ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(45), DOI: 10.17485/ijst/2016/v9i45/106482, December 2016* ISSN (Online) : 0974-5645

Predicting the Sentimental Reviews in Tamil Movie using Machine Learning Algorithms

Shriya Se*, R. Vinayakumar, M. Anand Kumar and K. P. Soman

Centre for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore Amrita Vishwa Vidyapetham, Amrita University, Coimbatore – 641112, Tamil Nadu, India; shriyaseshadrik.r@gmail.com, r_vinayakumar@cb.amrita.edu, m_anandkumar@cb.amrita.edu, kp_soman@amrita.edu

Abstract

Objective: This paper aims at classifying the Tamil movie reviews as positive and negative using supervised machine learning algorithms. **Methods/Analysis:** A novel machine learning approaches are needed for analyzing the Social media text where the data are increasing exponentially. Here, in this work, Machine learning algorithms such as SVM, Maxent classifier, Decision tree and Naive Bayes are used for classifying Tamil movie reviews into positive and negative. Features are also extracted from TamilSentiwordnet. **Findings:** The dataset for this work has been prepared. SVM algorithm performs well in classifying the Tamil movie reviews when compared with other machine learning algorithms. Both cross validation and accuracy of the algorithm shows that SVM performs well. Other than SVM, Decision tree perform well in classifying the Tamil reviews. **Novelty/Improvement:** SVM gives an accuracy of 75.9% for classifying Tamil movie reviews which is a good milestone in the research field of Tamil language.

Keywords: Machine Learning, Maxent Classifier, Sentimental Analysis, Support Vector Machine, Tamil Language, TamilSentiwordnet

1. Introduction

Sentiment analysis or opinion mining has drawn attention in recent years. Nowadays this is an important and blossoming field in Natural Language Processing (NLP) which generally concern with the refinement and cataloging the opinion from the narrative. Primarily, the opinions are categorized into positive and negative which are helpful in many fields[1]. Opinions are important for many day-to-day activities either it may be for getting a product or it may be about a movie. Mining the opinion is not an easy task as each individual have different opinion about the product or about the movie. Opinions are private expression, which are not observed by others directly[2]. But, these opinions can be observed from the statement written by the individuals. Sentiment analysis is prevailing for English language but for Indian languages it's a rare thing[3].

Tamil being a Dravidian language which has no standard annotated corpora for sentiment analysis apart from SAIL data. But the data is available only for Twitter data. Hence forth, the work for Tamil language in the field of Natural Language Processing is very limited. In Tamil language, most of the grammatical functions are embedded into the words in the form of inflection. Many parser and tagger are already available for English language. Being low on resource, to work on Tamil language has been difficult. Many web pages have started to provide information in Tamil.

Tamil cinema industries also known as Kollywood has drawn attention in recent years. This is due to the fact that they started producing world class films in these years. So many non-Tamil speakers started watching Tamil film. So, reviewing Tamil movies will be an appreciating work in the field of Tamil language processing. Internet plays a censorious role in gathering the reviews.

* *Author for correspondence*

A frame work for sentiment analysis in Hindi used Hindisentiwordnet to observe the sentiment related to a document and also the polarity of the word using Hindisentiwordnet and finally aggregating the polarity to find whether the document is positive, negative or neutral[4]. Sentiment analysis using lexicon acquisition uses distributional thesaurus, sentence level co-occurrences and also they expanded the existing sentiment lexicon for Indian languages for analyzing the polarity in the tweets which are in Indian languages[5]. Cross lingual sentiment analysis using linked wordnetsin which opinion expressed in one language is classified using the trained corpus which is written in another language. By which they got an accuracy of 72% for Hindi and 84% for Marathi languages[6]. In Sentiment analysis based on negation and discourse relation the influence of negation and discourse roles are scrutinized for Hindi sentiment analysis which gives an accuracy of 80.21%[7]. Sentiment analysis using neural network used to find the sentiment analysis in the document level a special hand crafted features are extracted and is used to improve the accuracy of the system in which the accuracy of the system is about 83.88%[8]. SVM have been successfully used in Tamil language processing such as Statistical Machine translation[9], Morphological analysis[10], Morpheme Extraction[11], Word Sense disambiguation[12], Entity extraction[13,14] and Sentimental Analysis[15–17]. The flow of the work is as follows: The next section deals about mathematical ideas which are implemented in the paper. The methodology is explained in Section 3 where the proposed system and feature extraction are explained in detail. Section 4 is about the experimental study of the proposed system. Finally the conclusion of the paper is discussed in last section.

