
VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

# School of Computer Science and Engineering

## Digital Assignment - 3

**Programme : B Tech**

**Course Title : Natural Language Processing**

**Course Code : CSE 4022**

**Slot : E1**

**Title : Tamil News Classification**

**Team Members:**

**SHRIKUMARAN P A | 20BCE1082**

**R. SHRUTHI | 20BCE1375**

**DASARI.NAGAVENI | 20BCE1715**

**Faculty: Dr. ILAKIYASELVAN N** **Sign:**

**Date:**

# Tamil News Categorisation


## Abstract

This paper deals with the methods implemented to develop a Tamil News Classification model. Tokenization is a widely used method in Natural Language Processing to split a sentence or paragraph into smaller units with separate meaning attached to it. Here we collected the news in Tamil and splitted it into training and testing data. Here, the sentences from each category are grouped and the individual tokens are extracted from it. The test data can be tokenized and matched with the words in each category. The category with which the maximum number of words have matched is the correct category.. Also KNN algorithm was tried for the classification purpose but yielded very few results with a low accuracy. The tokenization method however yielded a good accuracy score. One of the proposed models can only classify one news at a time and has a higher time complexity, whereas the other one has a lesser time complexity and can classify more data at a time. The tokenisation model can yield more accuracy than KNN or the other proposed model.


## Introduction

News is the day to day information we receive about the happenings all over the world. News and media play a major role in today's society. No day passes without hearing news. The news is taken by people through TV, Newspaper, and also social media. So categorisation of such news is important. Some people might like to know the news of happenings all over the world while some like to keep themselves updated on sports news. There is a definite structure and procedure for writing various forms of news and their headlines. The words used must be honourable and not abusive in any way. So the words used for each category of news is different separated by stopwords and other conjunctions. In English the common stopwords are "the", "and", and so on. Similarly, in other languages also there are numerous stop words that are used to form a sentence. So categorisation of the news can be done by removing the stopwords and the punctuations in the sentences. Natural Language Processing is used for this purpose of categorisation of the news by parsing the news for main words in the news. This paper discusses the methods that can be used for categorising news in Tamil. The following sections will discuss the related work, future scope, and the methodology used to build the efficient models for Tamil news classification. The reader has to have a prerequisite knowledge in implementing machine learning models and estimating their precision from the range of obtained correct classifications. Some knowledge in Tamil is also required to check if the classification is correct or not.

# Related Work

| Paper | Description | Methodology | Advantage and disadvantage |
|---|---|---|---|
| Comparative Analysis of Tamil and English News Text Summarization Using Text Rank Algorithm | Tamil news is summarised into a small paragraph with the use of Text Rank Algorithm. | Each sentence in the document is assigned a particular rank or ranking using the text rank algorithm. The value of a sentence in the document is indicated by its rank. This suggests that the higher the sentence's rank, the more relevant the sentence is. Sentences with a rank greater than or equal to a certain threshold are considered for summary generation | It uses Text mining and is very efficient for stop words removal, tokenization and stemming purposes. |
| Topic categorization of Tamil News Articles using PreTrained Word2Vec Embeddings with Convolutional Neural Network | Convolution neural network can be used to categorise the tamil news | Input for convolutions are fed from the embedding layer. Three type convolutions 3×3, 4×4, 5×5 are used through which features are formalised. The features are combined using a merge layer. After the merge, three dense layers i.e. fully connected networks are given with the output of the merge layer. In the end of connection using a | CNN with pre-trained embeddings for tamil language gives better results compared to SVM and NB trained with TFIDF feature vectors. |

| | | | |
|---|---|---|---|
| | | sigmoid activation function the classification is done. | |
| Text Summarization for Tamil Online Sports News Using NLP | Tamil Sports news are summarised using various methods. | Raw Tamil sports news will be fed as input into the preprocessing step, In this step data will be tokenized into sentence and words, furthermore regular expressions will be used to detect the character structures such as punctuation, time, date numbers, etc. After that stop words like “gy (pala)”, “ xU (oru)”, “vd;W (enru)”, etc... will be removed by analyzing the frequency of the tokens, in here a list of stop words in Tamil language will be used for this purpose. Thereafter PoS tag will be attached to the remaining tokens such as verb, noun, etc. | |

| | | | |
|---|---|---|---|
| News text classification model based on topic model | This paper focuses on news text classification models based on Latent Dirichlet Allocation (LDA). | Latent Dirichlet Allocation (LDA) is a kind of topic model algorithm based on probability model the data is sorted and preprocessed, and the topic distribution of the training set of news texts is obtained through the LDA model,it chooses the topics trail n_X of news text as the feature of the news text Both the train n_X and for the class label trai n_Y as the input value of the Softmax Regression algorithm, start modelling, training and get the news text classifier. | LDA topic model, which is mainly based on the latent topic information of the text. In the model, we make news text changed from VSM to the topic vector. |

