ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 6, June 2016

Sentiment Extraction for Tamil Political Reviews

D. Anish[1], V. Sumathy[2],
Student, Department of CSE, Rajalakshmi Engineering College, Chennai, India[1]
Assistant Professor, Department of CSE, Rajalakshmi Engineering College, Chennai, India[2]

Abstract: This paper proposes a Support Vector Machine (SVM) based approach for sentiment analysis from Tamil political reviews. Sentiment analysis, also called opinion mining, is the field of study in Natural Language Processing (NLP) that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. 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. Political review analysis using social media data, has witnessed a blooming interest over the past decade to understand the public opinion and trend especially during election time. Since Tamil is highly agglutinative language with rich morphology, it has many challenges such as noise, sarcasm, negation, etc. Using Support Vector Machine (SVM), the data is analyzed and classified on basis of positive, negative and neutral opinions. The weighted average for all sentiments (positive, negative and neutral) in the text is tabulated, and a conclusion is made based on the statistics.

Keywords*:* Sentiment Analysis, Opinion Mining, Tokenization, Sentiment Extraction, Part of Speech

I. INTRODUCTION

Web content is an important source of information to mine the sentiment/opinion of people on different products and services. The new era has resulted in the generation of large amount of user content on web. The developing advancement in technology has led to wide spread use of blogs, forums, e-news, reviews and the social networking platforms such as Facebook, Twitter. These social networking platforms has exponentially increased the amount of information generated on daily basis. Thus mining the data and identifying user sentiments, wishes, likes and dislikes is an important task that has attracted the focus of research community from last decade. Sentiment analysis is a task of greater importance among business and social media professionals because it can be used in aid to decision making, recommendation systems, product analysis, etc.

Sentiment Analysis involves capturing of user's behavior, likes and dislikes of an individual from the generated web content. 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 provide valuable insights about user's intentions, taste and likeliness. Sentiment Analysis focus on categorizing the text at the level of subjective and objective nature. Subjectivity indicates that the text contains opinion content because it is based on personal opinions or feelings rather than on external facts or evidence whereas, Objectivity indicates that the text is without opinion content because it describes something without being influenced by personal emotions or prejudices. Natural languages are highly unstructured in nature and thus understanding the meaning of a sentence automatically by a machine is a challenging task. But the applications of sentiment analysis are increasing every day. Manual training to solve this problem can achieve accuracy only to a satisfactory level and hence an automated system is needed to solve this problem.

Sentiment extraction is done either at document level, sentence level or aspect level. Both document level and sentence level classification methods are usually based on the classification of opinion words or phrases. This paper focus on sentence level sentiment extraction, because in most of the websites user comments are just a single

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 6, June 2016

sentence. Sentiment classification techniques can be roughly divided into machine learning (ML) approach, lexicon-based approach and hybrid approach[1]. The text classification methods using ML approach can be divided into supervised and unsupervised learning methods. The supervised methods make use of a large number of labeled training documents. The unsupervised methods are used when it is difficult to find these labeled training documents. Various ML algorithms for text classification are Naive Bayes Classifier, Maximum Entropy Classifier, Neural Networks, Support Vector Machine, etc. The lexicon-based approach depends on finding the opinion lexicon which is used to analyze the text. The hybrid Approach combines both approaches and is common with sentiment lexicons being the key role in most of the methods.

Tamil belongs to the Dravidian family, a large family of languages of South and Central India, and Sri Lanka. Tamil is an official language of two countries, Singapore and Sri Lanka, and also has official status in the Indian state of Tamil Nadu and the Indian Union Territory of Puducherry. Sentiment Analysis for Tamil political review is not an easy task because, it solely depends on the words that are used for expressing the feelings of the public over political matters. Since Tamil is highly agglutinative language with rich morphology, it has wide range of fluctuated words that express the same meaning. In this paper, we propose a Support Vector Machine (SVM) based approach for classifying the sentiments from the user comments.

