Tree Based Opinion Mining in Tamil for Product Recommendations using R

A. Sharmista
Research Scholar, Dept. of Computer Applications
Madurai Kamaraj University
Tamil Nadu, India
ssharmistasaravananpkn@gmail.com

M. Ramaswami
Associate Professor, Dept. of Computer Applications
Madurai Kamaraj University
Tamil Nadu, India
mrswami123@gmail.com.

***Abstract*-Sentiment analysis, the automated extraction of expressions of positive or negative attitudes from text has received considerable attention from researchers during the past decade. In addition, the popularity of internet users has been growing fast parallel to emerging technologies; that actively use online review sites, social networks and personal blogs to express their opinions. In this paper we discuss some of the challenges in sentiment extraction especially in Tamil language and some of the approaches that have been taken to address these challenges and our approach analyses sentiments from Twitter social media. Tamil, a Dravidian language has a very rich morphological structure which is agglutinative. Tamil words are made up of lexical roots followed by one or more affixes, mostly suffixes. Tamil is also a post positional inflectional language. We have developed a Parts of speech tagging system to handle nouns and verbs. So finding a word in a language like Tamil is very complex. We try to resolve this complexity by identifying the categorical ambiguities and developing decision tree classification techniques at word grammatical category and grammatical feature level. These techniques were used to annotate the corpora and trained using the R Tool. The results obtained in each level were encouraging.**

Keywords-Sentiment Analysis, Natural Language Processing, Data Mining, Opinion Lexicon, and Decision tree classification.

I. INTRODUCTION

Being a human, words are used to express our thoughts and language acts as a medium to communicate those thoughts locally and globally. A World language is a language spoken internationally and which is learned by many people as a second language. The world's most widely used language is English which has over 1.8 billion users worldwide. Today, India is one of the multilingual nations in the world. The 'unity in diversity' describes India completely because a number of languages and dialect are spoken by Indians. The Constitution of India has recognized 18 national languages, each of which has a history and richness of its own. Further India is home to 22 official languages and more than thousands of spoken languages. The languages of southern India are mainly of the Dravidian group. Dravidian family consists of Malayalam, Tamil, Telugu and Kannada languages.

Language is medium for communicating our ideas, feelings or emotions to one another. Written communication either online or offline makes use of different languages for exchange of thoughts or feelings with one another. Various techniques and methods are present in the field of opinion mining and sentiment analysis [1] to extract the emotions from text. Gathering feeling or emotions associated with text is known as Opinion mining. Opinion mining is extracting opinions from text. This paper presents an analysis of Tamil language family present in India.

The enormous amount of information stored in a digital form are in unstructured texts cannot simply be used by computers, which normally handle text data as simple sequence of strings. Therefore we need a specific pre-processing methods are required in order to extract meaningful information from the documents. Natural Language Processing (NLP) is a tool for tracking the feeling of the public about a particular product [2]. The sentiment found within comments, feedback or critiques provide useful indicators for many different purposes. These sentiments can be categorized either into two categories: either positive or negative; or into none scale, e.g., very good, good, satisfactory, bad, very bad. A sentiment analysis task can be used as a classification task where each category represents a sentiment. The overview of the proposed model is shown in Figure. 1.

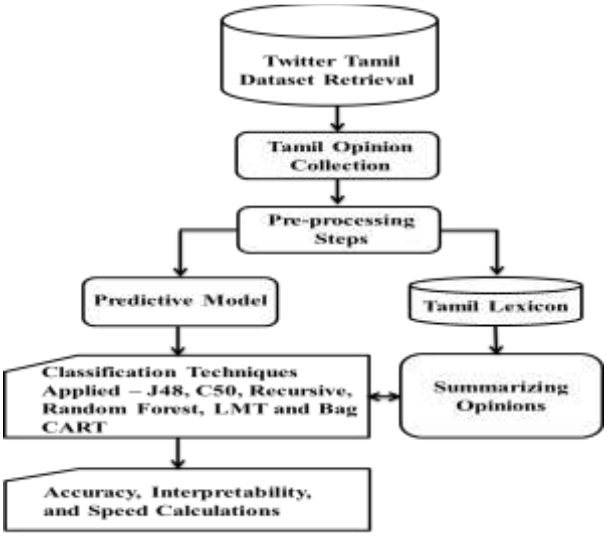
Twitter Tamil Dataset Retrieval
Tamil Opinion Collection
Pre-processing Steps
Predictive Model
Tamil Lexicon
Classification Techniques Applied – J48, C50, Recursive, Random Forest, LMT and Bag CART
Summarizing Opinions
Accuracy, Interpretability, and Speed Calculations

