KEC@DPIL-FIRE2016: Detection of Paraphrases on Indian Languages

R.Thangarajan
Professor
Department of CSE
Kongu Engineering College
Erode, Tamilnadu
+919443014942
rt.cse@kongu.edu

S.V.Kogilavani
Assistant Professor (SrG)
Department of CSE
Kongu Engineering College
Erode, Tamilnadu
+919486153223
kogilavani.sv@gmail.com

A.Karthic
UG Student
Department of CSE
Kongu Engineering College
Erode, Tamilnadu
+919443840800
karthic1011@gmail.com

S.Jawahar
UG Student
Department of CSE
Kongu Engineering College
Erode, Tamilnadu
+918973744171
jawahar273@gmail.com

ABSTRACT
This paper presents a report on Detecting Paraphrases in Indian Languages (DPIL), in particular the Tamil language, by the team NLP@KEC of Kongu Engineering College. Automatic paraphrase detection is an intellectual task which has immense applications like plagiarism detection, new event detection, etc. Paraphrase is defined as the expression of a given fact in more than one way by means of different phrases. Paraphrase identification is a classic natural language processing task which is of classification type. Though there are several algorithms for paraphrase identification, reflecting the semantic relations between the constituent parts of a sentence plays a very important role. In this paper we utilize sixteen different features to best represent the similarity between sentences. The proposed approach utilizes machine learning algorithms like Support Vector Machine and Maximum Entropy for classification of given sentence pair. They have been classified into Paraphrase and Not-a-Paraphrase for task1 and Paraphrase, Not-a-Paraphrase and Semi-Paraphrase for task2. The accuracy and performance of these methods are measured on the basis of evaluation parameters like accuracy, precision, recall, f-measure and macro f-measure. Our methodology got 2nd place in DPIL evaluation track.

Keywords
Natural Language Processing; Paraphrase Identification; Machine Learning Approach; Support Vector Machine; Maximum Entropy; Shallow Parser.

1. INTRODUCTION
Variability of semantic expression is a fundamental phenomenon of natural language where in the same meaning can be expressed by, or inferred from, different texts. Paraphrases are alternative ways to express or convey the same information. One can express a single event in many different ways in natural language sentences which depends on the command of the language the writer or speaker has on the language in consideration. A properly written paraphrase expresses the ideas of a fact or event in words and sentence structure inherent to the writer or speaker. It is similar to summarization but the key difference is that paraphrases include both key points and sub-points. Because a paraphrase includes detailed information it can sometimes be as long as the original source. A paraphrase typically explains or clarifies the text that is being paraphrased. This greatly adds to the difficulty of detecting paraphrases

Paraphrase detection is the task of determining whether two or more sentences represent the same meaning or not [1]. Paraphrase detection systems progress the performance of a paraphrase generation by choosing the best sentence from the list of paraphrase sentences. Plagiarism detection is another task which needs the paraphrase identification technique to detect the sentences which are paraphrases of others. Identifying paraphrases is an important task that is used in information retrieval, question answering, text summarization and plagiarism detection. This work mainly focuses on the detection of paraphrases in Tamil language. For example, the sentences in Table 1 express the same meaning therefore, they are paraphrases.

Table 1 Sample Sentence in Tamil Language

கேரளமாநிலம்திருச்சூரில்கூடல்மாணிக்கம் கோயில்திருவிழாதுவங்கியது.
கூடல்மாணிக்கம்கோயில்திருவிழாகோலாகலமாகதுவங்கியது.

Our proposed system utilizes two supervised machine learning approaches using a Support Vector Machine (SVM), Maximum Entropy (ME) and learns classifiers based on sixteen features like lexical and POS tagging features in order to detect the paraphrase sentences. The structure of the report is defined as follows: Section-2 describes literature review. Section-3 gives the task description Section-4 represents the overview of the proposed system. Section-5 presents the performance evaluation results and Section-6 concludes the work.

2. LITERATURE REVIEW
In this section, recent research work carried in the field of paraphrase identification in general, and paraphrase identification in Tamil language in particular is discussed. Many researchers on paraphrase identification make use of existing Natural Language Processing (NLP) tools to identify paraphrases. [2] exploits the NLP tools of a QA system to identify paraphrases. [3-6] have employed lexical semantic similarity information based on resources such as WordNet [7]. Tamil paraphrase detection is tried with deep learning method [18].

