] Nigam K., Lafferty J., and M cCallum A. using maximum entropy for Text Classification. In Proc of the IJCAI-99 Workshop on M achine Learning for Information Filtering(1999) | [6] Virendrakumar Dhotre and Balaji Jagtap , "SVM and HM M Based Hybrid Approach of Sentiment Analysis for Teacher Feedback Assessment," International journal of emerging Trends & Technology in Computer Science(IJETTCS), Volume3, Issue, M ay -June 2014 | [7] Anagha M , Raveena R Kumar, Sreetha K, "Fuzzy Logic Based Hybrid Approach for Sentiment Analysis of M alayalam M ovie Reviews," IEEE Spices , Feb 2015, pg 773-776