Most of the research work is done using machine learning (ML) approaches. One of the first noted work in this area was done by Peter D. Turney for English[2]. In this paper a simple unsupervised learning algorithm for classifying reviews as recommended (thumbs up) or not recommended (thumbs down) was presented. Sentiment extraction for automobiles to movie reviews at document level was implemented using semantic orientation. The algorithm achieved 74% of accuracy.

Another work on the sentiment extraction of movie reviews was done by Pang. The aim of this work was to find the best way to classify the sentiment from the text, either machine learning techniques or human-produced baseline. Three different machine learning techniques were analyzed Maximum Entropy, Support Vector Machine and Naive Bayes. Experiments with different variations of n-gram approach like unigrams presence, unigrams with frequency, unigrams+ bigrams, bigrams, unigrams+ POS, adjectives, most frequent unigrams, and unigrams+ positions were conducted. They concluded that machine learning technique are better when compared to human-produced baseline. The paper also said that the Naive Bayes approach tend to do the worst while SVMs is the best machine learning technique[3].

Chenhao Tan, Lillian Lee and Jie Tang, have demonstrated that user-level sentiment analysis can be significantly improved by incorporating link information from a social network. This paper explores into the social network structures to help sentiment analysis, represents an interesting research direction in social network mining. Markov Random Field was used, although the sparsity of the graphs may be an issue in applying such an approach[4].

Sandeep Chandran, Bhadran V K, Santhosh George and Manoj Kumar P, have developed a Feature-based domain independent approach for Document-level sentiment extraction for Malayalam. In order to achieve domain independence, they have increased the weight of domain free features and reduce the weight of domain specific features[5].

Deepu S Nair, Jisha P Jayan, Rajeev R and Elizabeth Sherly, have developed a Rule-based approach to perform Sentence-level sentiment classification for Malayalam movie reviews. Negation rule is applied for improving the sentiment classification. This system has achieved 85% of accuracy for movie reviews[6]. M. Ravichandran, G. Kulanthaivel and T. Chellatamilan, focused on sentence-level opinion classification using the unsupervised learning algorithm named bigram item response theory (BIRT) for tweets[7].

Abd Samad Hasan Basaria, Burairah Hussina, I Gede Pramudya Anantaa and Junta Zeniarja, attempts to use the messages of twitter to review a movie by using opinion mining. Support Vector Machine (SVM) binary classifier is used to identify positive and negative reviews and hybrid Partical Swarm Optimization (PSO) is used to select the best parameter for dual optimization and achieved 77% of accuracy[8].

Subjectivity and Sentiment Analysis (SSA) for Arabic social media has been developed by Muhammad Abdul-Mageed, Sandra Kubler and Mona Diab. A two-stage classification approach is adopted for this work. In the first stage (i.e., Subjectivity), a binary classifier is built to separate objective from subjective cases. For the second stage (i.e., Sentiment) binary classification is applied, that distinguishes positive from negative cases[9].

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 6, June 2016

Gurunath H Naragund, Santhosh Kumar K L and Jharna Majumdar, have proposed a Decision Making and Analysis (DMA) System which has been integrated into the Robotic Head, where the robot will be more interactive and useful in the real time. Various text pre-processing tasks such as stop-word removal, tokenization, POS tagging are also done. Using a sentiment dictionary with positive and negative opinion words are used for identifying the sentiment in each sentence[10]. Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou, have designed and developed a movie-rating and review-summarization system in a mobile environment using latent semantic analysis (LSA) to identify product features. The condensed descriptions of movie reviews are generated from the feature-based summarization[11].

This paper is organized into five different sections. The first section deals with the introduction and the major works carried out in this area. The second section explains the proposed work. The third section describes the processes involved for implementation. The fourth section shows the results obtained. The fifth section concludes the paper.

II. Proposed Work

In this work, Support Vector Machine based approach is implemented to analyze user generated sentences. The political reviews are collected from social media websites such as Facebook and Twitter for experimentation.