The ability to identify paraphrase, in which a sentence express the same meaning of another one but with different words, has proven useful for a wide variety of Many different natural language processing applications are there for detecting paraphrases. The different approaches can be categorized into supervised methods, i.e. [8-9], which are the most promising methods.

Most of the existing system utilizes thresholds to determine whether two sentences are similar and represents the same meaning. But the specification of exact threshold will depends up on the training data. Machine Learning (ML)techniques have been applied in order to overcome the problems in setting the threshold.

The benefit of applying the ML approach resides based on the morphologic, syntactic, semantic features of a sentence.

In general, supervised and unsupervised machine learning techniques are quite useful in paraphrase detection. In supervised learning technique, the dataset is labeled and trained to obtain a reasonable output which help in proper decision making. Unlike supervised learning, unsupervised learning process does not need any label data; therefore they cannot be processed easily. This report work presents the impact of two supervised learning methods on given dataset.

3. TASK DESCRIPTION

One of the most commonly used corpora for paraphrase detection is the Micro Soft Research in Paraphrase (MSRP) corpus [10], which contains 5,801 English sentence pairs from news articles manually labeled with 67% paraphrases and 33% non-paraphrases. Since there are no annotated corpora or automated semantic interpretation systems available for Indian languages till date, creating benchmark data for paraphrases and utilizing that data in open shared task competitions will motivate the research community for further research in Indian languages [11]. We participated in DPIL task which is focused on sentence level paraphrase identification for Indian languages (Tamil, Malayalam, Hindi and Punjabi). In this context, the task is divided into two subtasks.

3.1 Sub Task 1:

Given a pair of sentences, the system is required to assess if the two sentences carry the same meaning or not and to classify them into Paraphrase (P), or Not Paraphrase (NP) otherwise.

3.2 Sub Task 2:

Given two sentences from newspaper domain, the task is to identify whether they are completely equivalent (P) or roughly equivalent (SP) or Not Equivalent (NE). This task is similar to the subtask 1, but the main difference is 3-point scale tag in paraphrases.

The training data set given to us is the News Corpus, which contains 2,500 Tamil sentence pairs for subtask1 and 3,500 Tamil sentence pairs for subtask2. The test set given to us consisted of 900 Tamil sentence pairs for subtask1 and 1,400 Tamil sentence pairs for subtask2. Both the training and test corpus are available at the link specified in [12].

4. OVERVIEW OF PROPOSED SYSTEM

Subtask1 and subtask2 datasets are processed using Tamil shallow parser. The processed results are in text format; but for classification of sentences using the machine learning algorithms, the text values are converted into numerical matrix, which is then given as input into SVM and ME for further classification.

SVM: Given data are analyzed and decision boundaries are defined by having hyper planes. In two category case, the hyper plane separates the document vector of one class from other classes, where the separation is maintained to be large as possible [13].

ME: The training data are used to set constraint on conditional distribution [14]. Each constraint is used to express characteristics of training the data. These constraints then are used for testing the data. The results obtained from this analysis are compared using different performance evaluation measures.

