Impact Factor (SJIF): 5.301

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 6, Issue 4, April-2019

Tamil Document Classification and Topic Identification

Professor Dr. S. Saraswathi
Department of Information Technology
Pondicherry Engineering College
Puducherry, India
swathi@pec.edu

R. Santhiya
Department of Information Technology
Pondicherry Engineering College
Puducherry, India
santhiyaramdoss@pec.edu

B.L. Sanjeev Prasad
Department of Information Technology
Pondicherry Engineering College
Puducherry, India
sanjeev.bl97@gmail.com

G. Gnanaprakasam
Department of Information Technology
Pondicherry Engineering College
Puducherry, India
prakashyoshik@gmail.com

***Abstract*—Nowadays the number of documents in electronic form is huge and grows day by day with the rapid development of internet. It is extremely important to organize the documents according to the topic because huge number of documents are available nowadays. Commonly, this can be achieved by using classification techniques. Document classification is an important tool for applications such as web search engines. This proposal deals with classification of Tamil documents. Classification is a supervised learning process that organizes documents or text files into distinct groups. Stop words will be removed from the input text document to decrease the size of the document to be processed. In this project we have used naïve bayes algorithm, support vector machine algorithm to classify the documents and latent drichelet allocation for topic modelling.**

Keywords: text classification, naïve bayes, svm, LDA, machine learning.

1.INTRODUCTION

Text Classification is one of the widely used natural language processing task in different business problems. The major goal of text classification is to automatically classify the text documents into different categories. Some examples of text classification are sentimental analysis from social media, Detection of spam and non-spam emails and Classification of news articles into defined topics. Text Classification is a supervised machine learning task since a labelled dataset is used to train the classifier.

Many studies are available for classification of english documents or news articles. In this paper we do tamil news article classification. Various machine learning algorithms were used for document classification. We used naïve bayes algorithm and support vector machine to classify tamil news articles. For topic modelling we use Latent Drichlet Allocation method. In order to experiment with the models for Tamil language, we first had to build a new data set.

2. LITERATURE SURVEY

Automatic text classification is the task of assigning predefined categories to unclassified text documents. When an unknown document is given to the system it automatically assigns it the category which is most appropriate. The classification of textual data has more significance in effective document management. Nowadays amount of available online information increased and makes it difficult to manage and retrieve those documents without proper classification. There are two main approaches for document classification namely Supervised and Unsupervised learning. In supervised learning, the classifier is first trained with a set of training data in which documents are labeled with their category, and then the trained system is used for classifying new documents. The unsupervised learning is mainly based on clustering.

One of the popular approaches in supervised learning is the VSM [1]. This is based on assigning weights proportional to the document frequencies of a word in the current category as against the rest of the categories. The VSM represents the text documents as vectors where each distinct word is a separate component. It assigns some weight to each component of the vector depending on the importance of that component Before any digital text can be processed by a machine learning (ML) classifier, a mapping must be performed on the data that is somehow able to represent the required characteristics or 'features' into a more compact and computationally appropriate form. The most established and well-known method of the document weighting approaches is the vector space model. The most common feature used in text classification is tf-idf (term frequency–inverse document frequency) measure. TFIDF gives a weighting or relevance of how important a word is to a document. In Vector space model a document di is represented by a set of words (t1; t2 ...tn) wherein each tj is a word that appears in the text document di, and n denotes the total number of various words in the index used to identify the meaning of the text document. Word tj has a corresponding weight wi calculated as a combination of the statistics term frequency TF(tj, di) and inverse document frequency IDF(tj). The input Files are to be pre-processed before applying the clustering task, in order to reduce the size of the document and problem space. The pre-processing step gets the input document as input and all the data items are represented as vectors [2]. In English, the document is reduced to one third of the document size. In Tamil, the document size is reduced to only 12% to 18%. Removal of these stop words from the input document can reduce the noise in the file and increases the computational efficiency of the system.