Figure 1. Overview of the Proposed Model

The development of the internet led to an exponential increase in the amount of electronic documents not only in English, but also other regional languages. Therefore the need for automatic Tamil text classification of user feedback about a particular product is growing at a fast pace. Automatic classification of Tamil documents is still on the research field since it is s morphologically rich language and agglutinative in nature [3]. Automatic text classification is the task of assigning predefined categories to unclassified text documents. By considering these problems here decision tree classification of Tamil documents were experimented on J48, C50, Recursive decision tree, CART, Random forest and Logistic Model Tree techniques. When an unknown document is given to the system it automatically assigns it the category which is more appropriate. The classification of textual data has practical significance in effective document management. In particular, as the amount of available online information increases, managing and retrieving [4] these feedbacks of a product are difficult without proper classification.

Our classification learning results reveal that incorporating social-network information can indeed lead to statistically significant sentiment classification improvements over the performance of an approach based on decision tree classification having access only to textual features. In this paper analysis of Tamil reviews from Twitter social-network was done.

The rest of the paper is organized as follows: In section II, the need for sentiment analysis was discussed. In section III review mining is discussed. In section IV problem studies is discussed. Results and concluding remarks are provided in section V.

II. Need for Sentimental Analysis in Tamil

Tamil is one of the oldest languages and it belongs to the South Dravidian family. Of all Dravidian languages, Tamil has the longest literary tradition. The earliest records are cave inscriptions from the second century B.C. Tamil is a morphologically rich and agglutinative language. Inflections are marked by suffixes attached to lexical base, which may be augmented by derivational suffixes. When morphemes or words combine, certain morphophonemic changes occur. Words in Tamil have a strong postpositional inflectional component. For verbs, these inflections carry information on the person, number and gender of the subject. Further, model and tense information for verb are also collocated in the inflections. For nouns, inflections serve to mark the case of the noun. The inflectional nature of the Tamil words prevents a simple stemming process like the one which is used for English documents [5]. A complete morphological analysis to find the stem is also cumbersome since it requires a stem dictionary.

Computationally, each root word can take a few thousand inflected word-forms, out of which only a few hundred will exist in a typical corpus. Subject-verb agreement is required for the grammaticality of a Tamil sentence. Tamil allows subject and object drop as well as verb less sentences. In addition, the subject of a sentence or a clause can be a possessive Noun Phrase (NP) or an NP in nominative or dative case. As Tamil is an agglutinative language, several suffixes can be added to the root word [6], thus forming thousands of different word forms.

The verb is the chief constituent of a sentence. Verbs take different argument structures [7] based on their semantic nature. These argument structures define the case markers of noun phrases, which are, for instance, direct and indirect objects to the verb. In addition, the predicate which is in finite form has to agree with the subject in terms of person-number-gender (e.g. avan – vand - taan, he came_3PMS‟). In Tamil, two noun phrases can constitute a sentence (avanmaaNa- van, he student‟ He is a student‟). As Tamil has relatively free word-order, the constituents of a sentence can be shuffled retaining the same meaning (avanpazhamcaappiTTaan, he fruit ate_3PMS, pazhamavancaappiTTaan, fruit he ate_3PMS, He ate a fruit‟).

