A SUBJECTIVE FEATURE EXTRACTION FOR SENTIMENT ANALYSIS IN MALAYALAM LANGUAGE

Jisha P. Jayan[1], Deepu S. Nair[2], Elizabeth Sherly[3]

*[1]Virtual Resource Center for Language Computing(VRCLC), Indian Institute of Information Technology and Management-Kerala , Thiruvananthapuram
jisha.jayan@iiitmk.ac.in[1], deepu.s@iiitmk.ac.in[2], sherly@iiitmk.ac.in[3]

Abstract: In recent days, Sentiment Analysis has become an active research in NLP, which analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from writing language. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, and social network. In his paper, sentiment analysis of Malayalam film review is carried out using machine learning techniques CRF combined with a rule based approach. The system shows 82 % accuracy.

INTRODUCTION

Sentiment Analysis (SA) is a process that helps to extract the subjective or conceptual information from various sources. It deals with analyzing emotions, feelings and the attitude of a speaker or a writer from a given piece of text. In a broader sense, SA is a cognitive process which helps computer to understand and extract human behavior such as likes and dislikes, feelings and emotions and many other attributes and also to predict the behavioral aspects of human. It is also used for opinion mining, one of the hottest topics in NLP that helps to identify and extract subjective information in source materials and provides valuable insights about the user's intentions, taste and likeliness etc.

Facts and opinions are two main types of textual information in the world. Facts are objective expressions about entities, events and their properties while opinions are usually subjective expressions that describe people's sentiments, feelings toward entities, events and their properties. Opinion expressions convey people's positive or negative sentiments and it may be a neutral comment. Present research on textual information processing has been focused on mining and retrieval of factual information.

The web has dramatically changed the way that people express their views and opinions. Common men can now post reviews of products at various online product review sites and express their views on almost anything in Internet forums, discussion groups, social media and blogs, which are collectively called the user-generated content. This would lead to measurable sources of information, which helps to improve the quality of the product and for a better feedback and choice to the user. Such system can be further modified to an automatic textual analysis for sentiments, automatic survey analysis, opinion extraction, or a recommender system. Such system typically tries to extract the overall sentiment revealed in a sentence or document, either positive or negative, or neutral.

Malayalam belongs to the Dravidian family, a large family of languages of South and Central India, and SriLanka. Malayalam exhibits heavy amount of agglutination. Due to the agglutination and rich morphology of words along with high ambiguity of Malayalam language, research in NLP for Malayalam is always challenging and is same for sentiment analysis because of its high dependence on words that are used for expressing the feelings or other sentiments.

The sentence-level and document-level review has been considered. Focus on the sentence-level sentiment extraction is significant because in most of the websites, user comments are just a single sentence. Document -level provides the semantics of the entire document, but often fails to detect sentiment about individual aspects of the topic. A statistical approach using simple co-occurrence that commonly used machine learning techniques is a trivial approach, but fail to provide a better result, especially in cases where both negative and positive comes in two differ sentences in a document. In order to resolve such shortcoming, we propose a hybrid statistical model using rule based and extracting the grammatical features.

The paper is organized into different sections. First section dealt with the introduction about SA and the objective of the paper. The second section exposed the states of the art that provides some of the major work carried out in this area. The third section reveals the proposed work and the methodologies. The fourth section includes the implementation and the result obtained. The fifth section concludes the paper.

STATES OF THE ART - SENTIMENT ANALYSIS

There has been a wide range of work carried out on this topic. The main research carried out in the area of sentiment analysis is in the document and sentence level. Document and sentence level classification methods are usually based on the classification of review context or words. Most of the work done is by using either of these three methods, Semantic Orientation method, Machine Learning method or Rule Based approach.

One of the first attempts in this field was done by Alekh Agarwal and Pushpak Bhattacharyya [1] for English. In this paper they made an attempt to determine the overall polarity of a document, such as identifying for the appreciation or criticism of a movie. They presented machine learning based approach to solve the problem of determining the sentiments similar to text categorization. The movie review was selected for their experiments. Their paper concluded with an accuracy of over 90% for the first time.