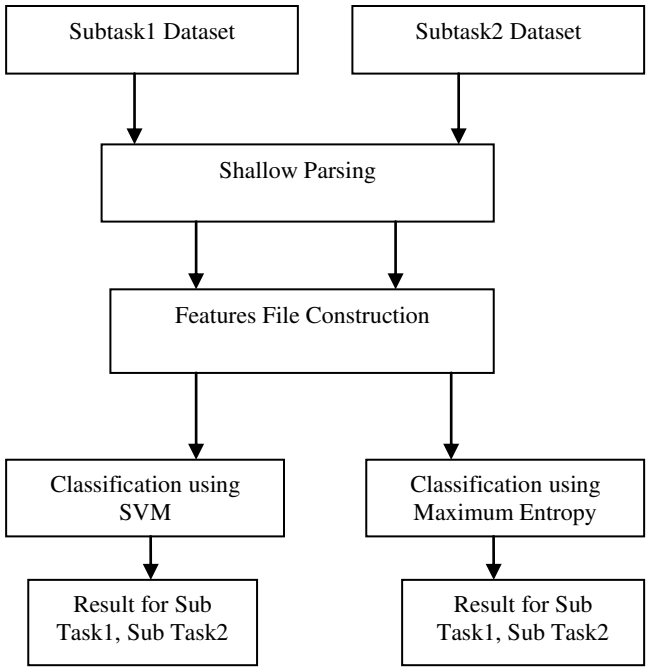

Figure 1: Overview of Proposed System

4.1 SHALLOW PARSING

Shallow parsing also called as chunking or light parsing is an analysis of a sentence which first identifies basic parts of sentences such as nouns, verbs, adjectives, etc., and then links them to higher order units that have discrete grammatical meanings. While the most elementary chunking algorithms simply link constituent parts on the basis of elementary search patterns, approaches that use ML techniques can take contextual information into account and thus compose chunks in such a way that they better reflect the semantic relations between the basic constituents. That is, these more advanced methods get around the problem that combinations of elementary constituents can have different higher level meanings depending on the context of the sentence. The proposed system utilizes Tamil Shallow parser developed by IIIT [15]. For example, the following Tamil sentence

"செயற்கைக்கோள்களை ஏவி இன்று இஸ்ரோ புதிய சாதனை படைத்துள்ளது."

would be chunked as in Table 2.

Table 2 Chunking Result of Sample Sentence

<Sentence id="1">			
1	((	NP	<fsaf='ஏவு,v,any,any,any,,இ,i' vbp="Y" head="ஏவி" poslcat="NM" paradigm="v5">
1.1	செயற்கைக்	NN	<fsaf='செயற்கை,n,any,sg,any,d,0,' case_name="nom" paradigm="n2">
1.2	கோள்களை	NN	<fsaf='கோள்,n,any,pl,any,d,ஐ,kalY_E' case_name="acc" paradigm="n17">
1.3	ஏவி	NN	<fsaf='ஏவு,v,any,any,any,,இ,i' vbp="Y" name="ஏவி" poslcat="NM" paradigm="v5">
	))		
2	((	NP	<fsaf='இன்று,n,any,sg,any,d,0,' head="இன்று" case_name="nom" paradigm="n6">
2.1	இன்று	NN	<fsaf='இன்று,n,any,sg,any,d,0,' name="இன்று" case_name="nom" paradigm="n6">
	))		
3	((	NP	<fsaf='இஸ்ரோ,n,any,sg,any,d,0,' head="இஸ்ரோ" case_name="nom" paradigm="n1">
3.1	இஸ்ரோ	NN	<fsaf='இஸ்ரோ,n,any,sg,any,d,0,' name="இஸ்ரோ" case_name="nom" paradigm="n1">
	))		
4	((	NP	<fsaf='சாதனை,n,any,sg,any,d,0,' head="சாதனை" case_name="nom" paradigm="n2">
4.1	புதிய	JJ	<fsaf='புதிய,adj,any,any,any,,,' paradigm="adj">
4.2	சாதனை	NN	<fsaf='சாதனை,n,any,sg,any,d,0,' name="சாதனை" case_name="nom" paradigm="n2">
	))		
5	((	VGF	<fsaf='படை,v,n,sg,3,,த்து_உள்_ள்,wwu_ulY_lY_awu' vbp="Y" head="படைத்துள்ளது" tense="PRESENT" paradigm="v11" finite="Y">
5.1	படைத்துள்ளது	VM	<fsaf='படை,v,n,sg,3,,த்து_உள்_ள்,wwu_ulY_lY_awu' vbp="Y" name="படைத்துள்ளது" tense="PRESENT" paradigm="v11" finite="Y">
	))		
5.2	.	SYM	<fsaf='˙ ,punc,,,,,,'>

4.2 FEATURES FILE CONSTRUCTION

From the output of shallow parsing process, feature file is constructed both for training and test datasets for subtask1 and subtask2. Sample feature file is presented in Table 3