2. Mathematical Background

2.1 Support Vector Machine

Support Vector Machine is classification method which is mainly in data classification and function approximation. Support Vector Machine constructs a hyper plane which separate two different classes. While separating the classes, the SVM tries to gain highest separation in between two classes. SVM prediction projects the test data where the train data are already projected. The SVM classifier classifies the data into positive and negative based upon the weight of the test data[18]. Linear SVM are designed for binary classification and it is given below. The training data is given by $X_k = \{x_1, x_2, \ldots\ldots, x_m\}$ and the label corresponding to each class is given by d_k where, $k = \{1,2, \ldots\ldots m\}, X_k \in R^D$, $d_k \in \{+1, -1\}$ and the formulation is given by:

$$\frac{\min 1}{2} w^T w \tag{1}$$

Subjected to,

$$d_k[w^T x_k - \gamma] \geq 1 \tag{2}$$

Where w is the weight vector. The SVM classifier's performance depends on choosing the γ values[19].

2.2 Maxent Classifier

The entropy classifier is a discriminative classifier most of the time used in traditional language processing, speech and data retrieval. This works well for text classification problems such as sentiment analysis. The Maxent classifier works on the principle of Maximum Entropy. The main goal behind highest entropy one can prefer most uniform model which also fulfill the given constrain[20]. In general, Maxent is used estimate the distributional probability. This paper is meant for classification problem, thus we limit to learning the conditional probability only from the training data. The learned conditional probability must have the following property.

$$\frac{1}{|X|}\sum_{x \in X} f_i(x, c(x)) = \sum_x P(x) \sum_c p(c \mid x) f_i(x, c) \tag{3}$$

Where D is the set of features in training data, P (d) is an unknown document distribution and we are interested in modeling it. Thus, we use our training data without unknown document as an observation. Then the above equation will be:

$$\frac{1}{|X|}\sum_{x \in X} f_i(x, c(x)) = \frac{1}{|X|} \sum_{x \in X} \sum_c P(c \mid x) f_i(x, c) \tag{4}$$

Where $f_i(x, c)$ is the feature of the document and the classes which is a real valued function[21].

2.3 Decision Tree

Decision tree is a popular and powerful tool for prediction and classification problem. Decision tree usually create a model tree using the available data and uses the tree for classifying the future data. In Decision tree the non-leaf nodes are decision node and leaf nodes contains the class

name[22]. Generally information gain which is called as Statistical property is used here to decide which attribute goes to the decision node and which are not. This gain measures how well the targeted classes are separated from training set from the given attribute[23]. Entropy measure the gain from the given attribute and it is given as:

$$Entropy(S) = \sum -p(I) \log 2p(I) \quad (5)$$

2.4 Naive Bayes

Naive Bayes is one of the best algorithm for classifying documents[24]. It has been widely used in the field of recovering of information and freshly it has been used in machine learning researches[25]. In recent years, Bernoulli model and multinomial model drawn attention[26]. The multinomial template represents the integer feature to represent document whereas in Bernoulli model vector of binary feature are attained from the document[27]. The Naive Bayes is arithmetically represented as:

$$P(c \mid d) = p(c) \prod_{1 \le k \le n_d \le n} p(t_k \mid c) \quad (6)$$

Generally, Laplacian smoothing (a small correction value) is included in order to normalize the error in Naïve Bayes[28,29].

