VOL. 10, NO. 8, MAY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

SUPERVISED METHODS FOR DOMAIN CLASSIFICATION OF TAMIL DOCUMENTS

Reshma U., Barathi Ganesh H. B., Anand Kumar M. and Soman K. P.
Centre for Excellence in Computational Engineering and Networking, Amrita Vishwa Vidyapeetham, Coimbatore, India
E-Mail: reshma.anata@gmail.com

ABSTRACT

The Era of digitization induces the need of domainclassification in both the on-line and off-line applications. The necessity of automatic text classification arises for utilizing it in diverse fields. Hence various methodologies like Machine Learningalgorithms were proposed to do the same. Here automatic document classification of Tamil documents have been proposed by considering the exponential growth of Tamil text documents in the form of unstructured data available as News, Encyclopedias, E-books, E-Governance, Social Media and much more. Max-Ent, CRF and SVM algorithms are used here to achieve more than 90 percentage average accuracy in both the sentence and document level classification of Tamil text documents. In this work Dinakarannewspaper dataset from EMILLE/CIIL Corpus has been utilized to experiment the ability of Machine Learning algorithms in Tamil domain classification.

Keywords: Tamil document classification, big data, Max-Ent classifier, CRF, SVM.

1. INTRODUCTION

The phenomenal growth of Tamil text documents in the web which is an electronic repository ensures the need of automatic domain classification to do the further processing and storage effectively. Availability of these documents in electronic repository increases exponentially on day to day basis. These large pool of documents exist in the form of newspapers, E-books, articles, digital libraries, E-Governance, social media and encyclopedias (Wikipedia 1, 97, 871 pages) [1]. Hence this big data has to be preprocessed for effective database management in distributed computation platform or cluster. This will reduce the computation while retrieving same for further applications like search engines and other information retrieval applications where domain classification plays a key role.

Domain classification is assigning a predefined set of classes to the documents. The intellectual way of classification involves abundant amount of manpower and time. Hence alternative methods like Machine Learning (ML) algorithms, rule based approaches and neural networks are proposed to do the automatic classification. By considering infeasibility of other methods, here ML algorithms were focused to do the automatic Tamil document classification. The automaticclassification in ML involves building a training model whichwill be trained by manually labeled documents to respective domains in the case of supervised learning and this model will be used to predict the unknown documents. Unsupervised learning will build the model without considering training data and will be omitted here by considering non-existence of key to evaluate the potential solution [2]. The way of building model introduces the various algorithms. Generally Generative model outperforms when compared to Discriminative model, hence staying with Discriminative model will lead better performance towards objective [3]. Some of the familiar algorithms are Support Vector Machine (SVM), Nave Bayes algorithm, Maximum Entropy models (Max-Ent), Hidden Marcov Models (HMM) and Conditional Random Fields (CRF).

Assigning the documents to multiple classes, extracting information from those and summarization of the same requires document classification. Some real world examplesare in classifying business names by industry, in differentiating a mail as spam and other, movie reviews and much more. Most of the ML algorithms were applied and experimented on English language and attained considerable accuracy in domain classification. Dennis Ramdass and Shreyes Seshasai successfully classified the MIT newspaper articles with NavieBayes, Max-Ent algorithms by having a probabilistic grammar parser and attained 77 percent accuracy [4]. By integrating labeled and unlabeled data Kamal Nigam et al reduced the classification error by 30 percent by combining Expectation-Maximization and a Nave Bayes algorithm [5]. Kamal Nigam et al proposed Max-Ent model with Gaussian prior to achieve the 78.8 percent average accuracy in classification of WebKB, Industry sector and Newsgroups datasets [6]. Comparison of Navie Bayes, Max-Ent, and SVM algorithms on sentiment classification of movie review done by Bo Pang et al by including various features and they concluded that SVM is best among them [7].

