POS Tagging Using Naïve Bayes Algorithm For Tamil

M. Rajasekar, Research Scholar, Dr. A. Udhayakumar, Professor

Abstract: Part-of-speech Tagging is the basic and major task in any Natural Language Processing Applications. Based on this POS Tagged corpus, it will be implemented to Named Entity Recognition, Information Extraction, and Machine Translation and so on. In this research paper the machine learning method Naïve Bayes algorithm is implemented to get optimal output of Tagged corpus. The source documents are women health issues related article from Internet. The result got from this approach is good.

Index Terms: POS Tagging, Machine Learning, Naïve Bayes, Women health issues

1. INTRODUCTION

In the world of Natural Language processing the most basic and important task is Part-of-Speech Tagging. The POS tagging is the essential part of all the NLP applications. Part-of-speech tagging is the process of marking up a word in a corpus as corresponding to a particular syntactic category known as part of speech[1], such as Noun, Verb, Adjective, Pronoun, Conjunction, etc.,

2. POS TAGGING

The parts of speech tagging is the task to assign the part-of-speech with a word in the sentences based on its syntactical structure of the Language. There are some techniques in POS Tagging.

- Rules based method
- Probabilistic method
- Deep learning method

Simply, Rules based method is assigning POS Tag based on predefined rules. The rules depend upon the Language structure. The probabilistic method is assigning POS Tags based on the probability of particular tag sequence occurring. The deep learning method is to assign the POS Tag by using recurrent neural networks.

2.1. POS Tagging in Tamil

The process of POS Tagging in such a morphologically rich language Tamil is very complex. Because the word formation structure in Tamil language is flexible. This paper discuss about the different approach to implement the POS Tagging in Tamil Language. The POS Tagging and grammartical structure is defined in Tamil language in 13th Century itself. The grammar book Nunnul is a Tamil Grammar guide written by Sage Pavanandi. The Nunnul is the first grammar guide in Tamil Language.

2.2. Overview of Nunnul:

In Nunnul the Sage Pavanandi has given detailed structure and rules for formatting the sentence from word by word. The overall structure and types for Noun, Verb and other parts[2].

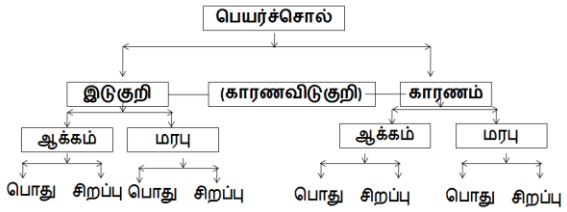

Figure 1. *Nouns derivation in Nunnul*

In the above figure the noun derivation is explained in nunnul. The noun (peyarsol) has two derivation. Iduguri peyar, karanappeyar. In both iduguri and kaaranam they has two types. Aakkam, marapu. Then these two types are classified as podhu, sirappu peyar.

Likewise the sage pavanandi gave the full description about the verb components, pagudhi, vigudhu, santhi, saariyari.

3. COMPUTATIONAL METHODOLOGY

Nowadays the Part-of-speech tagging process is fully implemented in computational methods. The computational algorithms and rules have developed to implement the POS Tagging of a particular language to give maximum accuracy with the manual knowledge about POS Tagging.

3.1 Need of computational methods

The main problem of the POS Tagging is ambiguity. That seems a word or token can act as two POS Tag. For example,

Sentence 1 ➔ The book is on the table.

Sentence 2 ➔ Book the tickets

In the sentence 1, the word book is meaning as noun. The same word book is meaning as verb in the sentence 2. By manually it is easy to identify the verb and noun. But in the implementation of computational methodology it is such a critical task. So, the machine learning methods were to be used to solve this issue and given the approximate output and accuracy. These computational methods has its own merits and demerits[3].

4. LITERATURE REVIEW

The task of POS Tagging for Tamil Language is already done in such case by the eminent researchers in the field of Natural Language Processing.