Tamil nouns are inflected for case and number (plural). The morph tactics of nominal forms of Tamil is as follows:

Noun + (Plural maker) + Case marker

Consider the morphological decomposition of kaalkaLai, legs_ACC‟:

kaalkaLai =>kaal<N> + kaL<plural> + ai<accusative case>

Verbs are inflected for tense and finite and non- finite markers. A finite verb shows subject- agreement marker.

Verb + Tense + Subject agreement marker (per- son-number-gender)

Consider the morphological decomposition of the word paTittaan, he read: paTittaan =>paTi<V> + tt<past tense> + aan<3rd person, singular, masculine>

The agreement marker can simultaneously represent three distinct grammatical features: person, number and gender. For example, the morpheme – aan itself indicates the third person, singular number and masculine gender. These grammars are handled by POS (Parts of Speech) tagger part of the algorithm.

Sentiment analysis is the process of determining whether social media publications are positive or negative. Due to ambiguities inherent in language it can be very challenging to program software analysis tools to accurately resolve whether a word is positive or negative. Most of the sentiment analysis [8] materials available are in English. So, to interpret sentiment in Tamil, for example, which is spoken by approximately 20 per cent of the population, involves a time-consuming and often unreliable process of machine translation before analysis can take place”. Hence the need for sentiment analysis in Tamil is essential for a particular person to take decision on whether to buy a particular product or not.

III. Review Mining

An overview of the proposed model for our opinion summarization system has given in Figure. 1. The system performs the summarization in two main steps: feature extraction and aspect list lookup identification. The inputs to the system are a product name and an entry page for all the reviews of the product and the output is the summary of the reviews. Given the inputs, the system first downloads (or crawls) all the reviews, and puts them in the review database. The feature extraction function, which is the focus of this paper, first extracts “hot” features that a lot of people have expressed their opinions on in their reviews, and then finds those infrequent ones. The aspect list lookup identification function takes the generated features and summarizes the opinions of the feature into 2 categories: positive and negative. POS tagging [9] is the part-of-speech tagging from natural language processing. Below, we discuss each of the functions in feature extraction in turn.

A. Part –of- Speech Tagging(POS)

Before discussing the application of part-of-speech tagging from natural language processing [9], we first give some example sentences from some reviews to describe what kinds of opinions that we will handle. Our system aims to find what people like and dislike about a given product. Therefore how to find out the product features that people talk about is an important step. However, due to the difficulty of natural language understanding, some types of sentences are hard to deal with. Let us see some easy and hard sentences from the reviews of a digital camera:

“படங்கள்மிகதெளிவாகஉள்ளன.”

“ஒட்டுமொத்தஒருஅற்புதமானமிகவும்சிறியகேமரா.”

In the first sentence, the user is satisfied with the picture quality of the camera, படங்கள் is the feature that the user talks about. Similarly, the second sentence shows that கேமரா is the feature that the user expresses his/her opinion. While the features of these two sentences are explicitly mentioned in the sentences, some features are implicit and hard to find. For example,

“ஒளிஎன்றாலும், அதைஎளிதாகபைகளில்பொருந்தும்.”

In the above scenario, the customer is talking about the size of the camera, but the word “அளவு” is not explicitly mentioned in the sentence. To find such implicit features, semantic understanding is needed, which requires more sophisticated techniques. However, implicit features occur much less frequent than explicit ones. Thus in this paper, we focus on finding features that appear explicitly as nouns or noun phrases in the reviews [10]. To identify nouns/noun phrases from the reviews, we use the part-of-speech tagging.

In this work, we use the Natural Language Processor(NLP) linguistic parser, which parses each sentence and yields the part-of-speech tag of each word (whether the word is a noun, verb, adjective, etc) and identifies simple noun and verb groups (syntactic chunking). The following shows a sentence with the POS tags.