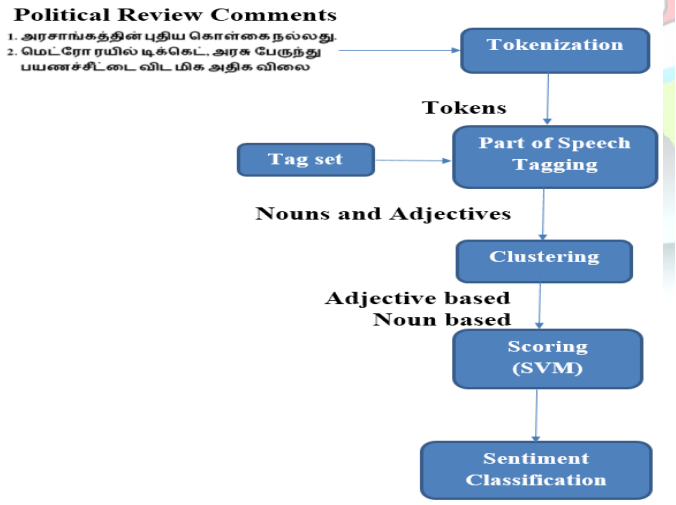

Fig. 1. Block Diagram

Sentence level opinion extraction of direct and comparative reviews, on political subjects in Tamil text, is done by pre-processing the input (political review comments) and by applying ML algorithms. Initially, tokenization process is done to split sentences to words. Part of Speech tagger is used to tag the words in each sentence according to Tamil grammar. Clustering algorithm is applied to perform noun based grouping of sentences. Support vector machine classifier is used to analyze and classify text on the criteria (positive, negative and neutral). Finally, the sentiment classification for the given input sentences is obtained.

III. Implementation

Textual data is only a textual interpretation or block of characters at the beginning. For information retrieval we require the words of the data set. The input sentences are given as a string and may contain special characters. These special characters are removed and the text is converted to a standard format.

Tokenization is a pre-processing step, carried out to provide proper input to the next module. The comments/sentences are split into words using the white space in the text.

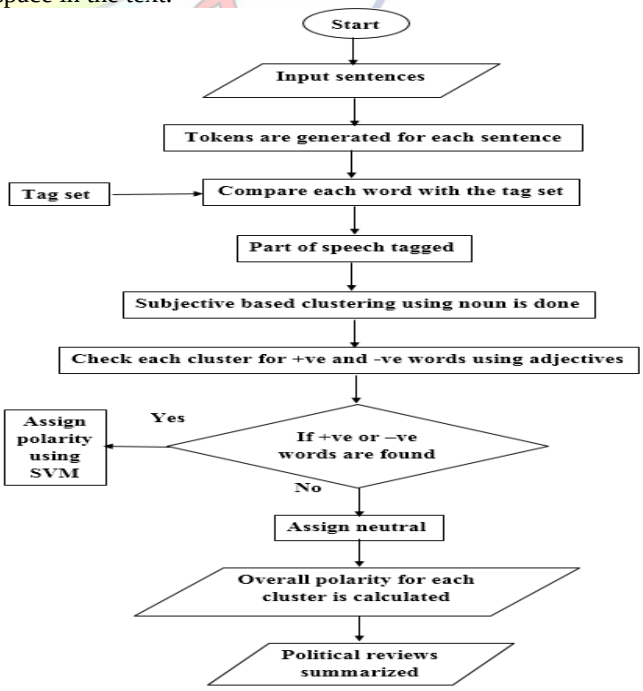

Fig. 2. Work Flow Diagram

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 6, June 2016

A. POS Tagging

POS tagging means assigning grammatical classes i.e. appropriate parts of speech tags to each word in a natural language sentence. Assigning a POS tag to each word of an un-annotated text by hand is very time consuming, which results in the existence of various approaches to automate the job. So POS tagging is a technique to automate the tagging process of lexical categories.

The process takes a word or a sentence as input, assigns a POS tag to the word or to each word in the sentence, and produces the tagged text as output. POS tagging is implemented using SVM algorithm and tag set. SVM is easy to train and flexible, hence provides higher accuracy. Own tag set was developed for training and testing the POS-tagger generators.

UNTAGGED CORPUS

வாழ்நாள் தண்டனை கொடுக்கபட்ட நபரின் தண்டனையை குறைக்க முடியும்.