Neural networks are networks of nodes, which are mathematical models of biological neurons [1]. These networks have self learning capability, they are fault tolerant and noise immune, and have applications in system identification, pattern recognition, classification, image processing and natural language processing, etc. Classification decision for any document of reasonable size is based on the combined evidence from many sources. Each word is a source to classify a document. There are many ANN models available namely backpropagation networks, counter propagation networks etc. Backpropagation ANNs are used for natural language processing tasks such as syntax analysis, language studies, text classification, a three-layer feedforward neural network with hyperbolic tangent (tanh) activation function in the hidden layer, followed by a linear output layer is employed. The neural network is trained with backpropagation algorithm. The inputs are the components of the document vector, and the outputs are the document categories. It is observed that NN models are effective in classifying Tamil documents. The performance of NN is better for more representative collection [3].

In text classification tasks, documents are characterized by the words that appear in them. Thus, one of the simplest ways to apply machine learning to text classification is to treat each word as a boolean variable. This is the first statistical language model called multi-variate Bernoulli naive Bayes (BNB) model. BNB assumes that a document is represented by a vector of binary feature variables indicating which words occur or not in the document, and thus ignores the information of the number of times a it occurs in a document. To overcome the shortcoming confronting BNB, the multinomial naive Bayes (MNB) model is proposed by capturing the information of the number of times a word occurs in a document. However, one systemic problem confronting MNB is that when one class has more training documents than the others, MNB selects poor weights for the decision boundary. This is due to an under-studied bias effect that shrinks weights for classes with few training documents. To balance the amount of training documents used per estimate and to deal with skewed training data, a complement class version of MNB, called complement naive Bayes (CNB) is proposed.

Although recent work in supervised learning has shown that these naive Bayes text classifiers, such as MNB, CNB and OVA, have achieved remarkable classification performance, all of them again make a similar naive Bayes assumption: that the probability of each word event in a document is independent of the word's context and position in the document. However, it is obvious that this independence assumption required by them is rarely true, which would harm their performance in the real-world text classification applications with complex dependencies among words. In order to weaken this independence assumption required by them, many enhancements to naive Bayes text classifiers have been proposed [5]. Deep feature weighting (DFW) to improve three state-of-the-art naive Bayes text classifiers multinomial naive Bayes (MNB), complement naive Bayes (CNB) and the one-versus-all-but-one model (OVA).

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. Generally Generative model performs well when compared to Discriminative model, therefore staying with discriminative model will give better performance towards objective [6]. Some of the familiar

algorithms are Support Vector Machine (SVM), Naive Bayes algorithm, Maximum Entropy models (Max-Ent). Assigning the documents to multiple classes and extracting information from those and summarizing the same requires document classification. Some real-world examples are in classifying business names by industry, in differentiating a mail as spam and other, movie reviews etc. Most of the ML algorithms were applied on English language and attained acceptable accuracy in domain classification. Dennis Ramdass and Shreyes Seshasai classified the MIT newspaper articles with NavieBayes, Max-Ent algorithms by having a probabilistic grammar parser and attained 77% accuracy. Classification of Tamil documents is still on the research field because it is a morphologically richlanguage and agglutinative in nature [6]. By considering these problems here automatic classification of Tamil documents was experimented on Max-Ent, CRF and SVM algorithms.

3.OVERVIEW OF THE SYSTEM

An end-to-end text classification pipeline is composed of three main components

3.1 Dataset Preparation

The first step is the Dataset Preparation step which includes the process of loading a dataset and performing pre-processing to remove the dataset. The dataset is then splitted into train and validation sets.

