**International Journal of Applied Sciences and Smart Technologies**

Volume 3, Issue 2, pages 153–160

p-ISSN 2655-8564, e-ISSN 2685-9432

# Text Classification on Tamil

Omprakash Yadav[1], Alcina Judy[1], Praveen D’souza[1],
Calvin Galbaw[1,*], Hinal Rane[1]

[1] *Department of Computer, Xavier Institute of Engineering, Mahim, Mumbai 400016, India*
[*] *Corresponding Author: calving2012@gmail.com*

(Received 01-09-2020; Revised 29-12-2021; Accepted 29-12-2021)

**Abstract**

By and large, we don't know to talk and read the territorial dialects that are spoken in our nation. So we have accepted Tamil language as it is our territorial and numerous doesn't get it. In our task, the content in Tamil language is stacked from Wikipedia. It is then sifted through and extraordinary characters are evacuated it is then characterized by the titles like id, title, URL, etc. It is then used to prepare the model utilizing CNN calculation and the dataset is created. Along these lines, you would now be able to test utilizing an irregular Wikipedia page and the content is grouped by the titles and anticipated.

**Keywords:** tamil text classification, feature classification, vocabulary set or bag-of-words, text mining, natural language processing

## 1 Introduction

For the most part, we don't comprehend huge numbers of the local dialects in our nation. So at whatever point an individual of various state language is spoken or composed, we were unable to get it. In this task, we characterize the content dependent on the sort like name, nation, id, and so on. Here, we use CNN to arrange the content and train the dataset. It is useful for individuals to order the sort and in any event, get a thought of what the content looks like.

It will be simple for the individual to know and recognize the various segments present in the information. The sort of information is helpful for various logical purposes for getting it.

# 2 Literature Survey

We refer to references [1], [2], [3], [4], [5]. In deep learning, a convolutional neural system (CNN or ConvNet) is a class of deep neural systems, most usually applied to investigating visual features. These utilize the spatial loads of channels to extricate highlights from the picture. They have applications in picture and video acknowledgment, recommender frameworks, picture arrangement, clinical picture examination, regular language handling, and money related time arrangement. Convolutional neural systems use convolutional layers as building squares to gain from the dataset. Alongside these, pooling layers and completely associated layers are utilized.

A convolution is the basic use of a channel to an info that outcomes in an activation. These channels slide over width and tallness to convolve the information and use actuation capacity to make a highlighted map. This guide can be passed to another convolutional layer to make an increasingly itemized map.

These component maps can be unfurled to take care of into a completely associated layer to get the explicit prescient displaying issue, for example, picture arrangement. Since information like pictures, recordings, and other multi-dimensional information have a quadratic number of highlights, an ordinary neural system needs to process a huge measure of straight capacities and enactments which takes a quadratic measure of time. Be that as it may, convolutional organize registers each weight in a straight time utilizing channels.

he outcome is profoundly explicit highlights that can be distinguished anyplace on input images [3].

1. Convolutional neural systems apply a channel to a contribution to make a component map that sums up the nearness of recognized highlights in the input [3].
2. Filters can be high quality, for example, line finders, yet the advancement of convolutional neural systems is to get familiar with the channels during preparing with regards to a particular forecast problem [3].
3. How to figure the component map for one-and two-dimensional convolutional layers in a convolutional neural system [3].

For regular language handling tasks, counterfeit neural systems, for example, intermittent neural systems (RNN) and long transient memory systems (LSTM) are favored because they go off past initiation or yield as a contribution to the following concealed states. This aids in recalling the word/character figured which helping I processing the following ward word. That is the reason these models are utilized most often in language models.

Since the attempted assignment is of order, convolutional neural systems are utilized which changes in input. Instead of contributing a picture, word installing can be utilized as the info. Word installing is made utilizing different models, for example, Word2Vec. Since the forecast will be made on Wikipedia information, we have made an installation on Wikipedia pages.

**Input Layers**: It's the layer where we contribute to our model. The quantity of neurons in this layer is equivalent to add up to estimate of the word implanting.

**Hidden Layer**: For grouping utilizing word implanting, for the most part, a single layer of completely associated layers are utilized to shape the concealed layer. Additionally, a single layer of convolutional layer followed by a completely associated layer can be utilized.

**Output Layer**: Since there are n number of words, the yield from the concealed layer is then taken care of into a calculated capacity of the softmax layer which changes over the yield of each word into the likelihood score of each class.

The information is then taken care of into the model and yield from each layer is acquired this progression is called feedforward, we at that point figure the blunder utilizing a mistake work, some normal mistake capacities are cross-entropy, square misfortune blunder and so forth. From that point forward, we back engender into the model by figuring the subsidiaries. This progression is called Backpropagation which fundamentally is utilized to limit the misfortune.

# 3 Existing System

Natural language processing represents computational techniques used for processing human language. The language can either be represented in terms of text or speech. NLP in the context of deep learning has become very popular because of its ability to handle text which is far from being grammatically correct. The ability to learn from the data has made the machine learning system powerful enough to process any type of unstructured text. Machine learning approaches have been used to achieve state of the art results on NLP tasks like text classification, machine translation, question answering, text summarization, text ranking, relation classification, and others. The focus of our work is text classification of Tamil language. Text classification is the most widely used NLP task. It finds application in sentiment analysis, spam detection, email classification, and document classification to name a few. It is an integral component of conversational systems for intent detection. There have been very few text classification works in literature focusing on the resource-constrained Tamil language. While the most important reason for this is the unavailability of large training data; another reason is the generalizability of deep learning architectures to different languages. However, Tamil is a morphologically rich and relatively free word order language so we investigate the performance of different models on the Tamil text classification task. Moreover, there has been a substantial rise in Tamil language digital content in recent years. Service providers, e-commerce industries are now targeting local languages to improve their visibility. An increase in the robustness of translation and transliteration systems has also contributed to the rise of NLP systems for Tamil text. This work will help in the selection of the right models and provide a suitable benchmark for further research in Tamil text classification tasks.