Even though various algorithms were proposed and experimented on English documents yet there is no standard algorithm fixed for domain classification. Hence this ensures that further research is required for domain classification problem. Automatic classification of Tamil documents is still on the research field since it is a morphologically richlanguage and agglutinative in nature. By considering these problems here automatic classification of Tamil documents were experimented on Max-Ent, CRF and SVM algorithms. Section 2 discusses the related work done in document classification on Indian languages. Section 3 deals with the mathematical model behind experimented supervised ML algorithms. Section 4 deals with the dataset, feature selection and experimental

results. Section 5 details about conclusion and future work towards the objective.

2. RELATED WORKS

There have been various models proposed for classification of documents in Indian languages like Kannada, Punjabi, Tamil, and Telugu. Consideration of Indian language is due to the probability sparseness caused by the agglutinative natureof languages. SVM seems to be the primary chosen method for Tamil language in linguistic applications like POS Tagging [8], Morphological Analysis [9], Word sense disambiguation [10] and Machine Translation [11].

Nidhi and Vishal Gupta proposed a hybrid model which attains 80 percent accuracy while classifying 184 Punjabi News Articles into 7 predefined classes. After stop word removal they proposed a combination of Nave Bayes and Ontology Based classification and came up a new method [12]. Narayana Swamyand Hanumanthappa utilized Naive Bayes, Decision tree and k-Nearest-Neighbor classifiers to classify the Kannada, Tamil and Telugu corpus with respect to the language [13]. Advanced Back Propagation Algorithm (ABPA) with Artificial Neural Network were applied on Tamil documents for classification and obtained 94.33 percent accuracy which shows greater performance than Vector Space Model (VSM) which yields only 90.33 percent accuracy. For this approach CIIL corpus was utilized by Kanimozhi and further preprocessing was done on it [14] . SVM was applied to do sentiment classification in Hindi done by Sneha Mulatakar. She utilized travel domain reviews for classifying whether the opinion is positive or negative [15]. Jayashree, Srikantamurthyand Anami utilized Nave Bayesian method for classification of Kannada language documents and the obtained results were improved using dimensionality reduction technique by applying stop word removal and restriction based on word occurrence [16]. K. Rajan *et al* classified 300 test documents into 5 classes by training VSM and Artificial Neural Network (ANN) with 100 documents. They showed that ANN as 93.3 percent accuracy compared to the VSM which yields 90.3 percent [17].This survey shows that lack of standard approach for the domain classification in Tamil language yields to experiment the available ML algorithms for classification of Tamil documents. Here both sentence level and document level classification is done using the Max-Ent, CRF and SVM ML approaches.

3. MATHEMATICAL BACKGROUND

A. Maximum entropy classifier

Maximum Entropy is a uniform model which assigns a probability distribution over the documents (D) and this will be constrained by having expected values (λ_i) of features. Here documents will be represented as a word count features (f_i) and by using this word count on class by class basis expected values are estimated from labeled training set. This can be mathematically expressed as (6),

$$logP(C|D,\lambda) = \sum_{(c,d)\epsilon(C,D)} log \frac{\exp(\Sigma_i, \lambda_i, f_i(c|d))}{\Sigma_{c'} \exp(\Sigma_i, \lambda_i, f_i(c|d))} \quad (1)$$

Denominator is a normalizing function. More precisely Max-Ent model is also like a CRF model except that here the model is built by chaining the local models instead of taking the entire sequence. Disadvantage of Max-Ent model are label biasing and casual competition bias. Algorithm for building a Max-Ent model can be expressed as follows:

a) Input
- Labeled documents
- Selection of features $f_i$from training set

b) Learning method
- Finding expected value and constraining conditional distribution with each feature.
- Optimization (Maximal Entropy) : Error reduction by L-BFGS algorithm (Computing λ_i)
- Smoothening (Gaussian Prior): Product over the Gaussian with all (λ_i)

c) Output of learning
- Probability distribution function with constrain
- Unlabeled data passed for prediction

Stanford classifier utilized for implementing Max-Ent classifier.

B. Conditional random fields

CRF is a combination of discriminative classification and graphical modeling to predict the multivariate output C (Classes) from observed D (Documents). This can be mathematically represented as [3],

$$argmaxP(C|D) \quad (2)$$

$$P(C|D,\lambda) = \frac{\exp(\Sigma_i \lambda_i f_i(c|d))}{\Sigma_{c'} \exp(\Sigma_i \lambda_i f_i(c'|d))} \quad (3)$$

C is set of cliques of the graph for given observed data D. Cliques are the edges and vertices of the graph i.e. labels and dependency between the labels. Denominator is a normalization factor computed as partition function.