<S><NG><W C='PRP' L='SS' T='w' S='Y'>நான்</W></NG><VG><W C='VBP'>நான்</W><W
C='RB'>முற்றிலும்</W></VG><W C='IN'>உள்ள</W><NG><W C='NN'>பிரமிப்பு</W></NG><W
C='IN'>என்ற</W><NG><W C='DT'>இந்த</W><W C='NN'>கேமரா</W></NG><W C='.'> . </W></S>

The NLP system generates XML output. For instance, <W C=‘NN’> indicates a noun and <NG> indicates a noun group/noun phrase. Each sentence is saved in the review database along with the POS tag information of each word in the sentence. A transaction file is then created for the generation of frequent features in the next step. In this file, each line contains words from a sentence, which includes only preprocessed nouns/noun phrases of the sentence. The reason is that other components of a sentence are unlikely to be product features. Here, preprocessing includes the deletion of stop words, comments and special characters”.

B. Purning

Not all frequent features generated by decision tree classification are useful or are genuine features. There are also some uninteresting and redundant ones. Pruning aims to remove these incorrect features.

C. Opinion Words Extraction:

Opinion words are words that people use to express a positive or negative opinion. Observing that people often express their opinions of a product feature using opinion words that are located around the feature in the sentence, we can extract opinion words from the review database [11] using all the remaining frequent features (after pruning). For instance, let us look at the following two sentences:

“பட்டா கொடுமையாக இருக்கிறது மற்றும் நீங்கள் அணுக வேண்டும் கேமராபாகங்கள் வழியில் பெறுகிறார்.”

“கிட்டத்தட்ட 800 படங்கள் பிறகு நான் இந்த கேமரா நம்ப முடியாத படங்களை எடுக்கிறது என்று கண்டுபிடிக்கப்பட்டுள்ளது.”

In the first sentence, பட்டா, the feature, is near the opinion word பயங்கரமான. And in the second example, feature படங்கள் is close to the opinion word நம்பமுடியாத. Following from this observation, we can extract opinion words in the following way:

For each sentence in the review database, if it contains any frequent feature, extract the nearby adjective. If such an adjective is found, it is considered an opinion word. A nearby adjective refers to the adjacent adjective that modifies the noun/noun phrase that is a frequent feature. As shown in the previous example, பயங்கரமான is the adjective that modifies பட்டா, and நம்பமுடியாத is the adjective that modifies படங்கள்.

D. Opinion Sentence orientation determination:

After opinion features have been identified, we determine the semantic orientation (i.e., positive or negative) of each opinion sentence. This consists of two steps: (1) for each opinion word in the opinion word list, we identify its semantic orientation using decision tree classification techniques and (2) we then decide the opinion orientation of each sentence based on the dominant orientation of the opinion words in the sentence.

IV. Problem Study

Nowadays social media becomes part of a person's life. Social media such as Facebook, Twitter, Instagram or Linkedin has a numeral number of the user and keeps growing every day. It is estimated that over 500 million people are interacting with social media. The number of social media users growing has attracted marketers. Marketers have recognized that social media marketing as an important part of their marketing communication strategies. Also, social media helps organizations to communicate with their customers. These interactions help marketers determine customer needs and understand what their market might look like. Key business factors [12] of social media allow consumers to estimate products, make recommendations to contacts or friends, and share any of the purchases through their social media.

Moreover, recommendations by friends or connections on social media also could help consumers on decision-making. Those recommendations could help brand attitudes, purchasing attitudes, and advertising attitudes. According to the recommendations on purchases, 59% of all users were using Twitter. Most of top brands and services notice it and started to focus on social media marketing. So we have used Twitter to collect all the user reviews for different mobile phone products because of good speed responses. The number of instances collected is 100 and it is based on the attribute like தயாரிப்பு (Product), மாதிரி (Model), விலை (Prize), கருத்து (Feedback), நபரின் பெயர் (Name of the Person), பால் (Gender), வயது குழு (Age group), பகுதி (Area), தொழில்முறை (Profession), and இலக்கு (Target).