TAGGED CORPUS

வாழ்நாள் <N>
தண்டனை<N>
கொடுக்கபட்ட<V>
நபரின்<N>
தண்டனையை<N>
குறைக்க<AJ>
முடியும்<AJ>.
.<DOT>

Fig. 3. Example of Tagged sentence

The present tag set consists 8 tags. A corpus size of one hundred words was used for training and testing the accuracy of the tagger generators.

B. Clustering

The tagged sentences from the POS tagger is obtained and is given as input for noun based clustering. Nouns are the words which denote a person, place, thing, time, etc. Based on the subjects taken for political review analysis, a set of nouns are predefined for subjective clustering. SVM algorithm yields best results for supervised learning problems, hence SVM based clustering algorithm is applied to obtain separate clusters for each subject.

C. Scoring

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in the Figure 4 of how positives and negatives are classified using the decision plane. Linear classifier is used which separates a set of objects into their respective groups (+1 and -1 in this case) with a line.

The noun based clusters are obtained and SVM based sentiment classification is performed. SVM is an optimal classifier in the sense that, for a given training data, it learns a classification hyperplane in the feature space, which has the maximal distance (or margin) to all training examples (except a small number of outliers).

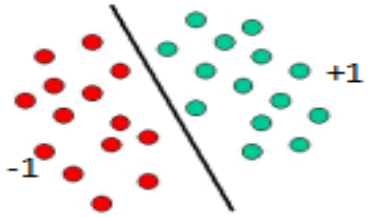

Figure. 4. SVM Classifier

The training data contains words that express any type of sentiment. A polarity is assigned to positive, negative and neutral sentiments for polarity evaluation. The sentiments in a sentence are usually the adjectives of the sentence, hence the SVM classifier uses the adjectives to identify the type of sentiment in each sentence.

D. Review Summarization

The SVM classifier, classifies the sentiments for each sentence and assigns a polarity for each cluster. The overall summarization based on various clusters and its polarity is tabulated and a graph is drawn to show the statistics.

IV. RESULTS

Tamil Unicode corpus was used for testing. The results obtained on implementation of the proposed system are as follows.

Example:

Input 1- *அரசாங்கத்தின் புதிய கொள்கை நல்லது.*
(Arasangathin pudiya kozhgai naladhu)
Output- *நல்லது* (naladhu) → **Positive**

Input 2- ***தமிழ்நாடு தேர்தல் மே முதல் வாரத்தில் நடைபெறும்.***
(Tamilnatu tertal me mutal vaarattil nataiperum)
Output- **Neutral**

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 6, June 2016

Input 3- பூகம்பம் பாதிக்கப்பட்ட பகுதிகளில் மீட்பு பணிகள் மிகவும் மோசமாக இருந்தது.
(Boogambam paathikapatta pagudhigalil meetpu panigal mosamaga erundhadhu)
Output- மோசமாக *(mosamaga)* → **Negative**

Table I. Result Analysis

Sl. No	Comments	Result
1.	□□□□ □□□□□□□ □□□□ □□□□□□□ □□□□□□□ □□□□□□□□□□ □□□□□□□□.	Correct
2.	□□□□□ □□□□□□□□□□□□□□ □□□□ □□□□□□□□□□□ □□□□□□□□□□□ □□□□ □□□□□ □□று□□.	Correct
3.	□□□□ □□□□□□□ □□□□□□ □□□□□□□□□□ □□□□□□□ □□□□□□.	Correct
4.	□□□□□□□□□□□□□ □□□□□ □□□□□□□ □□□□□□.	Correct
5.	□□□□ □□□□□□□ □□□□□□□ □□□□□□□□□□ □□□□□□ □□□□□.	Wrong
6.	□□□□ □□□□□□□ □□□□□□□ □□□□□□□□□□ □□□□□.	Correct
7.	□□□□□□□□□ □□□□□□□ □□ □□□□□ □□□□□□□□ □□□□□□□□□.	Correct
8.	□□□□ □□□□□□□ □□□□□□□□□□ □□□□□ □□□□□ □□□□□□□ □□□□□□□ □□□□□□□□.	Correct
9.	□□□□ □□□□□□□ □□□□□□□ □□□□□□ □□□□□□□□□□□□ □□□□.	Correct
10.	□□□□□□□□ □□□□□□□□□□□□□ □□□□□□□□□□ □□□□□□ □□□□□□ நடக்கிறது.	Wrong
11.	□□□□□□□□ □□□□□ □□□□□□ □□□□□ □□□□ □□□□□.	Correct
12.	□□□ □□□□□□□ □□□□□□□ □□□□□□□□□□□□□, □□□□□□ □□□□□□□□□□ □□□□□□□ □□□□□□□□□□□□□.	Correct
13.	புதிய வர்த்தகம் கொள்கை , சரிசெய்யப்பட வேண்டும்.	Correct
14.	□□□□ □□□□□□□ □□□□□□ □□□□□□□□□□ □□□□□□□□□□□□ □□□□□□.	Correct
15.	□□□□ □□□□□□□ □□□□□ □□□□□□□□□□□□□ □□□□□□□□ □□□□□□.	Correct