3.2 Feature Engineering

In feature engineering the raw dataset is transformed into flat features which can be used in a machine learning model also called as process of creating new features from the existing data. Here we used tf-idf vectors as features. This score represents the relative importance of a term in the document and the entire corpus. It is composed by two terms: the first one computes the normalized Term Frequency (TF), the second one is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

$$\text{TF(t)} = \frac{(Number\ of\ times\ term\ t\ appears\ in\ a\ document)}{(Total\ number\ of\ terms\ in\ the\ document)}$$

$$\text{IDF(t)} = \log \frac{(Total\ number\ of\ documents)}{(Number\ of\ documents\ with\ term\ t\ in\ it)}$$

TF-IDF Vectors can be generated at different levels of input tokens (words, characters, n-grams).

3.3 Model Building

The last step is the Model Building step in which a machine learning model is trained on a labelled dataset. There are many different machine learning models available which can be used to train a final model. We used two main machine learning algorithms for classification such as naïve bayes and support vector machine. Then we compare the model's performance based on accuracy, precision, recall, and F-score.

4. PROPOSED METHODOLOGY

Text classification is used to classify the documents according to the domain. In this paper we do tamil news classification for the predefined categories using naïve bayes, svm algorithm and topic modelling using Latent drichlet allocation method.

4.1 Dataset

The dataset used in this study is a collection of news articles from thinamalar website of different categories such as cricket, football, hockey, state news, central news.

Category	No of Documents
cricket	50
hockey	50
football	50
Indian Politics	50
State Politics	50

Figure. 4.1.1 Distribution of Datasets

4.2 Preprocessing

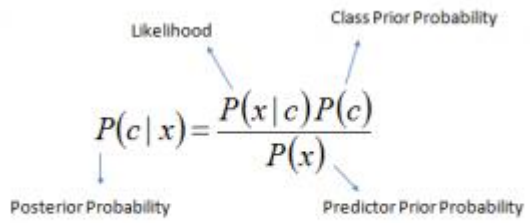
The given tamil documents will be preprocessed to remove the stop words which will reduce the size of the document. Here we use morphological analyser (Tamil shallow parser) to find the POS (parts of speech) of each word. Using those POS, we will remove the stop word from the tamil news articles.

4.3 Testing and Training

Since we are using supervised learning, we must train the classifier. We are dividing our dataset into two parts, 80% of the news article is used for training the model, remaining 20% is used for testing phase. Here we are using two machine learning algorithm such as naïve bayes and support vector machine.

A) Naïve Bayes

It is a classification technique based on Bayes's Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a feature in a class is not related to the presence of some other feature. Naive Bayes model is easy to implement and particularly useful for very large data sets. Naive Bayes is known to performs well even for highly sophisticated classification methods. For example, a fruit may be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other, all these properties independently contribute to the probability that this fruit is an apple and because of this it is known as 'Naive'. Bayes theorem paves a way for calculating posterior probability P(c|x) from P(c), P(x) and P(x|c)



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Where,

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class.*
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

B) Support Vector Machine

Support vector machines are supervised learning models with machine learning algorithms that analyze data which is used for classification and regression analysis. In a set of given training examples, each marked as belonging to one or the other of any two categories, it builds a model that assigns new examples to one category or the other and makes it a non-probabilistic binary. SVM represent the given examples as points in space as mapped points so that the examples of the separate categories are divided by a gap that is as wide as possible. Then new examples are mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

C) Latent Drichlet Allocation

Topic Modelling is a method to identify the collection of words (called a topic) from a group of documents that contains best information in the group[4]. Latent Dirichlet Allocation method can be used for generating Topic Modelling Features.LDA is technique which is based on matrix factorization technique. The given dataset or corpus will be converted into a document term matrix. It converts the document-term Matrix into two lower dimensional matrices M1 and M2.M1 is a document-topics matrix and M2 is a topic-terms matrix with dimensions (N, K) and (K, M) respectively, where N is the number of documents, K is the number of topics and M is the vocabulary size.It Iterates through each word "w" for each document "d" and tries to regulate the present topic – word assignment with a brand new assignment. A new topic "k" is assigned to word "w" with a chance P that is a product of two chances p1 and p2. Two probabilities p1 and p2 are calculated for every topic. P1 – p (topic t / document d) = the proportion of words in document d that are currently assigned to topic t. P2 – p (word w / topic t) = the proportion of assignments to topic t over all documents that come from this word w.The current topic-word assignment is replaced with a new topic with the probability, product of p1 and p2. LDA assumes that all the existing word-topic assignments except the current word are correct. This is basically the probability that topic t generated word w, therefore it makes sense to regulate the present word's topic with new probability. After several iterations, a steady state is achieved where the document topic and topic term distributions are good.