3. Methodology

3.1 Proposed System

The design of the proposed system is demonstrated in the Figure 1. The reviews of different movies are collected from the web pages, structured and analyzed and manually tagged into positive and negative. These classified reviews are given as a training data for the machine. The attributes are extracted from the Sentiwordnet and also from the training data. A model file is created using the Sentiwordnet and training data. For testing the system, movie reviews are further collected and structured. The model file along with the testing data is given as the input to the system. The Support Vector Machine (SVM), Maximum Entropy classifier (Maxent), Decision tree and Naive Bayes are used to categorize the reviews. The mathematical background is explained in detail in the above section where, -1 corresponds to negative classes and +1 corresponds to the positive classes.

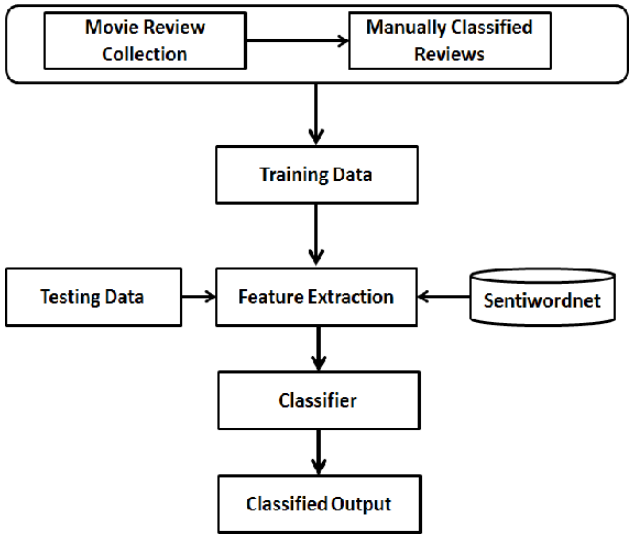

Figure 1. Flow diagram of the proposed system.

3.2 Feature Extraction

In this task, crucial process is given to feature extraction as the correctness of the classifier depends mainly on the extirpated feature. The context words of training data are taken as feature. The punctuations and apostrophe are considered as feature. TamilSentiwordnet is added as an extra feature for improving the accuracy of the system. Sentiwordnet has set of words which are classified into positive and negative. The words also contain the POS tag of the particular word. The terms in the Tamil Sentiwordnet are considered as a word feature. Along with the training data and Sentiwordnet, the system creates a model file which is further used for testing.

4. Experimental Analysis

4.1 About Dataset

The data are collected from different sources of webpages. These data are in unstructured format where the review contains some special characters, other languages fonts, some pictures of the movies, etc. This unstructured data are structured by removing the unwanted characters and other language fonts. After structuring, the data are manually tagged based on the comments, the reviews it contains. These manually tagged reviews are thoroughly reviewed by two different linguistic experts. The reviews are tagged

into positive and negative which are saved in two different folders. Each file in the folder contains the review of a particular movie which is written in Tamil language. Up to the above step the preprocessing of the data gets over. Each folder has 267 unique text. The positive and negative average word count is about 6110 and 8646 respectively for each file in the folder. Table 1 describes about the reviews that are considered for training and testing the proposed system.

Table 1. Detailed description of reviews that are collected and their count

Reviews Description	File Count	Reviews for testing	Reviews for training	Average Word Count
Negative Reviews	267	210	57	8646
Positive Reviews	267	210	57	6110

4.2 Experimental Results and Analysis

The Tamil movies which are collected from different sources are structured and manually tagged. Entirely, 534 reviews about different Tamil movies are collected and are manually tagged. These reviews contains 267 unique positive reviews and negative reviews respectively. Table 1 gives a detail description of the reviews. For training, the data are given to the system and the results are decipated in the table. Along with the training data Sentiwordnet are given as extra feature to the system. The words of Sentiwordnet contain tagging for each word and also cointain the POS tag of those words. The words from Sentiwordnet is taken as feature for classifying the system. A model file is created using the training data and also using the features. For testing, the data along with the model files are used. The correctness of the classifier is examined using F-score measure. F-score can be calculated using percision and recall.

$$F-Score = 2\frac{pr}{p+r} \tag{7}$$

$$where \quad p = \frac{tp}{tp+tr}, \; r = \frac{tp}{tp+fn}$$

Presicion is the ratio of true positive (tp) to all predicted positive (tp+fp) and recall is the ratio of true positive (tp) to all actual positive (tp+fn). The accuracy and cross validation accuracy of each classifier using Sentiwordnet as feature and without using Sentiwordnet as feature are given in theTable 2 and Table 3. It is noticed from the table, for different classifiers different accuracy are obtained. Above all the classifiers Support Vector Machine (SVM) perform well for classifing the data. The comparisons of accuracies using Sentiwordnet and without using Sentiwordnet are described in the Figure 2.