The decision tree is a popular classification method and it displays relationships found in the training data. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. Popular decision tree algorithms like Recursive decision tree, C4.5 (J48), C50, LMT, Random Forest and Bag-Cart were used in this paper to select the next best attribute. The goal of these algorithms is to learn the decision function stored in the data and then use it to classify new inputs. The class label used is இலக்கு (Target).

To implement this machine learning approach the trees are constructed in a top-down recursive divide-and-conquer manner.After training and testing these three methods with a labeled twitter corpus, we selected the decision tree classifier for sentiment analysis since it’s faster and at the same time it gives little more accuracy.

We evaluate the effect of decision-tree algorithms on tweets dataset containing emoticons in classifying process. Decision Tree algorithm generates a tree-like graphical representation of the model it produces [13]. Usually accompanied by rules of the form “if condition then outcome,” which constitute the text version of the model, decision trees have become popular because of their easily understandable results. Some commonly implemented decision tree algorithms include Classification and regression trees (CART).When choosing an algorithm for a predictive model, we must weigh three important criteria: accuracy, interpretability, and speed.

*Accuracy:*We measure accuracy by generating predictions for cases with known outcomes and then compare the predicted value to the actual value. For classifiers a prediction is either right or wrong, so we can state the accuracy as percentage correct, or as an error rate (percentage wrong). Despite the claims you may encounter from various software vendors, no “most accurate” algorithm exists [13]. In some cases, Decision tree will produce the most accurate classifier. Thus, we usually try to apply at least six algorithms to a data set to see which has the best accuracy. Our experience shows that LMT and Random forest frequently have somewhat higher accuracy than other algorithms.

Interpretability:Decision tree model produce the most extensive interpretive information. Decision trees can find and report interactions—for example, “Customer reviews in Tamil are extracted from Twitter on particular product they have purchased and reports of the reviews of many users who purchased the same product was generated. From this we can find whether to purchase that product or not.

Speed: Processes associated with predictive modeling emphasize speed: the time it takes to train a model and make predictions about new cases.

V. Results and Analysis

We can use to do our sentiment analysis is by utilizing the R package. R is a sophisticated statistical software package [14], which provides new approaches to data mining. Rattle, an R GUI, an open source tool for analysis of data mining algorithms. Rattle is built on the statistical language R. Rattle is simple to use, quickly to deploy, and allows to rapidly work through the data processing. In this work, we analyzed an effect of product review data set obtained from *Opinion Lexicon* that contains Tamil corpus. The algorithm is executed using Rattle to predict the best product by identifying the total number of nermarai (Positive) opinions as in the figure 2 and their results are in the table I. The number of instances used for analysis of product data is 100.

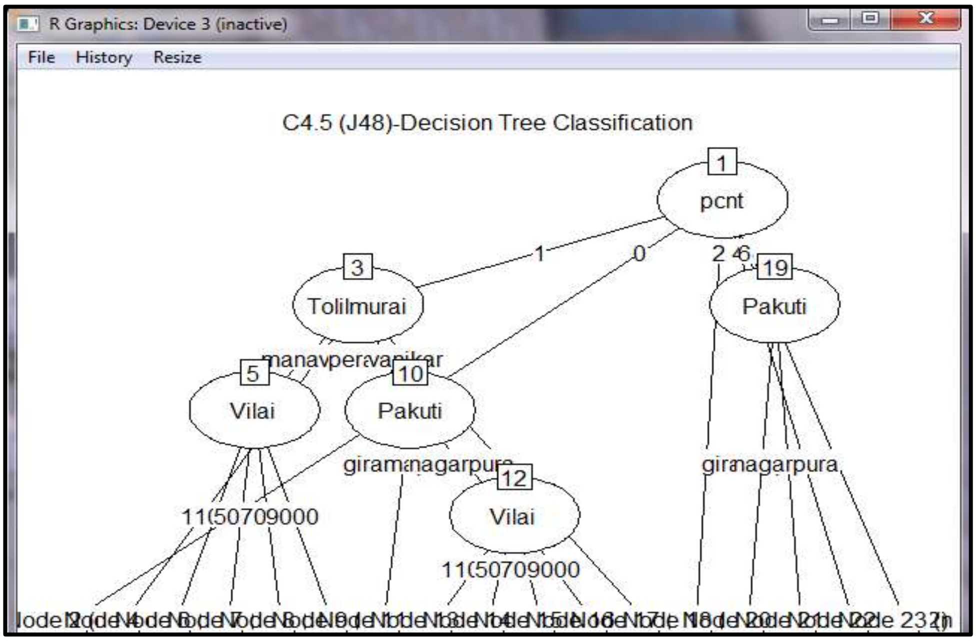
R Graphics: Device 3 (inactive)
File
History
Resize
C4.5 (J48)-Decision Tree Classification
1
pcnt
1
0
2
6
3
Tolilmurai
19
Pakuti
manavperavanikar
5
Vilai
10
Pakuti
giramnagarpura
12
Vilai
giranagarpura
11(50709000
11(50709000
lode2Node4Node6Node7Node8Node9Node11Node13Node14Node15Node16Node17Node18Node20Node21Node22Node 23(n