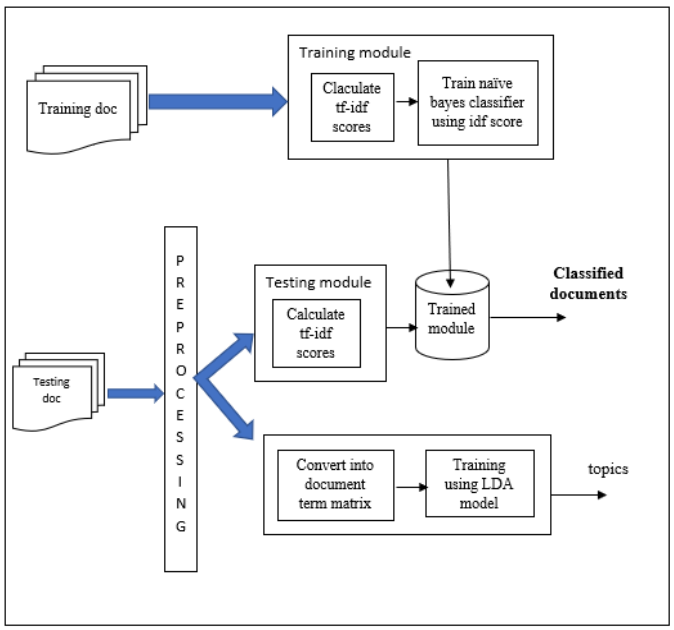


Figure. 4.3.1 Architecture Diagram for Text Classification and Topic Identification

4.4 Result Analysis

We used our test set for evalution of our models.In order to see the results we computed confusion matrix where number of correct prediction will be on the diagonal of the matrix, while the number of wrong predictions will be on the other position. We then calculated performance metrics precision, recall, f-measure and overall accuracy. These measures are calculated by using the

true positives TP (observation that are correctly predicted by the model) and false positives (observation that are classified as one category which is not belong to that category) false negative (observations that should have been classified as one category, but were not). Precision measures the proportion between correct predictions and the total predicted observation in the category. It is computed by the formula TP/(TP+FP). Recall measures the proportion between correct predictions and the total number of articles of the category. It is computed by the formula TP/(TP+FN). F-score is the harmonic average of precision and recall. It is computed by the formula 2∗(precision∗recall)/(precision+recall. Accuracy is the proportion of correct predictions and the total number of observations. We are getting 72% accuracy in Naïve Bayes and 81% accuracy in SVM.