The above table shows the performance of SVM for Sentiment Analysis. When 15 sentences i.e., political reviews were given to the system as an input for analyzing the sentiments, the system performed correctly for 13 sentences while it gave a wrong analysis for 2 sentences.

V. Conclusions And Future Work

The existing systems for political reviews in Tamil is not adequate in understanding and analyzing the sentiments in a comment/review. In this paper, Support Vector Machine algorithm is used to classify the text on basis of positive, negative and neutral reviews. In future work use of various other machine learning methods such as CRF, Maximum Entropy for sentiment extraction can be proposed.

References

[1] Walaa Medhat, Ahmed Hassan and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. Elsevier. 2014 May, 1093-1113.

[2] Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the Association for Computational Linguistics. 2002 July, 417–424.

[3] Pang, Bo, Lillian Lee, and 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. 2002, Volume 10.

[4] Chenhao Tan, Lillian Lee and Jie Tang. User-Level Sentiment Analysis Incorporating Social Networks. ACM Publications. 2011 August, 21-24.

[5] Sandeep Chandran, Bhadran V K, Santhosh, George and Manoj Kumar P. Document Level Sentiment Extraction for Malayalam (Feature Based Domain Independent Approach). International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. II. Special Issue XXV, 2015 April.

[6] Deepu S. Nair, Jisha P. Jayan, Rajeev R.R and Elizabeth Sherly. SentiMa - Sentiment Extraction for Malayalam. IEEE. September 2014, 1719-1723.

[7] M. Ravichandran, G. Kulanthaivel and T. Chellatamilan. Intelligent Topical Sentiment Analysis for the Classification of E-Learners and Their Topics of Interest. Hindawi Publishing Corporation, The Scientific World Journal. 2015

[8] Abd. Samad Hasan Basaria, Burairah Hussina, I. Gede Pramudya Anantaa and Junta Zeniarja. Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Published by Elsevier Ltd. 2013, 453-462.

[9] Muhammad Abdul-Mageed, Sandra Kubler and Mona Diab. SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media. Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Association for Computational Linguistics. 2012 July, 19-28.

[10] Gurunath H Naragund, Santhosh Kumar K L and Jharna Majumdar. Development of Decision Making and Analysis on Customer Reviews using Sentiment Dictionary for Human-Robot Interaction. International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 8, 2015 August.

[11] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou. Movie Rating and Review Summarization in Mobile Environment. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 42, No. 3. 2012 May.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 6, June 2016

BIOGRAPHY

D. Anish received B.E in Computer Science and Engineering Specialization in April 2014. She is currently pursuing her M.E in Computer Science and Engineering Specialization in Rajalakshmi Engineering College, Chennai- Tamil Nadu. Her research interests are Natural Language Processing & Machine Learning.

V. Sumathy received B.E in Computer Science and Engineering Specialization in April 2002. She was awarded Honor in M.E CSE in the April 2010. She is a Life Member in professional Bodies like ISTE. She is working as Asst. Professor in the Department of CSE in Rajalakshmi Engineering College, Chennai-Tamil Nadu. Her research interests are in the areas of Networks & Network Security.