Another work on the sentiment extraction of movie was done by Pang [2]. The ultimate aim of that work was to find the best way to classify the sentiment from text, either standard machine learning techniques or human-produced baseline. Three different machine learning techniques explained were mainly Maximum Entropy, Support Vector Machine, and Naive Bayes. In their experiment, they tried different variations of n-gram approach like unigrams presence,

unigrams with frequency, unigrams with bigrams, bigrams, unigrams with POS, adjectives, most frequent unigrams, unigrams with positions They concluded that machine learning techniques are quite good in comparison to the human generated baseline. The paper also remarked that the Naïve Bayes approach tend to do the worst while SVM performs the best.

Manurung, and Ruli [3] work was carried out in 2008 for Indonesian Language using machine learning method. In this work, he initially translated English movie review into Indonesian language and then applied to the machine learning approach such as Naive Bayes, SVM, and maximum Entropy method to perform the sentiment classification. He reached at the conclusion that SVM is the best classification method giving 80.09% accuracy.

Saggion and Funk [4] used senti-wordnet to perform opinion classification. They calculated positive and negative score for a review and based on the maximum score, the polarity of the review was assigned. They also extracted features and used machine learning algorithms to perform classification of the sentiments from the text.

Turney [5] also worked on part of speech (POS) information. He used tag patterns with a window of maximum three words using trigrams. In his experiment, he considered JJ, RB, NN, NNS POS-tags with some set of rules for classification of product reviews. He used adjectives and adverbs for performing opinion classification on reviews. PMI-IR algorithm is used to estimate the semantic orientation of the sentiment phrase. He achieved an average accuracy of 74% on 410 reviews of different domains collected from opinion.

Barbosa [6] designed a 2-step automatic sentiment analysis method for classifying tweets. They used a noisy training set to reduce the labelling effort in developing classifiers. First, they classified tweets into subjective and objective tweets. Then subjective tweets are classified as positive and negative tweets. Celikyilmaz [7] design a pronunciation based word clustering method for tweet normalization. In pronunciation based word clustering, words having similar pronunciation are clustered and assigned common tokens. They also used text processing techniques like assigning similar tokens for numbers, html links, user identifiers, and target organization names for normalization. After doing normalization, they used probabilistic models to identify polarity lexicons. They performed classification using the BoosTexter classifier with these polarity lexicons as features and obtained a reduced error rate.

In Malayalam, the works on sentiment analysis is in its infant stage. Geethu Mohandas [8], had proposed a semantic orientation method for extraction of the mood from any sentence. In their study, they have applied the semantic orientation method using an unsupervised learning technology for classifying the input text for classification. They used a tag set which includes the tags sorrow, joy, anger and neutral for tagging the manually created corpus and then calculated the semantic orientation by semantic association using SO-PMI (Semantic Orientation from Point wise Mutual Information). They concluded their paper with a conclusion that the SO-PMI method gives about 63% accuracy.

PROPOSED WORK

The proposed work concentrates on sentiment analysis to find the positive, negative or neutral opinions from the user's writings at the document level. The polarity of sentence and rating of individual category, such as film, direction, acting, song, script etc. is individually computed. The different suggestions, opinion and feedback about the film by considering different factors improve the overall ranking in a more meaningful manner and also item wise scoring. This work has been implemented on a hybrid approach combining the machine learning technique with rules. Since Malayalam is a highly agglutinative language with rich morphology, and also of free order, it has a wide range of fluctuated words with the same meaning. Also such reviews, there are a number of colloquial usages, short forms and broken sentences. Here we used Conditional Random Field (CRF) techniques for proper extraction and classification of sentiments.

Conditional Random Fields (CRFs) is a probabilistic framework for labeling and segmenting structured data, such as sequences, trees and lattices. The underlying idea is that of defining a conditional probability distribution over label sequences, given a particular observation sequence, rather than a joint distribution over both label and observation sequences. The primary advantage of CRFs over Hidden Markov Models is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. Additionally, CRFs avoid the label bias problem, a weakness exhibited by Maximum Entropy Markov models (MEMMs) and other conditional Markov models based on directed graphical models.