Table 2. Accuracy and cross validation accuracy of each classifier without using Sentiwordnet

Classifiers	Cross Validation of classifiers(%)	Accuracy of classifiers (%)
Naïve Bayes	50.41	61.79
Maxent	52.10	59.55
Decision tree	56.17	64.04
SVM	56.81	71.91

Table 3. Accuracy and cross validation accuracy of each classifier using Sentiwordnet

Classifiers	Cross Validation of classifiers(%)	Accuracy of classifiers (%)
Naïve Bayes	55.28	66.17
Maxent	56.95	64.04
Decision tree	56.39	66.29
SVM	57.92	75.96

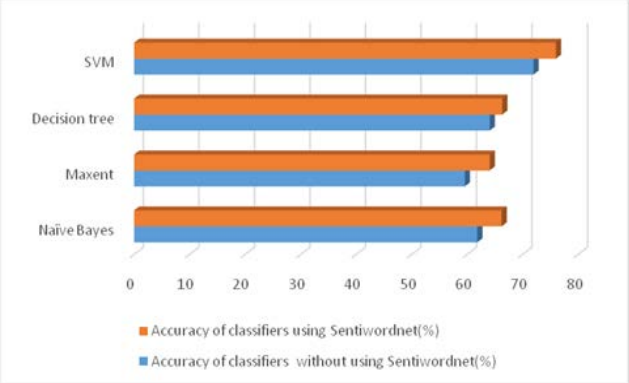

Figure 2. Comparisons of accuracies using Sentiwordnet and without using Sentiwordnet.

5. Conclusion and Future Work

We presented a proposal for classifying the Tamil movie reviews using supervised algorithms, namely SVM, Maxent classifier, Decision tree and Naive Bayes. Feature are extracted and are given to the system for better classification and to improve the accuracy of the system. The TamilSentiwordnet (word) features are influencing the file level sentiment analysis in a greater level. The accuracy of different classifiers with word feature are decipated in the table. The SVM classifies better than other classifiers and the results are decipated in the above table. Collecting the movie reviews is one of the major part of this work. As there are only few work for Tamil language,

this will be a great work towards the improvement of Tamil language for sentiment analysis. For future work, fine grained tagging can be done where, these reviews can be fine grained into very positive, positive, very negative, negative and neutral. Unsupervised way of implementing data can also be done as forthcoming work.