Figure 2. J48 with Opinion lexicon for Product Dataset

Table I. Confusion Matrix of J48 with Opinion Lexicon

Actual Class	**Predicated Class**		
	Ethirmari	***Nermarai***	***Yarum***
	11	3	0
	5	28	12
	13	4	23

Figure 3 shows results of LMT decision tree with opinion lexicon for product reviews dataset run on R platform.Table II. Shows results of LMT with opinion lexicon for product reviews dataset run on R platform. It depicts the nermarai, ethirmarai and yarum reviews which are classified based on the number of instances 100. It is predicted that there are 27 ethirmarai reviews, 30 yaru classified reviews and 35 nermarai reviews in case when matched with opinion lexicon.

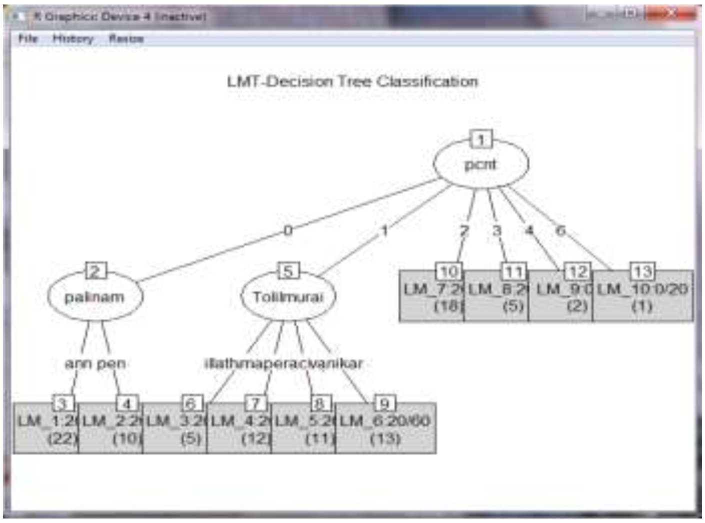

Figure 3. LMT with Opinion lexicon for Product Dataset

Table II. Confusion matrix of LMT with Opinion Lexicon

	Predicated Class		
	Ethirmari	***Nermarai***	***Yarum***
Actual Class	27	0	2
	2	35	3
	0	0	30

Figure.4. shows results of CART decision tree with opinion lexicon for product reviews dataset run on R platform.