IMPLEMENTATION

A document level feature extraction and analysis is carried out for finding the polarity and rating of the film reviews. The polarity indicates positiveness, negativeness or neutrality of the document and the rating gives the rate in each category separately. The categories mainly dealt with our song, acting, direction, script and film. POS tagging is performed to the sentence, but for many of the attributes in sentiments requires additional tagsets, that is being included in the proposed method for better analysis and prediction. The training and testing process using CRF is depicted in Figure 1and 2.

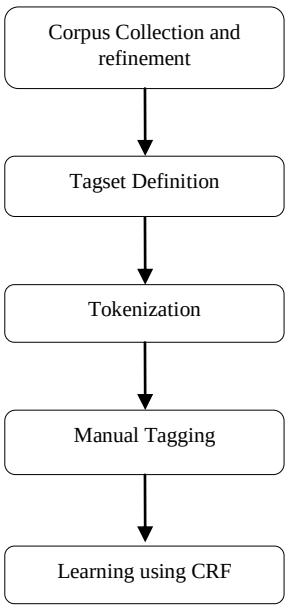

Figure1: Training Phase

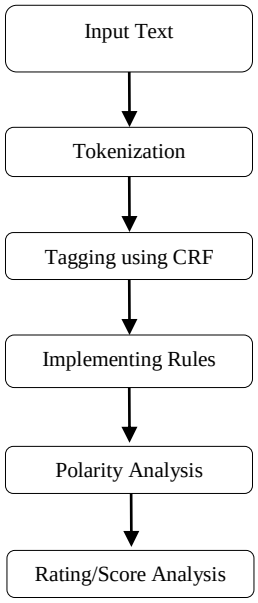

Figure 2: Testing Phase

TAGSET DEFINITION

The additional tagset definition for sentiments is an important task in this work. There is no exact and standard tagset for the sentiments in Malayalam presently, without that proper tagging is difficult. We have defined additional 10 tags for sentiment analysis in our study. Tagging not only depends on word but also the context of that particular document. The same words have the different tags in different contexts.

Table I. Sentimental Tags

Sl No.	Tags	Example	Description
1	CC_CCD	പക്ഷെ	Conjunction
2	DIR	സംവിധാനം , സ്ക്രിപ്റ്റ്	Direction /Script
3	INEG	എന്നല്ല	Inverse Negative
4	INTF	വളരെ	Intensifier
5	NEG	മോശമാണ്	Negation
6	NEU	ഒരുക്കമായി	Neutral
7	POS	മികച്ചതാണ്	Positive
8	RD_PUNC	.	Sentence Ending
9	SPCL	സിനിമയാണ് , ചിത്രത്തിന്	Film
10	TST	കഥാപാത്രം	Acting / Song

MACHINE LEARNING USING CRF

The collected and refined corpus has been tagged manually for training the engine. Here the engine has the capability of recalling the previous experience, there by learns for better classification. About 30000 tokens and its tags were used for training. The rules were also implemented appropriately with respects to the various semantics.

Algorithm

Step 1: Take Input

Step 2: Classification Using CRF (Tagging)

Step 3: Analyze the tagged output

Step 4: Apply 9 rules for finding the polarity of the sentence or document (Positive, Negative, Neutral)

Step 5: Find the rating of the individual category (Excellent, good, not bad, bad, not good, worst)

Step 6: Results

Step 7: Exit

RESULT

The system has been analyzing the sentiments from Malayalam film review at document level. Find out the polarity and rating of individual category. The categories which are Film, Direction/Script, Song/Acting, and other factors and attained an overall accuracy of 82 %.