Table 3 Sample Sentence- Feature File

Para phrase Id	ben	nom	RB	Present	past	acc	future	loc	gen	dat	JJ	NNP	NN	soc	VM	count	class
TAM 0001	0	10	2	0	2	0	2	3	0	0	2	2	11	0	6	6	P
TAM 0002	0	8	1	0	2	0	0	1	0	0	1	1	9	0	2	4	P
TAM 0003	0	12	1	2	3	0	0	0	0	5	0	0	19	0	5	6	P
TAM 0004	0	15	2	0	1	0	3	2	0	3	2	0	20	0	3	4	P
TAM 0005	0	11	0	1	0	0	0	0	0	0	1	3	16	0	1	7	P
TAM 0006	0	13	3	2	1	0	0	0	0	1	0	0	17	0	7	3	P
TAM 0007	0	6	0	0	1	0	1	3	1	0	0	0	14	0	2	2	P
TAM 0008	0	5	1	1	1	0	0	3	0	2	1	0	12	0	2	7	P
TAM 0009	0	8	0	1	0	0	1	0	0	3	1	1	12	0	2	4	P
TAM 0010	0	7	0	0	2	1	0	4	0	0	0	1	15	0	3	7	P

4.3 TEXT CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

When supervised machine learning algorithms are considered for classification purpose, the input dataset is desired to be a labeled one. In this study, two different supervised learning techniques are applied for classification purpose such as SVM and ME.

Classification of sentences may be categorized into two types, i.e., binary sentence classification and multi-class sentence classification [16]. For the given dataset, in binary classification type, each sentence pair is classified as a label C, where C = {P, NP}. In this, P denotes the given sentence pair is a paraphrase and NP denotes that the given sentence pair is not a paraphrase. In multi class sentence classification, each sentence pair is classified as a label *C*, where *C* = {P, NP, SP}. In this, P specifies that the sentence pair is paraphrase, NP denotes that the given sentence pair is not a paraphrase whereas SP specifies that the given sentence pair is a semi paraphrase.

4.3.1 SVM Classification Method

SVM is based on the structural risk minimization principle from computational learning theory. This method analyzes data and defines decision boundaries by having hyper-planes. In binary classification problem, the hyper-plane separates the given vector in one class from other class, where the separation between hyper-planes is desired to be kept as large as possible. One property of SVM is that their ability to learn can be independent of the

dimensionality of the feature space. SVM measure the complexity of hypotheses based on the margin with which they separate the data, not the number of features [17]. Since SVM requires input in the form of a vector of numbers, the constructed feature file is given as input to SVM.

4.3.2 ME Classification Method

ME is a general technique for estimating probability distributions from data. The over-riding principle in maximum entropy is that when nothing is known, the distribution should be as uniform as possible, that is, have maximal entropy. Labeled training data is used to derive a set of constraints for the model that characterize the class-specific expectations for the distribution. Constraints are represented as expected values of "features" any real-valued function of an example. The improved iterative scaling algorithm finds the maximum entropy distribution that is consistent with the given constraints.

Due to the minimum assumptions that the ME classifier makes, we regularly use it when we don't know anything about the prior distributions and when it is unsafe to make any such assumptions. Moreover ME classifier is used when we can't assume the conditional independence of the features. This is particularly true in text classification problems where our features are usually words which obviously are not independent. The ME method requires more time to train comparing to Naive Bayes, primarily due to the optimization problem that needs to be solved in order to estimate the parameters of the model. Nevertheless, after computing these parameters, the method provides robust results and it is competitive in terms of CPU and memory consumption.

In our text classification scenario, maximum entropy estimates the conditional distribution of the class label given pair of sentences. Entire document is represented by a feature file. The labeled training data is used to estimate the expected value on a class-by-class basis.

5. PERFORMANCE EVALUATION PARAMETERS AND RESULTS

The parameters which are helpful to evaluate performance of supervised machine learning algorithm is based on the element from a matrix known as confusion matrix or contingency table. It is used in supervised machine learning algorithm to help in assessing performance of any algorithm. From classification point of view, terms such as "True Positive"(TP), "False Positive" (FP), "True Negative" (TN), "False Negative" (FP) are used to compare label of classes in this matrix as shown in Table 4.

Table 4 Confusion Matrix Format

Predicted Status	Actual Status	
	Not a Paraphrase	Paraphrase
Not a Paraphrase	TN(True Negative)	FN(False Negative)
Paraphrase	FP(False Positive)	TP(True Positive)

In Table 4, True Positive represents the number of sentences those are paraphrase and also classified as paraphrase by the classifier, where as False Positive indicates paraphrase sentences, but classifier does not classify it as paraphrase. Similarly, True Negative represents the sentences which are not paraphrase also classified as not paraphrase by the classifier, whereas False Negative are not paraphrase sentences but classifier classify it as paraphrase. Table 5 and Table 6 represents confusion matrix for sub task 1 sentences and sub task 2 sentences.