# Methodology

## Dataset Description

The dataset has been extracted with the features “News in English”, “News in Tamil”,”Category in English” and “Category in Tamil”. The dataset has been preprocessed for punctuations, stopwords and has been removed of all punctuations and stopwords. The duplicate records have also been removed. Stopwords in Tamil include '□□□□□', '□□□□□', '□□□□', '□□□□□□□□□□' have been removed. The punctuations such as ,./;:’-_!@# have been removed and only the tokens have been extracted to train the models.

**Pseudocode for Preprocessing:**

For removing punctuations:

1. Load the data
2. For each sentence in the field “News in Tamil”:
    1. Remove the punctuation.
    2. Write it back in the dataset.
3. End.

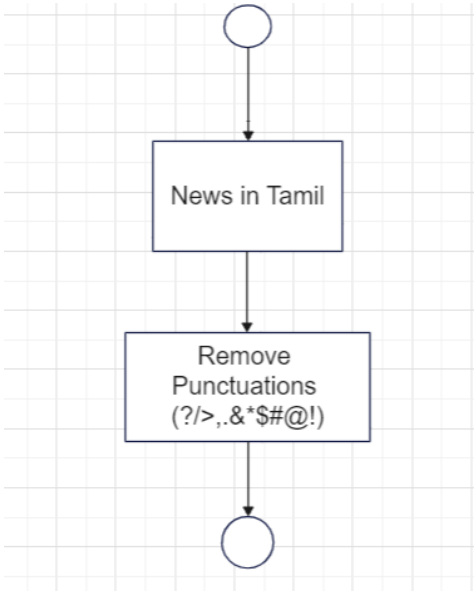


Fig.1

**For removing Stopwords and Tokenisation:**

1. Load the data that has been removed of punctuations
2. For each sentence in the field “News in Tamil”:
    1. Remove the stopwords.
    2. Extract each word to tokens
3. End.

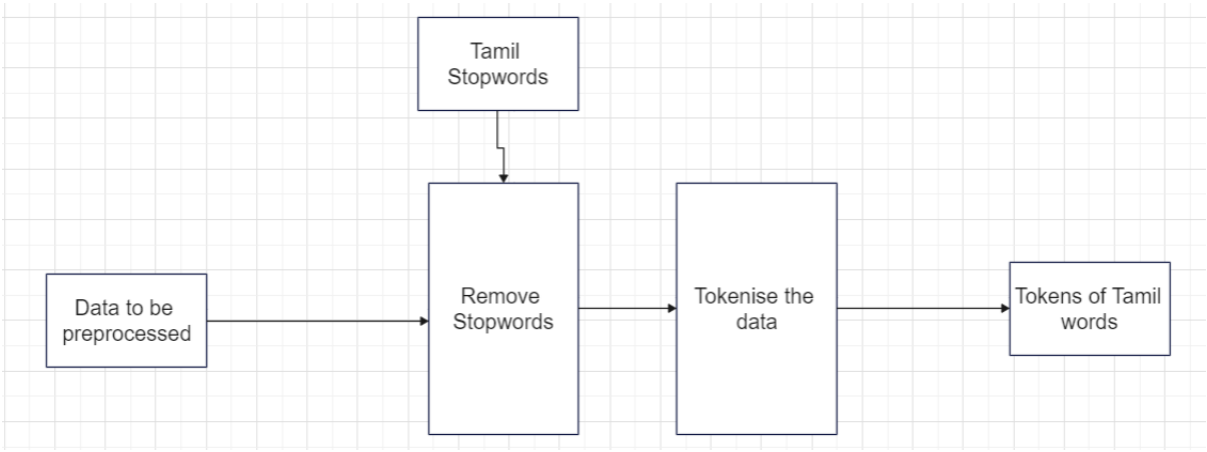


Fig.2

## Proposed Methodology

Three methods have been proposed to categorise the Tamil News. The first method involves taking a single record from a test data and preprocessing it. The tokens extracted are matched with the entire training data and for each data the number of matching tokens are counted. The category that has matched with the maximum number of words is the category of the text data. This is a brute force method and can only categorise one news at a time.