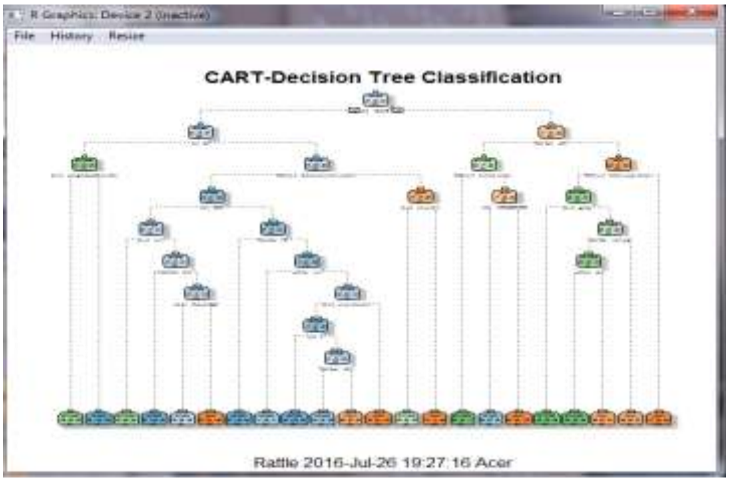
R Graphics: Device 2 (inactive)
File History Resize
CART-Decision Tree Classification
Rattle 2016-Jul-26 19:27:16 Acer

Figure 4. CART with Opinion lexicon for Product Dataset

Table III shows results of CART with opinion lexicon for product reviews dataset run on R platform. It depicts the nermarai, ethirmarai and yarum reviews which are classified based on the number of instances 100. It is predicted that there are 18 ethirmarai reviews, 23 yarum classified reviews with 10 error rate and 29 nermarai reviews in case when matched with opinion lexicon.

Table III. Confusion matrix of CART with Opinion Lexicon

	Predicated Class		
	Ethirmari	***Nermarai***	***Yarum***
Actual Class	18	2	2
	5	29	10
	6	4	23

Table IV shows results of Recursive decision tree with opinion lexicon for product reviews dataset run on R platform. It depicts the nermarai, ethirmarai and yarum reviews which are classified based on the number of instances 100. It is predicted that there are 20 ethirmarai reviews, 30 yarum classified reviews with 6 error rate and 30 nermarai reviews in case when matched with opinion lexicon.

Table IV. Confusion matrix of Recursive Decision Tree with Opinion Lexicon

Actual Class	**Predicated Class**		
	Ethirmari	***Nermarai***	***Yarum***
	20	4	1
	4	30	4
	5	1	30

Table V. shows results of Random forest with opinion lexicon for product reviews dataset run on R platform. It depicts the nermarai, ethirmarai and yarum reviews which are classified based on the number of instances 100. It is predicted that there are 26 ethirmarai reviews, 32 yarum classified reviews with 2 error rate and 34 nermarai reviews in case when matched with opinion lexicon.

Table V. Confusion matrix of Random Forest with Opinion Lexicon

Actual Class	**Predicated Class**		
	Ethirmari	***Nermarai***	***Yarum***
	26	0	1
	2	34	2
	1	1	32

Table VI. shows results of C50 with opinion lexicon for product reviews dataset run on R platform. It depicts the nermarai, ethirmarai and yarum reviews which are classified based on the number of instances 100. It is predicted that there are 7 ethirmarai reviews, 17 yarum classified reviews and 29 nermarai reviews in case when matched with opinion lexicon.

Table VI. Confusion matrix of C50 with Opinion Lexicon

Actual Class	**Predicated Class**		
	Ethirmari	***Nermarai***	***Yarum***
	7	4	0
	9	29	18
	13	2	17

The result of LMT and Random forest classification techniques shows better result when compared to other techniques of classification with product review data set obtained from *Opinion Lexicon* that contains Tamil corpus.

Table VII. Balanced Accuracy Calculation for Decision Tree Classification Techniques

Balanced Accuracy Calculation						
Attrubite	**Decision Tree Classification Techniques**					
இலக்கு (Target)	***C4.5(J48)***	***LMT***	***Bag-Cart***	***Recursion***	***Random Forest***	***C50***
Ethirmari (Negative)	0.6682	0.9512	0.7818	0.8091	0.9411	0.5921
Nermarai (Positive)	0.7672	0.9609	0.7971	0.8661	0.9545	0.7033
Yarum (None)	0.6958	0.9286	0.7504	0.8817	0.9415	0.6257

These results were calculated for all decision tree classification techniques such as J48, LMT, Bag-Cart, Recursive, Random Forest and C50. Accuracy was predicted over Ethirmarai (Positive), Nermarai (Negative) and Yarum (None) for each technique. Logistic Model Tree (LMT) and Random Forest shows more accuracy when compared to other techniques.

