International Journal of Intelligent Computing and Technology
Vol. (1), No. (1), 2017, pp. 01-8

Comparison of Statistical Approaches for Sentiment Analysis - Malayalam Film Review

Deepu S.Nair[1], Jisha P. Jayan[2] and Dr. Elizabeth Sherly[3]
deepu.snair7@gmail.com, jishapjayan@gmail.com, sherly@iiitmk.ac.in

ABSTRACT

Sentiment Analysis is one of the most active research areas in NLP, which analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, and social networks. Sentiment analysis enables computers to automate the activities performed by human for making decisions based on the sentiment of opinions, which has wide applications in data mining, Web mining, and text mining. This work has been carried out to find the sentiments from Malayalam film review. This paper is a combinational approach comprising machine learning techniques with rule based approach. This work would help to assign the rank and popularity of the new arrival films and also to the users for expressing their feelings after watching new films.

Keywords: *Sentiment Analysis, Support Vector Machine, Conditional Random Field, TnT*

1. INTRODUCTION

Sentiment Analysis (SA) deals with analyzing emotions, feelings and the attitude of a speaker or a writer from a given piece of text. Sentiment analysis otherwise known as opinion mining refers to the application of natural language processing, computational linguistics, and text analytic to identify and extract subjective information in source materials. It involves capturing of user's behavior, likes and dislikes of an individual from the generated web content and also considered as thoughts, views and attitude of a person arising mainly based on the emotion instead of a reason. Sentiments are considered as the manifestation of our feelings and emotions. This field of computer science deals with analyzing and predicting the hidden information stored in the text. This hidden information provides valuable insights about user's intentions, taste and likeliness. SA focuses on categorizing the text at the level of subjective and objective nature. Subjectivity indicates that the text contains opinion content whereas Objectivity indicates that the text is without opinion content. Facts and opinions are two main types of textual information in the world . Facts are objective expressions about entities, events and their properties. Opinions are usually subjective expressions that describe people's sentiments, feelings toward entities, events and their properties.

Malayalam belongs to the Dravidian family, a large family of languages of South and Central India, and Sri Lanka. In quarter 9th century A.D Malayalam language had formed. Sanskrit and Prakrit was influenced to the formation of Malayalam language. A good part of Malayalam language vocabulary is also of Sanskrit origin.

Due to the ambiguity of Malayalam language, researchers face many problems as the same word can convey different meaning in different situation.

Sentiment Analysis of the Malayalam film review is not an easy task because it is highly depended on the words that are used for expressing the feelings of the public after watching. Since Malayalam is highly agglutinative language with rich morphology, it has wide range of fluctuated words that expresses the same meaning. Focus on the sentence-level sentiment extraction is significant because in most of the websites, user comments are just a single sentence. This is a very interesting topic. In this paper, we propose a machine learning with some rules for extracting the sentiments of users from their writings and a comparison of two machine learning approaches.

This paper is organized into different sections. First section dealt with the introduction part. The second section deals with the major works carried out in this area. The next section explains the proposed work. The fourth section includes the implementation and the result obtained. The fifth section concludes the paper.

3. **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 any 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.

3. METHODOLOGY

The ultimate aim of this work is to find the positive, negative or neutral opinions from user's writings. The sentiment analysis is handled at Sentence level . This work will give the polarity of sentence with the rating of overall film. It is very helpful to give ranking that helps to analyze the current status of that film. Each individuals can make the feedback about new film, they have the privilege for criticize and also write there suggestion about the film. This work is mainly carried out to extract the sentiment from users feedback. This work has been implemented as hybrid approach ie, it is the combined approach of machine learning techniques and rules. SVM and CRF has been used for machine learning.

3.1 ALGORITHM

Step 1 : Accept input

Step 2 : Training using SVM and CRF

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 film (Excellent,good,not bad,bad,not good,worst)

Step 6 : Results

Step 7 : Exit.

3.2 IMPLEMENTATION

In this work statistical approach has been applied. In this work mainly focuses the sentence level analysis with SVM and CRF. To find out the polarity and rating of the film reviews. The polarity indicates the sentence positive, negative or neutral.