Eg: Input text: കുറ്റകൃത്യവും ദുരൂഹതയും അന്വേഷണവും സിനിമയില് മാന്യമായി പറയാനറിയാവുന്ന ആളാണ് താനെന്ന് ആദ്യചിത്രമായ ‘ ഡിറ്റക്ടീവി ’ലൂടെയും നാലാമത്തെ ചിത്രമായ ‘ മൈമറീസി ’ലൂടെയും തെളിയിച്ച സംവിധായകനാണ് ജീത്തു ജോസഫ് . എന്നാല് മലയാളികള്ക്ക് അത്ര പരിചിതമല്ലാത്ത കുടുംബത്തിലുള്ള ഗണത്തിലാണ് പുതിയ ചിത്രമായ ‘ ദൃശ്യം ’ ജീത്തു ഒരുക്കിയിരിക്കുന്നത് . മലയോരഗ്രാമത്തിലെ സാധാരണ കുടുംബത്തിലുണ്ടാകുന്ന ഗൗരവകരമായ പ്രതിസന്ധി തന്ത്രപരമായി കൈകാര്യം ചെയ്യുന്നതെങ്ങനെയെന്ന് ത്രില്ലടിപ്പിക്കും വിധം പറഞ്ഞാണ് ‘ ദൃശ്യ ’ത്തെ സംവിധായകന് സമ്പന്നമാക്കുന്നത് . കൂട്ടിന് മോഹലന്ലാലിനെ അഭിനയവഴക്കവും. ഇവ രണ്ടുമാകുമ്പോള് ‘ ദൃശ്യം ’ ദൃശ്യാനുഭവമാകുന്നു . മൂഡ് അനുസരിച്ചുള്ള ഗാനങ്ങളും പശ്ചാത്തലസംഗീതവും ചിത്രത്തെ ഉന്മേഷമുള്ളതാക്കുന്നു . മോഹന്ലാലിന്റെ അനായാസമായ അഭിനയ മികവ് തന്നെയാണ് കര്മ്മയോദ്ധായുടെ സവിശേഷതയെന്ന് നിസ്സംശയം പറയാം . ഡയറക്ഷന്റെ പോരായ്മ ചിത്രത്തെ ശരിക്കും ബാധിച്ചു . ഓര്ത്തെടുത്ത് പറയാവുന്ന സന്ദര്ഭങ്ങളും സംഭാഷണങ്ങളും ചിലതുണ്ട് ചിത്രത്തില് . മോഹന്ലാലിന്റെ അനായാസമായ അഭിനയ മികവ് തന്നെയാണ് കര്മ്മയോദ്ധായുടെ സവിശേഷതയെന്ന് നിസ്സംശയം പറയാം . അത്യാവശ്യത്തിന് പ്രേക്ഷക വെറുപ്പ് നേടാന് ആ വില്ലന് കഥാപാത്രത്തിന് കഴിഞ്ഞു എങ്കിൽ അത് മുരളി ശര്മ്മയുടെ വിജയം . ജനപ്രിയമായ ഒരു ടെലിവിഷന് കോമഡി ഷോയിലെ കഥാപാത്രങ്ങളെയൊക്കെ തന്റെ സിനിമയില് ഉള് പ്പെടുത്തിയിട്ടുണ്ട് സത്യന്അന്തിക്കാട് . സിനിമയുടെ ഒഴുക്കിനെ വല്ലാതെ ബാധിക്കുന്നുണ്ട് ഈ മാധ്യമങ്ങളുടെ ഇടപെടല് . വിരസത കൂടാതെ സിനിമ തീര്ക്കാന് സിദ്ധിഖിന് കഴിഞ്ഞു എങ്കിലും രണ്ടാം പകുതിയില് ഒട്ടൊന്ന് ദിശാബോധം നഷ്ടപ്പെട്ടുവോ എന്ന് തോന്നിപ്പിക്കുന്നുണ്ട് ഈ ജെന്റില് മാന് . ആദ്യപകുതിയുടെ ആവേശം ഇടയിലെവിടെയോ കെട്ടുപോവുന്നു . രസകരമായ കുറേയേറെ നിമിഷങ്ങളും ഹൃദയസ്പര്ശിയായ സംഭാഷണ ശകലങ്ങളും അസ്വാദ്യകരമായ നര്മ്മങ്ങളും ഇമ്പമാര്ന്ന ഗാനങ്ങളുമായി സിദ്ധിഖിന്റെ സ്ത്രീകളും മാന്യനായ മനുഷ്യനും വിഷ്ക്കാല ആഘോഷത്തിന് തിടമ്പേറ്റും . സിനിമയുടെ മിഴിവിന്റെ മികവിനു തെളിവായി സതീഷ് കുറുപ്പിന്റെ ഛായാഗ്രഹണം . കെ.ആർ.ഗൗരി ശങ്കര് വിദഗമായി എഡിറ്റിംഗ് നിര്വ്വഹിച്ചിരിക്കുന്നു . ഇമ്മാനുവലായി മമ്മൂട്ടി കാണികളെ മുഷിപ്പിച്ച് വിയര്പ്പിക്കുന്നുമില്ല . റഫീഖ് അഹമ്മദ് എഴുതി അഫ്സല്യൂസഫ് സംഗീതനിര്വ്വഹണം ചെയ്ത പാട്ടുകള് ഇമ്മാനുവല് എന്ന സിനിമയ്ക്ക് ആവശ്യമേയില്ലായിരുന്നു .സുനില് സുഖദയും സുകുമാരിയും മുക്തയും തങ്ങളുടെ ചെറുവേഷങ്ങള് മനോഹരമാക്കി . കോര്പ്പറേറ്റ് കപടതകളുടെ നടുവിലും നന്മ നിറഞ്ഞ ചിന്തകളോടെ ഒരുവനു വാഴാം എന്ന ശുഭസൂചകമായ ഒരു പാഠം ഇമ്മാനുവല് നമ്മെ ഓര്മ്മപ്പെടുത്തുന്നു എന്ന പോസീറ്റീവ് ചിന്തയോടെ നമുക്ക് പിരിയാം . പരിപൂര്ണ്ണത എന്ന അവസ്ഥയ്ക്ക് ഒരു ദൃശ്യ , ശ്രവ്യ ശുപമുണ്ടെങ്കില് അതാണ് ആമേന് ഒരു സിനിമയെ ഓരോ പ്രേക്ഷകനും വ്യത്യസ്തമായ ഭാവതലങ്ങളില് നിന്നാണ് കണ്ടെടുക്കുന്നത് . ചിലര്ക്ക സിനിമ വെറുമൊരു കാഴ്ചയാവാം . ആമേന് പരിപൂര്ണ്ണതയെ സ്പര്ശിക്കുന്നു എന്നുതന്നെ .