# 4 Proposed Methodology

The proposed methodology is as follows.

**Step 1: Obtain the text from Wikipedia for Tamil pages**

Go to https://ta.wikipedia.org/wiki/முதற்_பக்கம் from this extract the text and convert it into csv file this file is then taken for further processing.

**Step 2: Filtering and removal of special characters:**

The special characters and the ambiguity present in the text are removed such as comma, semicolon, asterisk mark, brackets and so on. This will help the text to be simplified for further processing of data.

**Step 3: Classify using titles**

The text is classified according to the titles such as id, name, title, url, recursive words etc.

**Step 4: Train the dataset using CNN**

The dataset is trained using Convolutional Neural Networks (CNN) is one kind of feed forward neural network. CNN is an efficient recognition algorithm which is widely used in pattern recognition and image processing. It has many features such as simple structure, less training parameters and adaptability.

**Step 5: Test using random Wikipedia page**

Now we are able to test any random Wikipedia page and the text is classified according to the titles and predicts the results.

# 5 Implementation

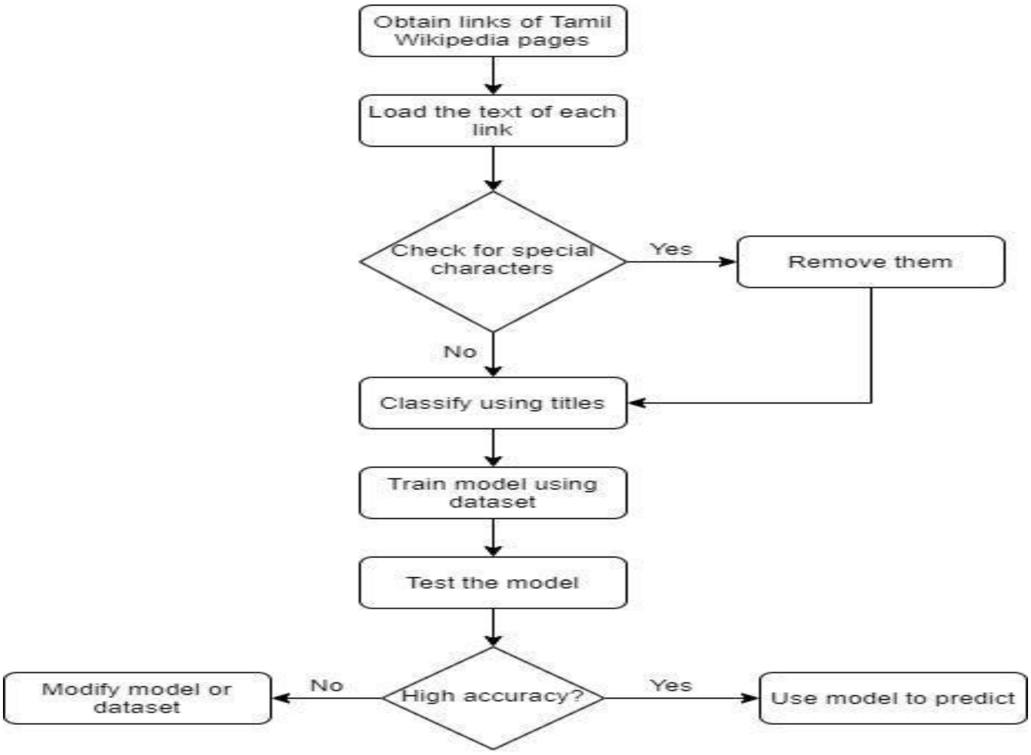
Obtain links of Tamil Wikipedia pages
Load the text of each link
Check for special characters
Yes
Remove them
No
Classify using titles
Train model using dataset
Test the model
High accuracy?
No
Modify model or dataset
Yes
Use model to predict

**Figure 1.** Flowchart.

The above Figure 1 is the Flowchart of our system. The working of our system is as follows:

1. The text from Tamil Wikipedia pages are extracted and checked for special characters.
2. Such characters create problem while classifying that is these special characters are not important to be classified.
3. The text is classified according to the title, tags, key words, and what the text is about.
4. This is used as the dataset for the model to be trained on.
5. Once we achieve high accuracy on the model, the user can use this model to get the details of an unknown Tamil text such as titles, etc.

# 6 Conclusion

In this report, we have introduced a Tamil language text arrangement that encourages the client to distinguish the sort of text and create a dataset by expelling all the ambiguities in the content and preparing the dataset which will be useful to test any irregular Wikipedia page.

# References

[1] E. Annamalai and S. B. Steever. Modern Tamil in Dravidian languages. *Newyork: Routledge Publication*, 1999.

[2] R. K. Belew, "Adaptive information retrieval." *In Proceedings of the 12th annual international ACM/SIGIR conference on research and development in information retrieval*, NY, 11–20, 1989.

[3] L. Chanunya and R. Peachavanish, "Automatic Thai language essay scoring using neural network and latent semantic analysis." *In Proceedings of the first Asia international conference on modeling and simulation*, 2007.

[4] C.H. Li and S.C. Park, "Text categorization based on artificial neural networks." *In ICONIP*, **4234**, LNCS 302–311, 2006.

[5] C.H. Li and S.C. Park, "Neural network for text classification based on singular value decomposition." *In Seventh international conference on computer and information technology*, 47–52, 2007.

This page intetntionally left blank