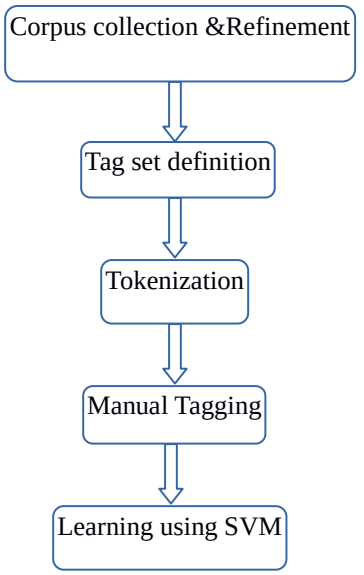

Fig 1. Training Phase

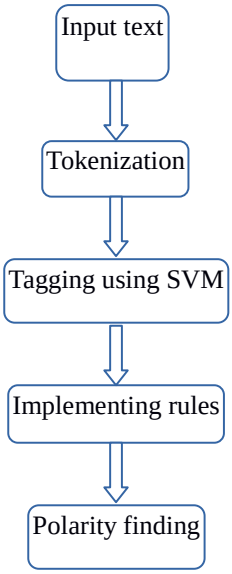

Fig 2. Testing Phase

3.3 TAGSET DEFINITION

Second face is tagset definition, it is an important task in this work, major problem in this field is nowhere found exact and standard sentiment tag set in Malayalam. The tagset definition is a very challenging task in this work, and also need very time consumption. Mainly 10 tags have been used for tagging the corpus.

Tagging process is not only depends on word but also must consider the context of that particular document. The same words have the different tags in different context. The tag is strictly depends on the context.

Table.1 Sentimental Tagset

Tags	Example	Description
CC_CCD	പക്ഷെ	Conjunction
DIR	സംവിധാനം ,സ്ക്രിപ്റ്റ്	Direction and Script
INEG	എന്നല്ല	Inverse Negative

INTF	വളരെ	Intensifier
NEG	മോശമാണ്	Negative
NEU	ഒരുക്കമായി	Neutral
POS	മികച്ചതാണ്	Positive
PUNC	.	Sentence Ending
SPCL	സിനിമയാണ്, ചിത്രത്തിനു	Films
TST	കഥാപാത്രങ്ങളിലൂടെ	Acting and Song

3.4 TRAINING WITH MACHINE LEARNING TOOLS

The machines have the capability to remember the previous experience. The refined corpus should be tagged with proper tag manually. The improper tagging causes the unjustifiable answers. The important matter in tagging is, the tagging must be done based on the context. After the manual tagging process, train the tagged corpus by appropriate machine learning technique such as CRF and SVM for training.

Both CRF and SVM have its own characteristics and its own model. When training the corpus, corresponding models are generated for further processing. The ultimate aim of the machine learning is to classify the input text with appropriate tag. Mainly 10 tag was defined for this work. Almost 30000 tokens are tagged properly. If a input text is gives to the engine, it will generate the tagged output with their generate the tagged output with their previous experience. The engine is checked the given sentence is already trained, if that exact sentence is trained earlier then it will gives accurate and proper tags on individual words. The tagging process strictly depends on context. If the input sentence is not trained previously, then the engine is checked the same or almost same context has trained earlier, then the engine is predicted the individual tags on each word. It may or may not be correct, because it's a prediction. If exact context never trained, it will return the almost similar tag with previous knowledge of similar context which is already trained. The experience of the engine is influence the correctness of the output. If engine have high experience the result is almost accurate. If the engine have low experience then the engine exhibit very poor performance. Major problem in machine learning is may be repeat the same word more than one times and the same word which convey the different meaning in different context. In different context the same word hold different tag this multi choice may confuse the engine. In this circumstantial situation the system will check whether the exact context is there, if found it ill return proper tag on each of them. In case that exact context is not there then it will return most probable tag to each tokens, it's depends on the experience of engine.

The working behind the classification is engine has check which tag is most proper in a context. It will cross checked with it's own previous experience. Format of the training set is very important in Machine Learning process. If any token violate the format engine has prevent the training process. So the corpus refinement and tagging is very crucial task.

4. RESULTS AND DISCUSSION

Sentence level sentiment analysis is treated in different way. In this experiment the ultimate intention is to find out the polarity of the input text. The review may contain the suggestion about film, song, direction, script and so on. With this information the system can predict the overall rating and score of new arrival film. Some peoples express their suggestion with in one sentence, some others are write detailed review about the film. The system will return the polarity of the input text, then calculate the rating and score of the film. An example is shown below.

Eg:1 Input: സിനിമയിൽ വളരെ രസകരമായ കുറേയേറെ നിമിഷങ്ങളും ഹൃദയസ്പർശിയായ സംഭാഷണ ശകലങ്ങളും അസ്വാദ്യകരമായ നർമ്മങ്ങളും ഇമ്പമാർന്ന ഗാനങ്ങളുമായി സിദ്ധിഖിന്റെ സ്ത്രീകളും മാന്യനായ മനുഷ്യനും വിഷുക്കാല ആഘോഷത്തിന് തിടമ്പേറ്റും .

Output : POSITIVE

Rating : EXCELLENT

Eg: 2 Input: എനിക്ക് പടം ഇഷ്ടമായില്ല

Output : NEGATIVE

Rating : BAD

SVM and CRF are most popular machine learning techniques having its own advantages and characteristics. Machine learning technique is mainly used for classifying the data in to appropriate class. For example the machine has been trained with a data “boy” is a “people” class. After the training phase, should test the performance and remembrance power of the machine with appropriate test data. If the test data is “boy” ,with the previous experience the machine will classify the data in to people class. This is the purpose of machine learning in classification process. If the machine is not classify in correct class, which means the machine does not learn properly or the efficiency of that particular learning approach is very poor. So the main challenge is to identify the appropriate machine learning approach for fulfill our goal.