Algorithm for building a CRF model can be expressed asfollows,

a) Input
- Labeled documents

b) Learning method
- Repeat until < Classification Rule >

- Add new features after error reduction (Maximizing y)
- Error reduction by L-BFGS algorithm (Computing $\lambda_i$parameter)

c) Output of learning
- Set of rules generated from observed data
- Unlabeled data passed for prediction

Advantage of CRF is it considers whole sequence to build conditional model rather than chaining of local models. It avoids the casual competition biasing problem but the only disadvantage is its takes time to train. The training time is directly proportional to the number of classes defined for classification. This CRF algorithm implemented using Machine Learning for Language Toolkit (MALLET) [19] [20].

C. Support vector machines

A non-probabilistic classifier model which maps the objective function points to the space. Hence this space will be separated by constraining the number of hyper planes according to availability of number of classes to categorize. Then test function is also mapped to the same space and respective classes will be predicted depending upon side of the gap they fall on which is formed by hyper planes. Can be expressed as [21],

$$c = sgn(f(d, w, b)) \quad (4)$$

$$f(d, w, b) =< w.d > +b \quad (5)$$

The two elements of SVM are weight vector (w) and the bias (b) which is the distance of hyper plane from origin. Algorithm for building SVM model can be expressed as,

a) Input
- Labeled documents
- Feature selection

b) Learning method
- Conversion of multiclass problem into binary classification problem
- Map the data points to the space with hyper plane
- Optimize to have large margin

c) Output of learning
- Model with training data in hyper plane separated space
- Test data to fall on respective gap in the same space

The main disadvantage of the SVM model is that it takes longer time to test with the larger feature space. We have used LIBSVM to build the SVM classifier model [22].

4. EXPERIMENTAL WORKS AND RESULTS

A. Dataset

A standard dataset Dinakaran newspaper articles and news from EMILLE/CIIL Corpus was utilized to experiment the Tamil document classification [23]. Precisely from the multilingual corpus in Tamil three domains from Dinakaran newspaper taken for the experiment. This dataset included news and articles about cinema, politics and sports [24]. After removal of English texts which was present in each document to explain details about it, the documents were manually tagged depending upon their content. Further details about training documents are given in the Table 1 and 2.

Table-1. Details of training set.

Domain	Total no. of documents	Total no. of sentences	Avg no. of sentences/ document	Avg no. of words/ sentence
Cinema	843	62889	74.6	11.91
Politics	1024	271487	260.54	12.51
Sports	990	68335	69.02	12.32
Total	2875	402711	140.07	12.38

Table-2. Details of test set.

	Total no. of documents	Total no. of sentences	Avg no. of sentences/ document	Avg no. of words/ sentence
Cinema	91	7441	81.76	11.47
Politics	90	20945	232.72	12.26
Sports	100	8120	81.20	11.38
Total	281	36506	129.91	11.91

Dataset experimented for classification contains 3156 documents and 3 domains. Among them 281 randomly picked documents were used as test documents and remaining 2875 used for training. Average sentence per documents is comparatively high (261) in politics documents but average words per sentence is well balanced. This led to classify the documents in sentence level.

B. Experiment

Algorithms were trained in document level and also sentence level by assigning each classifier with two models. In early stages these algorithms were experimented with smaller datasets to check the performance ability towards various biasing problems [6]. Training and testing time consumed by all these algorithms were averagely equal in this case. Then the standard EMILLE/CIIL dataset was trained and tested to experiment the performance of the ML algorithms. While training Max-Ent classifier to increase the accuracy level various n-gram level feature selections were performed. While utilizing Max-Ent Gaussian prior for smoothening σ = 0:154 value fixed to achieve greater accuracy [25]. SVM classifier was trained with ineffective feature selection and hybrid feature selection methods to reduce the probability sparseness [26]. The Kernel used in SVM is the Gaussian radial basis function and the soft margin parameter C, gamma were selected by a grid search method.