4.1. Hierarchal POS tagging for Tamil

The paper Hierarchical POS tagging for Tamil language using Machine learning approach, Dr.V.Dhanalakshmi et

al[4], is dealing the problem with Support vector machine(SVM) tool. They have implemented the hierarchical tagging method for two types of category. The word grammatical category and grammatical feature level[4]. They have got good results.

4.2. POS Tagger and Chunker for Tamil Language

The paper POS Tagger and Chunker for Tamil Language Dhanalakshmi V. et al[4], is deal with POS Tagging and chunking of words from sentences. The SVM tool have implemented for this research work. The accuracy level is high POS tagger (95.64%) and chunker (95.82%).

4.3. POS Tagging for classical Tamil Text

The research work, POS Tagging for classical Tamil Text, R. Akilan, et al[6], has its own method. It follows the Novel methodology with pre defined rules to tag the words. They deals with the classical Tamil literature text.

4.4. Advanced Tamil POS Tagger for Language Learners

The research work Advanced Tamil POS Tagger for Language Learners, M. Rajasekar et al[7], deals with the POS Tagging for Tamil text with derived POS Tags. It explains that more deep in the Tag sets. They have used rules based tagging methods. They have achieved good results.

4.5. Limitations of reviewed work

All the above research work has done for global word set of Tamil language. All the work dealt with the general Tamil words, the special fields of source text is not available. Domain specific word set is unavailable. Mostly SVM and Hierarchical methodology were used.

5. OBJECTIVES

The objectives of the proposed research work are as follows

- To produce the optimum accuracy of the POS Tagger
- To implement the probabilistic methodology for POS Tagging
- To develop domain specific corpus set in the field of Women health issues articles.
- To enable the researchers to use the generated corpus for NLP applications

6. PROPOSED RESEARCH WORK

In this proposed research work, the source documents are collected as women health related articles from internet. It provides the facility to move forward the development of NLP applications in the domain women health issues. There are 75 sets of articles were collected from internet to process in the proposed POS Tagger.

7. METHODOLOGY FOR POS TAGGING

The POS tagging is the process of classification of word into the particular category. The classification methods are as follows.

7.1. Text Classification

The term text classification is the process of assigning tags or categories to text according to its content[8]. It's one of the basic and prominent tasks in Natural Language Processing (NLP) with extensive applications such as sentiment analysis, spam mail detection, topic analyzing[8]. Unstructured data in the form of raw text is all over the technology world. E books, social media text, web pages, news, survey responses, chat messages and E mails. Text can be in the form of unstructured or semi structured in all the above sources. But users cannot access that text in their needed format or structure. To provide the text in the form of user preferred format is cost effective and time consuming process. This process is much hard for the famous language like English. In India, multi lingual country is the text extraction process is much harder in the regional language that can understand by the people living in the particular region. In this research work, the needed knowledge (text) is extracted from the southern ancient language Tamil.

7.2 How Does Text Classification Work?

The text classification can be done in two methods[8], Manual and automatic text classification.

Manual: a human can interprets the exact content of the raw text and categorizes it accordingly.

Automatic / Machine Learning: the natural language processing techniques are applied to classify text and it is faster and cost effective.

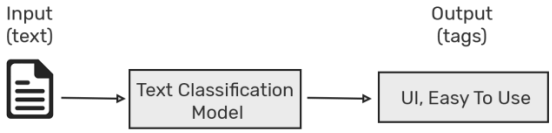

Figure2. *Text Classification*

There are many automatic text classification methods, they are as follows:

- Rules based classification
- Machine Learning based classification
- Hybrid method

7.2.1. Rules based classification

The handheld rules to be applied to classify the text into the predefined groups. When the system identify the word, it will compare the content with the predefined group tags, which is most appropriate to match with the current word. There are some disadvantages in this method. Deep learn about the knowledge of the domain, time consuming, frequently reconstruct the rules.

7.2.2. Machine learning base classification

Instead of following the manually constructed rules, the machine learning classification method uses the past observations. This method follows the earlier identified examples in the training document; it will make decision to classify the text into the expected group.

A. Word to Vector representation

The first step in machine learning method is to develop the classifier for the set of training data. The given set of words transformed into the vector representation.