The accuracy of machine learning is depends on the size of the corpus. Same word has different tag in different context. If the same word holds more than one tag then the machine may confuse which tag is proper for that word in current context. If the learning of machine is poor then it may be predict the improper tag because may be machine does not handle these type of context. So the prediction will correct if same context is train in different way, then the confusion may be decrease.

Here an attempt has made to compare the SVM and CRF for finding the accuracy and performance. In this experimental comparison, SVM and CRF are trained with various size of and 500 tokens are taken as test data. Precision, recall and F Score are calculated for the tags. The following table and are shown in the table below.

The F score can be interpreted as a weighted average of the precision and recall, where an F score reaches its best value at 1 and worst score at 0.

Table 2 : Precision and Recall Of SVM and CRF

Tag	CRF			SVM		
	Precision	Recall	F-Score	Precision	Recall	F-Score
CC_CCD	0.500	1.00	0.666	1.00	1.00	1.00
DIR	1.00	0.529	0.692	0.705	1.00	0.827
INEG	0.600	0.461	0.521	0.769	0.909	0.833
INTF	1.00	0.800	0.888	0.866	1.00	0.928
NEG	0.785	0.407	0.536	0.481	0.866	0.619
NEU	0.779	0.976	0.867	0.996	0.836	0.909
POS	0.756	0.607	0.673	0.568	0.906	0.698
RD_PUNC	1.00	1.00	1.00	1.00	1.00	1.00
SPCL	1.00	0.533	0.695	0.733	1.00	0.846
TST	1.00	0.210	0.347	0.947	1.00	0.972

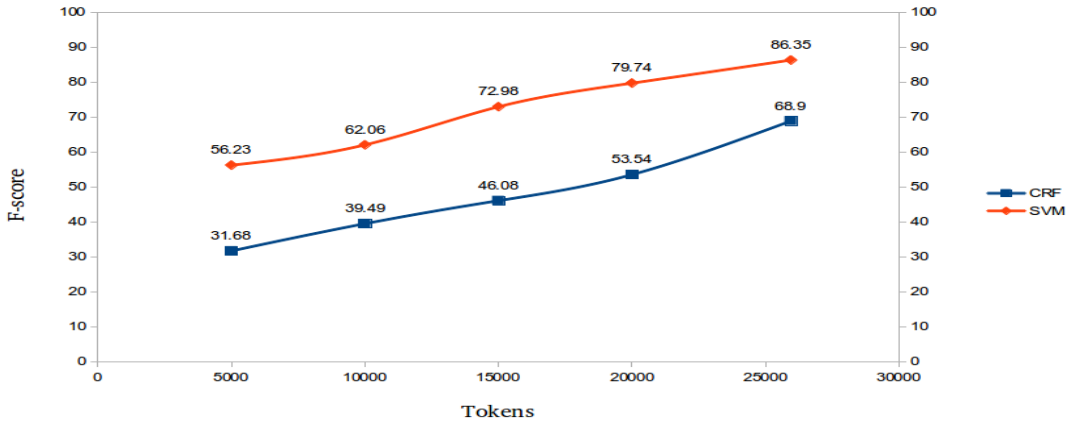

Fig 3 : F-score

5. CONCLUSION AND FUTURE ENHANCEMENT

The sentiment analysis can be considered as a part of cognitive science. This work proposes a method of extracting the sentiment from Malayalam film review. The three commonly used methods for sentiment analysis of a given text is machine learning method, semantic orientation method and rule based approach. In this work we have compared two statistical methods namely SVM and CRF for analyzing the sentiments of Malayalam movie reviews. From this study, we have found that SVM outperforms the CRF. 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. This work can be further extended to find the rating and score of the overall film and also individual categories like song, acting, direction, script and so on. 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.

ACKNOWLEDGEMENTS

Through this acknowledgment, I express my sincere gratitude to all the peoples from Virtual Resource Center for Language Computer (VRCLC) who have been associated with this assignment and have helped me with it and made it a worthwhile experience. I would like to express my deep gratitude to Professor Dr. Elizabeth Sherly , my research supervisor, for her patient guidance, enthusiastic encouragement and useful critiques of this research work.

REFERENCES

1. Alekh Agarwal & Pushpak Bhattacharyya. “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), 2005.

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

3. Manurung & Ruli, “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, 2008.

4. H. Saggion & A. Funk. Interpreting sentiwordnet for opinion classification. In LREC, 2010.

5. Turney & Peter D, “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, 2002.

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

7. Celikyilmaz, Asli, Dilek Hakkani-Tur & Junlan Feng, “Probabilistic model-based sentiment analysis of twitter messages”, Spoken Language Technology Workshop (SLT), 2010.