VI. CONCLUSION

It is a very important fact to analyze how people think in different context about different things. This becomes more important when it comes to the business world because business is dependent on their customers and they always try to make products or services in order to fulfill customer requirements. So knowing what they want, what they think and talk about existing products, services and brands is more useful for businesses to make decisions such as identifying competitors and analyzing trends. Both because people express their ideas on social media and it can access those data, it has been enabled in some way to do the above mentioned things by using those data. Finally it does data mining with the extracted sentiments so that decision making can be done by the customer on whether to buy the particular product or not. The first thing that was found to address this challenge was a lexical data source which is called Opinion lexicon, in that it has positive and negative words. During the implementation of the sentiment module we had to consider several issues such as, the comment by the user of a product or a brand can be not only in English but also mix with other language (Tamil), with emotional symbols etc., the comment may not completely match with what exactly user need to express about the product or brand, identifying the entity, identifying the relation of a particular comment with previous comments, ambiguity of words of the comment, human language is noisy and chaotic and the users may use different jargon or slang communications. But with the implementation machine learning techniques, it could achieve more accurate results after building classifiers training on large labeled data sets but still there are some issues of processing natural language. Finally, using the above mentioned techniques for sentiments regarding particular product or service with the user's information, it could successfully profile the products, analyze trends and forecasting. So, as overall, the system is capable of saying that how a set of people of a particular age range, a particular area with a particular profession think about a particular product or service and how it will change it the future which are most useful information when it comes to business world.

REFERENCES

[1] Bing Liu, Sentiment Analysis and Opinion Mining, Morgan &Claypool Publishers, 2012.

[2] Bo Pang and Lillian Lee, Opinion mining and sentiment analysis, pp.1–135, 2008.

[3] Annamalai, E., & Steever, Modern Tamil in Dravidian languages, Newyork: Routledge Publication, 1999.

[4] Pang, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieva, pp.1–135, 2008.

[5] M.Thangarasu and Dr.R.Manavalan, Design and Development of Stemmer for Tamil Language: Cluster Analysis, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.3, No.7, 2013.

[6] Rajendran, S., S. Arulmozi, B. Kumara Shanmugam, S. Baskaran, and S. Thiagarajan, Tamil WordNet, Proceedings of the First International Global WordNet Conference, pp.271- 274, 2002.

[7] Lehmann, Thomas, A grammar of modern Tamil. Pondicherry, India: Pondicherry Institute of Linguistics and Culture, 1993.

[8] Liu B., Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing, 2010.

[9] Dave K., Lawrence S, and Pennock D.M., Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, Proceedings of the 12th international conference on World Wide Web(WWW), pp.519–528, 2003.

[10] Tang H, Tan S, and Cheng X, A survey on sentiment detection of reviews, Expert Systems with Applications: An International Journal, pp.10760–10773, 2009.

[11] Zheng-Jun Zha, Jianxing Yu, Jinhui Tang,Meng Wang, and Tat-Seng Chua, Product Aspect Ranking and Its Applications, IEEE Transactions on knowledge and data engineering, Vol.26, No.5, pp.1211-1224, 2014.

[12] T.T. Thet, J. Cheon, and C. Khoo, Aspect-based sentiment analysis of movie reviews on discussion boards, Journal of Information Science, Vol.36, pp.823-848, 2010.

[13] M. J. Lee and A. S. Hanna, Decision tree approach to classify and quantify cumulative impact of change orders on productivity, Journal of Computing in Civil Engineering, pp.132–144, 2004.

[14] O'Reilly Media,Sebastopol, http://radar.oreilly.com/archives/2006/12/web_20_compact.html, 2007.