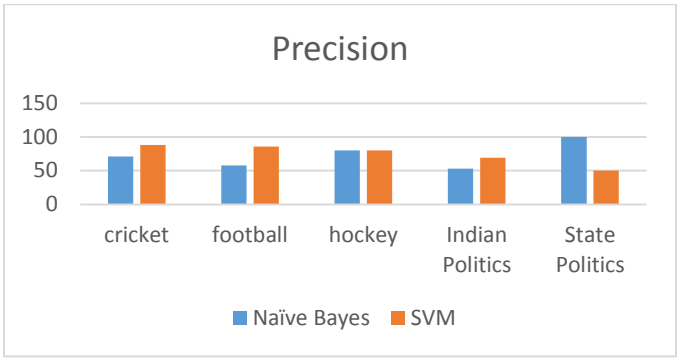


Figure. 4.4.1 Precision Comparison graph

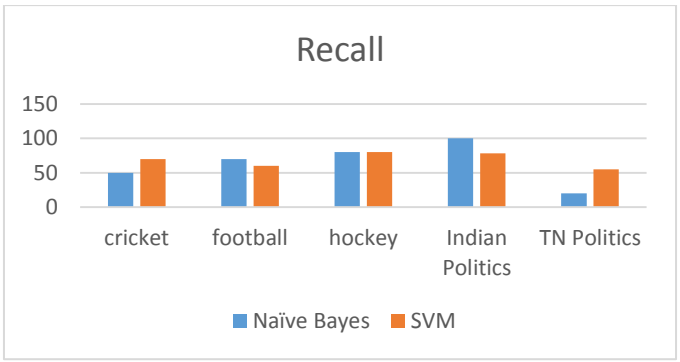


Figure. 4.4.2 Recall Comparison graph

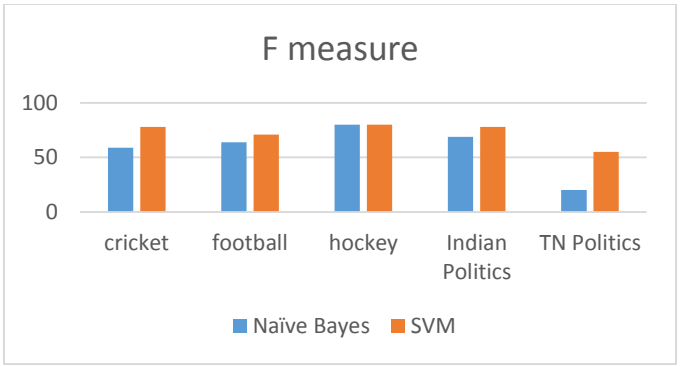


Figure. 4.4.3 F measure Comparison graph

5. CONCLUSION

In our proposed work, we explore the problem of tamil news article classification using two major machine learning algorithms such as naïve bayes and SVM which classify the news articles into different predefined categories (central, cricket, football, hockey, state). After that using LDA we do topic modelling which gives the collection of words which defines the domain it belongs to. In our future work we try to classify the news articles based on the images in it.

REFERNCES

[1] Rajan. K, Ramalingam. V, Ganesan. M, Palanivel. S & Palaniappan. B. "Automatic classification of Tamil documents using vector space model and artificial neural network" Expert Systems with Applications, Elsevier, pp 10914–10918, 2009.

[2] Syed Sabir Mohamed and Shanmugasundaram Hariharan. "Experiments on document clustering in Tamil language" ARPN Journal of Engineering and Applied Sciences, vol. 13, pp 3564 -3569, May 2018.

[3] Pooja Bolaj, Sharvari Govilkar "A Survey on Text Categorization Techniques for Indian Regional Languages" International Journal of Computer Science and Information Technologies, Vol. 7 (2), pp 480-483, 2016.

[4] Vijay Singh, Mangey Ram, Bhasker Pant "Identification of Zonal-wise Passenger's Issues in Indian Railways using Latent Drichlet Allocation (LDA): A Sentiment Analysis Approach on Tweets" Frontiers in Information Systems, Vol 2, pp 265-276,2018.

[5] Liangxiao Jiang, Chaoqun Li, Shasha Wang, Lungan Zhang "Deep feature weighting for naive Bayes and its application to text classification" Engineering Applications of Artificial Intelligence, Elsevier, pp 26–39,2016.

[6] Reshma U, Barathi Ganesh H. B, Anand Kumar M. and Soman K. P. "Supervised Methods for Domain Classification of Tamil Documents" ARPN Journal of Engineering and Applied Sciences, Vol. 10, May 2015.

[7] A. Vijaya Kathiravan, P. Kalaiyarasi "Sentence-Similarity Based Document Clustering Using Birch Algorithm" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 5, May 2015.