For example,

If we have set of word in our corpus, {The, cat, cow, rat, is, running, fast, with, calf, bad, not, basketball}, and the testing data set has, the sentence, "The cow is with calf". So that the vector representation is (1,0,1,0,1,0,0,1,1,0,0,0).

Then the vector format is forwarded to the feature extraction module, it will apply the Machine learning algorithm to form a classification model.

8. MACHINE LEARNING ALGORITHMS

The list of commonly used machine learning algorithms for text classification are,

- Decision Trees
- Naive-Bayes classifier
- Support Vector Machines
- K Nearest Neighbors
- Fuzzy C-Means

A. Decision Trees

This the supervised machine learning method, to classify the text from raw text document. In this method the data or sentence is continuously split according to a certain parameter[9]. The tree can be defined by two entities, decision nodes and leaves. The decision nodes are the data is to be split, the leave are the decisions or the final result.

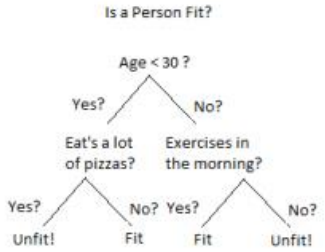

Figure 3. *Decision Tree*

B. Naïve Bayes classifier

This the simple "probabilistic classifier" based on applying Bayes theorem with strong assumptions between the features or groups[10].

Baye's theorem, the conditional probability can be as

$$p(C_k|X) = \frac{p(C_k)p(x|C_k)}{p(x)}$$

the equation can be written as

$$posterior = \frac{prior\ X\ likelihood}{evidence}$$

C. Support Vector Machines

Support vector machine (SVM) is a supervised machine learning algorithm. It is mostly used in classification problems. The each data item as the point, in n-dimensional space with the value of each data, then the machine perform the classification by finding the hyper-plane [11].

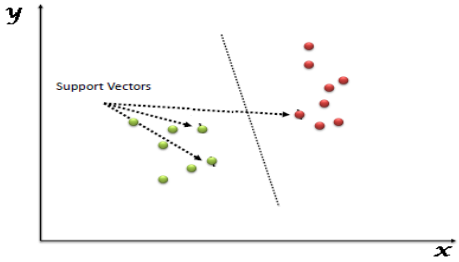

Figure 4. *SVM Model*

D. K nearest Neighbors

K- nearest neighbor algorithm, can be used for classification and regression predictive problems[12]. The K NN algorithm can be understand by the following example.

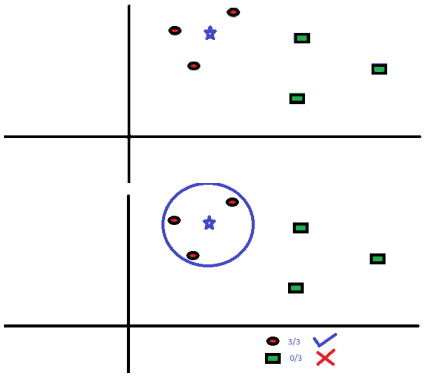

Figure 5. *KNN algorithm*

The above figure has spread of red circles (RC) and green square(GS). The problem is to find out the group of the blue star (BS). In this, the "K" is the nearest neighbor to take vote from, K=3. The circle can be made with three elements in the plane. Get the maximum number of feature, from the K value 3. It is 3 Red Circle (RC). So, it concludes the blue star will be grouped into red circle.

E. Fuzzy C- Means

Fuzzy C-means (FCM)is a method for classification which allows one part of data to belong to two or more groups[13]. It is based on minimization of the following objective function:

$$\arg\min_{C} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} \|\mathbf{x}_i - \mathbf{c}_j\|^2,$$

Where,

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|\mathbf{x}_i - \mathbf{c}_j\|}{\|\mathbf{x}_i - \mathbf{c}_k\|} \right)^{\frac{2}{m-1}}}.$$

9. POS Tagging from Women health related documents

In the proposed research work, the women health issues related text documents in Tamil language is used to implement the POS Tagging. The source data is raw text, the methodology is used to extract the information, is supervised machine learning algorithm, Naive bayes algorithm.