അഭിനന്ദനത്തിനുമേല് എത്ര അഭിനന്ദനങ്ങള് ചൊരിഞ്ഞാലും മതിയാകില്ല .

Table II. Result of Individual Category

Category	Polarity	Rating
Film	POSITIVE	GOOD
Direction/Script	NEGATIVE	BAD
Song/Acting	POSITIVE	EXCELLENT
Others	POSITIVE	EXCELLENT
Overall	POSITIVE	GOOD

Over all Result: POSITIVE

Over all Rating: GOOD

CONCLUSION AND FUTURE WORK

The sentiment analysis is the part of cognitive science that gives the artificial intelligence power to the machine. This work proposes a method of extracting the sentiments from the Malayalam film review. We have been implementing a hybrid approach for finding the sentiment from given sentence or document. This work would help to assign the rank and popularity of the new arrival film and also to the users for expressing their feelings after watching new films. Also help to find the rating and the score of the film. The polarity and rating of the individual categories like song, acting, direction, and script are also done. Presently, the sentiments can be extracted only from the movie reviews. This work can be enhanced for extracting the emotions from other areas like story, novels, product reviews and so on. The other machine learning approaches can also be used in this study.

REFERENCES

[1] Alekh Agarwal, and Pushpak Bhattacharyya. 2005 . 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).

[2] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics.

[3] Manurung, and Ruli, 2008 Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews, In Proceedings of the International Conference on Advanced Computational Intelligence and Its Applications

[4] H. Saggion and A. Funk. 2010. Interpreting sentiwordnet for opinion classification. In LREC.

[5] Turney, Peter D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics.

[6] Barbosa, Luciano, and Junlan Feng.2010. Robust sentiment detection on twitter from biased and noisy data, Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics.

[7] Celikyilmaz, Asli, Dilek Hakkani-Tur, and Junlan Feng.2010. Probabilistic model-based sentiment analysis of twitter messages", Spoken Language Technology Workshop (SLT).

[8] Mohandas, Neethu, Janardhanan P.S. Nair, and V. Govindaru.2012. Domain Specific Sentence Level Mood Extraction from Malayalam Text, Advances in Computing and Communications (ICACC), 2012 International Conference on. IEEE.