Table 5 Confusion Matrix for Subtask1 Sentences

	Actual Status (SVM)		Actual Status (ME)	
Predicted status	NP	P	NP	P
NP	409	129	420	120
P	117	245	106	254

Table 6. Confusion Matrix for Subtask2 Sentences

	Actual Status (SVM)			Actual Status (ME)		
Predicted status	NP	P	SP	NP	P	SP
NP	499	47	125	514	43	116
P	51	113	150	44	131	164
SP	95	113	207	87	99	202

In SVM method, out of 900 sentence pairs, 409 NP sentences are correctly classified as NP sentences and 129 NP sentences are misclassified as P sentences. Similarly, 245 sentences are correctly classified as P sentences and 117 P sentences are misclassified as NP sentences. We believe that this misclassification is mainly due to higher lexical similarity in false paraphrase pairs, which makes them hard to be differentiated from true paraphrase pairs. In ME method, out of 900 sentence pairs, 420 NP sentences are correctly classified as NP sentences and 120 NP sentences are misclassified as P sentences. Similarly, 254 sentences are correctly classified as P sentences and 106 P sentences are misclassified as NP sentences. The ME system performs fairly well compared to SVM method at identifying true paraphrase pairs, as given in Table 5for sub task 1 and Table 6 for sub task2.

Based on the values obtained from confusion matrix, other parameters such as "precision", "recall", "f-measure", and "accuracy" are found out for evaluating performance of any classifier.

•*Precision:*

It measures the exactness of the classifier result. It is the ratio of number of examples correctly labeled as paraphrase to total number of paraphrase sentences in sub task1 dataset. It can be calculated using the equation 1.

$$Precision = \frac{\text{TP}}{(\text{TP}+\text{FP})} \quad (1)$$

•*Recall:*

It measures the completeness of the classifier result. It is the ratio of total number of paraphrase sentences to total sentences which are truly paraphrase. It can be calculated using the equation 2.

$$Recall = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad (2)$$

•*F-Measure:*

It is the harmonic mean of precision and recall. It is required to optimize the system towards either precision or recall, which have more influence on final result. It can be calculated using the equation 3.

$$F - Measure = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (3)$$

The tables 7 and 8 represent Precision, Recall and F-Measure summary for sub task 1 and sub task2 in identifying paraphrase and not a paraphrase sentences. In Table 7 SVM-NP stands for SVM used for identifying Not Paraphrase (NP) sentences and SVM-P stands for SVM used for identifying Paraphrase (P) sentences. Similarly ME-NP stands for ME used for identifying Not Paraphrase (NP) sentences and ME-P stands for SVM used for identifying Paraphrase (P) sentences. The results show that precision, recall and F-Measure values are high for Non Paraphrase identification by both SVM and ME systems.

Table 7 Precision, Recall, F-Measure summary for SubTask1

Method	Precision	Recall	F-Measure
SVM – NP	0.76	0.78	0.77
SVM – P	0.68	0.66	0.67
ME - NP	0.78	0.80	0.79
ME – P	0.71	0.68	0.69

In Table 8, SVM-SP stands for SVM used for identifying Semi Paraphrase (SP) sentences and ME-SP stands for ME used for identifying Semi Paraphrase (SP) sentences. The results show that both the systems identify sentences in the order NP, SP and P.

Table 8 Precision, Recall, F-Measure summary for SubTask2

Method	Precision	Recall	F-Measure
SVM – NP	0.74	0.77	0.75
SVM – P	0.36	0.41	0.38
SVM – SP	0.50	0.44	0.47
ME - NP	0.76	0.80	0.78
ME – P	0.39	0.48	0.43
ME – SP	0.52	0.42	0.46

•*Accuracy:*

It is the most common measure of classification process. It can be calculated as the ratio of correctly classified sentences to total number of sentences. It can be calculated using the equation 4.

$$Accuracy = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (4)$$

•*F1 Macro Measure:*

F1 macro measure can be used when you want to know how the system performs overall across the sets of data.

The comparative analysis based on results obtained using proposed approaches are shown in Table 9. It can be analyzed that the accuracy obtained using ME method is better than that of SVM because of dependent feature and also high dimensionality and sparseness of text data.