C. Results and observations

Though this work was experimented with large number of dataset, the problem of probability sparseness has not raised due to the formulation of utilized Max-Ent and CRF algorithms. But SVM shows degraded performance towards probability sparseness problem while doing classification in sentence level. This is confirmed by observing the outcomes of respective tools experimented. CRF based classifier model avoids the casual competition biasing which was confirmed by observing classification accuracy in sentence level and document level classification. Max-Ent classifier provides greater accuracy in document level classification since the objective function is smoothened by both the Gaussian prior and the number of features [25]. As the number of features gets increased by having n-gram feature selection in Max-Ent classifier, there is a high degree of decrease in misclassification in sentence level classification, yet it can be observed that it suffers from the label biasing and problem of non- smoothening.

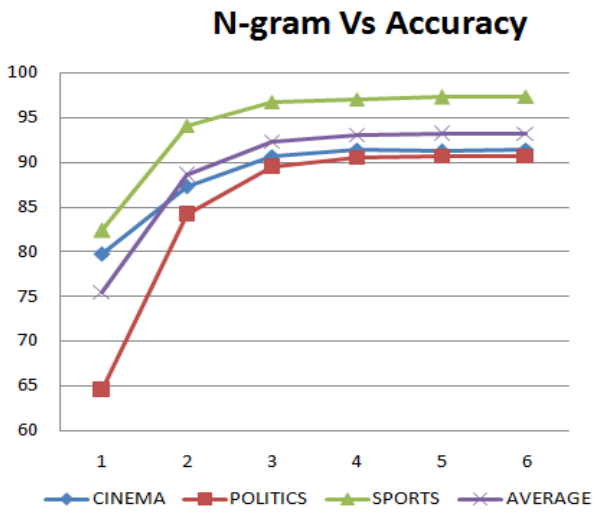

Figure-1. Max-Ent sentence level classification accuracy per class.

SVM model with ineffective feature selection provides varying accuracy during classification. With the cross validation and hybrid feature selection i.e. based on content this method provides greater accuracy compared to the previous one experimented [26]. In sentence level classification the space related to the feature i.e. Hilbert Space is sparse due to the size of the dataset which yields to degraded performance. This can be solved by including linguistic methods like stemming in preprocessing stage and dimension reduction methods before doing the classification. Results observed for document classification and sentence classification are shown in the Table 2 and 4.

Table-3. Average accuracy of documents classification.

Domain	MALLET	STANFORD	SVM
Cinema	98.9011	100	98.9
Politics	100	100	82.2
Sports	100	100	99.9
Total	99.6441	100	93.7

Table-4. Average accuracy of sentence classification.

Domain	MALLET	STANFORD	SVM
Cinema	98.9865	91.4	88.08
Politics	99.9951	90.8	79.83
Sports	100	97.3	94.69
Total	99.9945	93.1667	87.533

From the observations it is to be noted that Sports and Cinema are in a lead when compared to Politics. This is because the average numbers of sentences in the former are less than the latter which ensures the reduction in probability sparseness.

5. CONCLUSION AND FUTURE WORK

The Max-Ent, CRF and SVM algorithm were experimented to classify Tamil documents and attained very good accuracy as shown in Section IV. Among them CRF shows excellent performance in both the document and sentence level classification with same time consumed by other tools that were experimented. There was no misclassification in document level classification while using Max-Ent classifier but biasing problem arises in sentence level classification. SVM shows reduced performance comparing other methods because of probability sparseness in feature. As the size of the dataset and number of classes gets increased it is obvious that the classification becomes computationally heavy along with escalated time. Hence the future objective will be to solve this Big Data problem in distributed computation platform by implementing these algorithms in Apache Mahout and SPARK.

REFERENCES

[1] Wikipedia [Online]. Available: http://ta.wikipedia.org/ [Accessed: 09-Mar-2015].