Pseudocode:

1. Take test data ‘X’.
2. Tokenise ‘X’.
3. For each record ‘Y’ in training data, do:
    1. Tokenise ‘Y’.
    2. Check if any word in ‘X’ is in ‘Y’.
    3. If a word is present, do:
        1. Increment count
    4. If count exceeds maximum value, change maximum value to count
4. Return category of maximum.

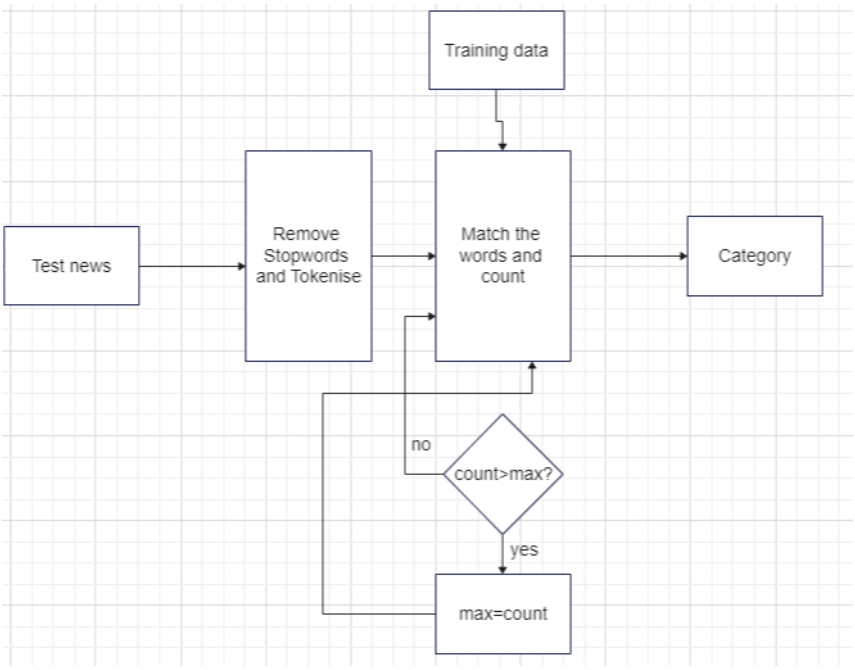


Fig.3

## The second method is the use of KNN algorithm:

The K - nearest neighbours algorithm can be used to predict the category of the Tamil News by taking the count of matching words in each and attaching them to the dataset as a separate column. Then the dataset is trained for 3 nearest neighbours and the category is predicted. The output will be the category of the test data. Since this uses 3 Nearest neighbour values the categorisation might slightly vary in accuracy.

Pseudocode:

1. Start
2. Tokenize test data “X”.
3. For each “News in Tamil” in training data, do:
    1. Find the count of matching tokens.
    2. Add it to the column
4. Train the model for 3 neighbours.
5. Return the category
6. End.

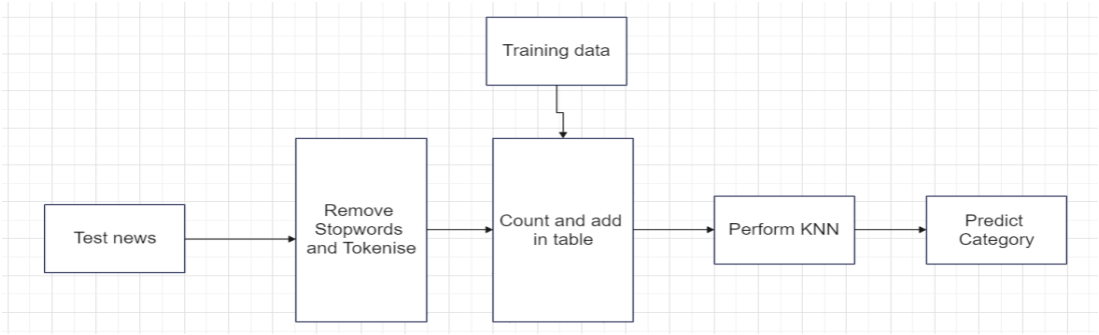


Fig.4

## Third Method:

This method involves keeping a stock of the words in each category stored previously. The words from the test data were matched with each category and the category that matched the most is the classified category. The most prominent words in each category were extracted from the training data and were stored in separate variables to check with the test data. This is a fast approach and purely uses the concept of tokenization. It can classify more records at a moment of time but is very less accuracy. It is because there is only one check in the number of words. It is almost an update of the first method where the difference is that instead of searching each record, only the words of each category are searched and the category is predicted.