Table 9 Accuracy of Machine Learning Approaches

Method/Task	Accuracy	F1 Measure / F1 Macro Measure
SVM – Sub Task1	0.73	0.72
SVM – Sub Task2	0.59	0.53
ME – Sub Task1	0.75	0.74
ME – Sub Task2	0.61	0.56

6. Conclusion

This work makes an attempt to classify given sentence pairs into paraphrases or not using two supervised machine learning algorithms, such as SVM and ME. In this paper we utilize sixteen different semantic features to best represent the similarity between sentences. Two machine learning algorithms such as Support Vector Machine and Maximum Entropy have been considered for classification of given sentence pair into Paraphrase (P)and Not-a-Paraphrase(NP) for task1 and Paraphrase(P), Not-a-Paraphrase(NP) and Semi-Paraphrase (SP) for task2. The accuracy and performance of these methods are measured on the basis of parameters such as accuracy, precision, recall, f-measure and macro F-measure. The results show that ME method outperforms than SVM to identify paraphrases.

7. References

[1] Qayyum, Z., and Altaf, W. 2012. Paraphrase Identification using Semantic Heuristic Features, Research Journal of Applied Sciences, Engineering and Technology, 4(22) (pp.4894-4904).

[2] Duclaye, F., Yvon F., Collin O., and Cedex L. 2002. Using the Web as a Linguistic Resource for Learning Reformulations Automatically. In proceedings of the Third International Conference on Language Resources and Evaluation (pp.390-396).

[3] Finch, A., Hwang, Y., and Sumitha, E. 2005. Using Machine Translation Evaluation Techniques to determine Sentence-level semantic Equivalence. In proceedings of the Third International Workshop on paraphrasing (pp.17-24).

[4] Mihalcea, R., Corley and Strapparava C.2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In proceedings of 21st National Conference on Artificial Intelligence, Vol:1 (pp:775-780).

[5] Fernando, S., and Stevenson, M.2008. A semantic similarity approach to paraphrase detection. In proceedings of 11th Annual research colloquium of the UK special interest group for computational linguistics (pp.45-52).

[6] Malakasiotis, P. 2009. Paraphrase recognition using machine learning to combine similarity measures. In proceedings of

the ACL-IJCNLP 2009 Student Research Workshop (pp.27-35).

[7] Miller, G.A.,(19950. WordNet: A lexical database for English. Communications of the ACM, Vol;38, No:11.

[8] Madnani, N., and Chodorow, M.,2012.Reexamining Machine Translation Metrics for Paraphrase Identification. In proceedings of NAACL HLT 2012 (pp.182-190).

[9] Socher R. Huang E and Manning C.D 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for paraphrase detection, Science (pp.1-9).

[10] William, B.Dolan and Chris Brockett, "Automatically constructing a Corpus of Sentential Paraphrases", In Proceedings of IWP, 2005.

[11] Anand Kumar, M., Singh, S., Kavirajan, B., and Soman, K P. 2016. DPIL@FIRE2016: Overview of shared task on Detecting Paraphrases in Indian Languages, Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India.

[12] nlp.amrita.edu/dpil_cen/index.html#dataset

[13] Hsu, C., W., Chang, C. C.,and Lin, C. J. 2003. A practical guide to support vector clas- sification. Simon Fraser University, 8888 University Drive, Burnaby BC, Canada, V5A 1S6.

[14] Nigam, K., Lafferty, J., and McCallum, A. 1999. Using maximum entropy for text classification. In IJCAI-99 workshop on machine learning for information filtering: 1 (pp. 61–67).

[15] http://ltrc.iiit.ac.in/analyzer/tamil/

[16] Tang, H., Tan, S.,and Cheng, X. 2009. A survey on sentiment detection of reviews. Expert Systems with Applications, 36 (7), (pp.10760–10773).

[17] Zhang, Jian, Hai Zhao, Liqing Zhang and Bao-Liang Lu. "An empirical Comparative study on two large-scale Hierarchical Text classification approaches", International Journal of Computer processing of Languages, 2011.

[18] Mahalakshmi, S., Anand Kumar, M., and Soman, K.P 2015. Paraphrase detection for Tamil language using deep learning algorithm. Int. J. of Appld. Engg. Res., 10 (17), 13929-13934.