[2] B. Baharudin, L. H. Lee, and K. Khan, “A review of machine learning algorithms for text-documents classification,” J. Adv. Inf. Technol., vol. 1, no. 1, pp. 4-20, 2010.

[3] J. Lafferty, A. McCallum, and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” 2001.

[4] D. Ramdass and S. Seshasai, “Document classification for newspaper articles,” 2009.

[5] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, “Text classification from labeled and unlabeled documents using EM,” Mach. Learn., vol. 39, no. 2-3, pp. 103-134, 2000.

[6] K. Nigam, J. Lafferty, and A. McCallum, “Using maximum entropy for text classification,” in IJCAI-99 workshop on machine learning for information filtering, 1999, vol. 1, pp. 61–67.

[7] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002, pp. 79–86.

[8] K. P. Soman and others, “Morphology based prototype statistical machine translation system for English to Tamil language,” 2013.

[9] V. Dhanalakshmi, M. Anand Kumar, R. U. Rekha, C. Arun Kumar, K. P. Soman, and S. Rajendran, “Morphological Analyzer for Agglutinative Languages Using Machine Learning Approaches,” in Advances in Recent Technologies in Communication and Computing, 2009. ARTCom’09. International Conference on, 2009, pp. 433-435.

[10] M. Anand Kumar, K.P. Soman, and S. Rajendran, "Tamil word sense disambiguation using support vector machines with rich features," International Journal of Applied Engineering Research (IJAER), vol. 9, pp. 7609-7620, Number 20 (2014).

[11] M.Anand Kumar, V.Dhanalakshmi, K.P. Soman, and R. S, "Factored statistical machine translation system for english to tamil language," Journal of Social Sciences and Humanities, pp. 1045-1061, 2014.

[12] V. G. Nidhi, “Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach,” in Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, COLING, 2012, pp. 109-122.

[13] M. N. Swamy, M. Hanumanthappa, and N. M. Jyothi, “Indian Language Text Representation and Categorization Using Supervised Learning Algorithm,” in Intelligent Computing Applications (ICICA), 2014 International Conference on, 2014, pp. 406-410.

[14] S.Kanimozhi, "Web based classification of tamil documents using abpa," International Journal of Scientific and Engineering Research, vol. 3, May 2012.

[15] S.Mulatkar, "Sentiment classification in hindi," INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH. vol.3, pp. 204 - 206, 2014.

[16] R. Jayashree, K. Srikantamurthy, and S. A. Basavaraj, “Suitability of Naïve Bayesian Methods for Paragraph Level Text Classification in the Kannada Language using Dimensionality Reduction,” Int. J. Artif. Intell. Appl. IJAIA, vol. 4, no. 5, pp. 121-131, 2013.

[17] K. Rajan, V. Ramalingam, M. Ganesan, S. Palanivel, and B. Palaniappan, “Automatic classification of Tamil documents using vector space model and artificial neural network,” Expert Syst. Appl., vol. 36, no. 8, pp. 10914-10918, 2009.

[18] Stanford [Online]. Available: http.nlp.stanford.edu/software/classifier [Accessed: 09-Mar-2015].

[19] Mallet [Online]. Available: http.mallet.cs.umass.edu [Accessed: 09-Mar-2015].

A. K. McCallum, “MALLET: A Machine Learning for language toolkit,” 2002.

[20] J. Gimenez and L. Marquez, "SVMTool Technical Manual v1. 3," 2006.

[21] LIBSVM [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm/ [Accessed: 09-Mar-2015].

[22] P.Baker, A. Hardie, T. McEnery, H. Cunningham, and R.J. Gaizauskas, "Emille, a 67 - million word corpus of indic languages: Dta collection, mark-up and Harmonisation," in LREC, 2002.

[23] EMILLE/CIIL [Online]. Available: http://metashare.elda.org/repository/browse/the-emilleciil-corpus/ [Accessed: 09-Mar-2015].

[24] S. F. Chen and R. Rosenfeld, "A Gaussian prior for smoothing maximum entropy models," DTIC Document, 1999.

[25] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin, "LIBLINEAR: A library for large linear classification," J. Mach. Learn. Res., vol. 9, pp. 1871-1874, 2008.