Pseudocode for categorising the words:

1. Initialise 5 variables, one for each category to 0.
2. For each record in training data, do:
    1. Check the category in tamil
    2. According to the category, tokenize the news and store the words in the correct variable.
3. For the test data, check with the categorised word and the category with maximum.
4. Return maximum category.

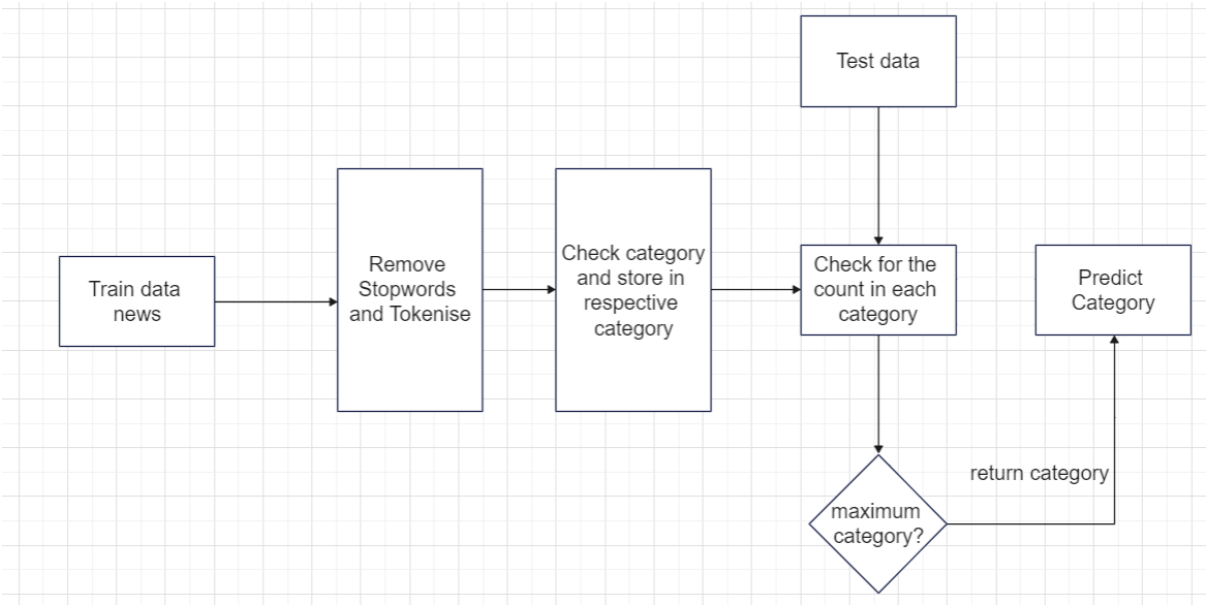


Fig.5

## Results and Discussions

The methods were checked with the test data. The performance metric used for this purpose is accuracy. The results are tabulated in the below table:

| **Method** | **Accuracy** | **Time complexity** |
|---|---|---|
| Normal Tokenization | 0.74 | O(n*n) |
| KNN | 0.58 | O(n*n) |
| Categorisation by grouping words | 0.1 | O(n) |

Majority of the research papers have performed only Tamil news summarisation but this is different as this paper purely focusses on News Categorisation. As the results show us, more than KNN, or grouping, normal tokenization works better for Tamil News Categorisation. The future scope of this project can be extended towards developing models for categorisation of news in other regional languages as well as develop models with less time and space complexity.

## Conclusions

From the table, we infer that normal tokenization with pre -trained embeddings for tamil language gives better results compared to KNN and grouping trained with count words feature vectors. The reason for this may be due to the occurrence of new tokens in politics test data than in cinema and sports. In future, we can accommodate the same methodology for other social media data as done for news web data. Also, sentiment of the data can be analysed after topic categorization. This will help in further improving the standard of the project.

## References


Santhanam, Ramraj. (2020). Topic categorization of Tamil News Articles using PreTrained Word2Vec Embeddings with Convolutional Neural Network. 10.1109/CISPSSE49931.2020.9212248.

Z. Li, W. Shang and M. Yan, "News text classification model based on topic model," *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, Okayama, Japan, 2016, pp. 1-5, doi: 10.1109/ICIS.2016.7550929.

Comparative Analysis of Tamil and English News Text Summarization Using Text Rank Algorithm Sarika M Dr.Rajeswari K C and Lavanya A P Turkish Journal of Computer and Mathematics Education Vol.12 No.9 (2021).

Topic categorization of Tamil News Articles using PreTrained Word2VecEmbeddings with Convolutional Neural Network Conference Paper · October 2020

Text Summarization for Tamil Online Sports News Using NLP Conference Paper · December 2018