6. References

1. Hutto CJ, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International AAAI Conference on Weblogs and Social Media; 2014 May 16.
2. Fink CR, Chou DS, Kopecky JJ, Llorens AJ. Coarse- and fine-grained sentiment analysis of social media text. Johns Hopkins APL Technical Digest. 2011 Jan; 30(1):22–30.
3. Selvan A, Anand Kumar M, Soman, KP. Sentiment analysis of Tamil movie reviews via feature frequency count. IEEE International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS 15); 2015.
4. Pandey P, Govilkar S. A framework for sentiment analysis in Hindi using HSWN. International Journal of Computer Applications. 2015 Jan; 119(19):23–6.
5. Kumar A, Kohail S, Ekbal A, Biemann C. IIT-TUDA: System for sentiment analysis in Indian languages using lexical acquisition. Mining Intelligence and Knowledge Exploration. 2015 Dec; 9468:684–93.
6. Balamurali AR. Cross-lingual sentiment analysis for Indian languages using linked wordnets. CiteSeer. 2012.
7. Mittal N, Agarwal B, Chouhan G, Bania N, Pareek P. Sentiment analysis of Hindi review based on negation and discourse relation. Proceedings of International Joint Conference on Natural Language Processing; 2013. p. 45–50.
8. Timmaraju A, Khanna V. Sentiment analysis on movie reviews using recursive and recurrent neural network architectures. Semantic Scholar; 2015.
9. Anand Kumar M, Dhanalakshmi V, Soman KP, Rajendran S. Factored statistical machine translation system for English to Tamil language. Pertanika Journal of Social Science and Humanities. 2014; 22(4):1045–61.
10. Anand Kumar M, Dhanalakshmi V. A novel approach to morphological analysis for Tamil language. Germany: University of Koeln Koln; 2009 Oct.
11. Anand Kumar M, Soman KP. AMRITA-CEN@FIRE-2014: Morpheme extraction and lemmatization for Tamil using machine learning. ACM International Conference Proceeding Series; 2014 Dec. p. 112–20.
12. Anand Kumar M, Rajendran, S, Soman KP. Tamil word sense disambiguation using Support Vector Machines with rich features. International Journal of Applied Engineering Research. 2014; 9(20):7609–20.
13. Abinaya, N, Anand Kumar M, Soman KP. Randomized kernel approach for named entity recognition in Tamil. Indian Journal of Science and Technology. 2015; 8(24).
14. Abinaya N, John N, Barathi Ganesh HB, Anand Kumar M, Soman KP. AMRITA-CEN@FIRE-2014: Named entity recognition for Indian languages using rich features. Proceedings of ACM International Conference Series; 2014 Dec. p. 103–11.
15. Patra BG, Das D, Das A, Prasath R. Shared task on sentiment analysis in Indian languages (sail) tweets - An overview. Mining Intelligence and Knowledge Exploration; 2015 Dec. p. 650–5.
16. Shriya S, Vinayakumar R, Kumar MA, Soman KP. AMRITA-CEN@ SAIL2015: Sentiment analysis in Indian Languages. Mining Intelligence and Knowledge Exploration; 2015 Dec. p. 703–10.
17. Kumar SS, Premjith B, Kumar MA, Soman KP. AMRITA_CEN-NLP@ SAIL2015: Sentiment analysis in Indian Language using regularized least square approach with randomized feature learning. Mining Intelligence and Knowledge Exploration; 2015 Dec. p. 671–83.
18. Rueping S. SVM classifier estimation from group probabilities. Proceedings of the 27th International Conference on Machine Learning (ICML-10); 2010. p. 911–8.
19. Rosset S, Tibshirani R, Zhu J, Hastie TJ. The entire regularization path for the Support Vector Machine. Advances in Neural Information Processing Systems; 2004. p. 561–8.
20. El-Halees A. Arabic text classification using maximum entropy. The Islamic University Journal (Series of Natural Studies and Engineering). 2007; 15(1):157–67.
21. Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification. IJCAI-99 Workshop on Machine Learning for Information Filtering. 1999 Aug; 1:61–7.
22. Quinlan JR. Induction of Decision trees. Machine Learning. 1986 Mar; 1(1):81–106.
23. Yuxun L, Niuniu X. Improved ID3 algorithm. 2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT). 2010 Jul; 8:465–8.
24. Ting SL, Ip WH, Tsang AHC. Is Naive Bayes a good classifier for document classification? International Journal of Software Engineering and Its Applications. 2011 Jul; 5(3):37–46.
25. Panda M, Abraham A, Patra MR. Discriminative multinomial Naive Bayes for network intrusion detection. 2010 Sixth International Conference on Information Assurance and Security (IAS); 2010 Aug. p. 5–10.
26. Juan A, Ney H. Reversing and smoothing the multinomial Naive Bayes text classifier. PRIS; 2002 Apr. p. 200–12.
27. Lewis DD. Naive Bayes at forty: The independence assumption in information retrieval. Machine Learning: ECML-98; 1998 Apr. p. 4–15.
28. Amor NB, Benferhat S, Elouedi Z. Naive Bayes vs. Decision trees in intrusion detection systems. Proceedings of the 2004 ACM symposium on Applied Computing; 2004 Mar. p. 420–4.
29. Das A, Gamback B. Sentimantics: Conceptual spaces for lexical sentiment polarity representation with contextuality. Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. Association for Computational Linguistics; 2012 Jul. p. 38–46.