9.1. Why Naïve bayes theorem?

- Simplicity: Actually it is quite transparent, easy to apply and fast.
- Light to train, Need less training data.
- Easily updateable if more training data is added

- Can be used for in classification problems in binary and multi-class level
- Small memory needed
- Its performance is very good even critical situation

9.2. Basics of Natural Language processing for Text

A. Tokenization

The sentence is divided into number of words. These words are considered as basic elements to generate the extraction model. The input sentence is as follows Figure.5.

இயற்கையான முறைக்கு மாறாக சில சூழல்களில் அறுவைசிகிச்சை பிரசவம் மேற்கொள்ளப்படுகிறது.

Figure 6. Initial Raw text

இயற்கையான
முறைக்கு
மாறாக
சில
சூழல்களில்
அறுவைசிகிச்சை
பிரசவம்
மேற்கொள்ளப்படுகிறது
.

Figure 7. *Output of the Tokenization*

B. Normalization

From the list of words in the wordset, each word is to be normalized, as remove punctuation marks, removing stops words and inflection of the root words. Normalization is the process to trim the words into roots by removing articulation through crumbing unnecessary characters[10]. In this Normalization process the system will remove the meyyazhuthukkaL, dot letters in Tamil languages, except, (ன், ல், ண், ம், ள், ழ், ஸ), and special characters, (.,,,:.;,”,’,-) Figure shows the input and output view of normalization.

இயற்கையான
முறைக்கு
மாறாக
சில
சூழல்களில்
அறுவைசிகிச்சை
பிரசவம்
மேற்கொள்ளப்படுகிறது
.

Figure 8. *Input for Normalization*

இயற்கையான
முறைக்கு
மாறாக
சில
சூழல்களில்
அறுவைசிகிச்சை
பிரசவம்
மேற்கொள்ளப்படுகிறது

Figure 9. *Output for Normalization*

C. POS Tagging

The POS Tagging is the processing of assign the suitable parts of speech tag to a word in the documents. The morphologically rich language Tamil, POS Tagging is not easy. The Bureau of Indian Standards has registered for standardization of POS Tags for Indian language on 2010. There are 47 Tags for Tamil language.

D. Ambiguity

A word can behave more than one POS tag is called ambiguity. For Example,
Book the Ticket – the word book is verb
The book is on the Table – the same word book is noun
To solve this ambiguity problem, need machine learning method to get most probable
POS tag for the given word. The supervised machine learning method Naïve Bayes thorem is used to classify the words.

E. Naïve Bayes Thorem

Finding the most probable result, the equation is,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Finding the Probability of Event A, when Event B is True
P(A) – Priori Probability (Probability of event before event B)
P(A|B) - Posterior Probability (Probability of event after event B is true)
To explain this model, the sample data is used as follows,
The Data set,

Text	Category
A great game	Sports
The Electron was over	Non Sports
Very clean match	Sports
A clan but forgettable game	Sports
It was a close electron	Non Sports

Table 1*. Dataset for Naïve Bayes*

Total words in Sports-11
Total words in Non Sports-9
Total unique words-14
Aim to classify the sentence “A very close game” Whether Sports or Non Sports cateory.
Concept: To find the probability of given sentence in two categories.
P (Sports | A very close game) = ?
P (Non Sports | A very close game) = ?
Finding Probability
Given Sentence is -→ “ A very close game”
P(A very close game) = P(A) X P(very) X P(close) X P(game)
Feature Engineering
P(A | Sports) = P(A| Sports) X P(very| Sports) X P(close| Sports) X P(game| Sports)
P(A | Non Sports) = P(A| Non Sports) X P(very| Non Sports) X P(close| Non Sports) X P(game| Non Sports)
Calculating Probabilities
Sentence: A very close game
P(a| sports) =2/11, P(very | sports) = 1/11, P (close| sports) = 0/11, P (game |sports) = 2/11
Hence,
P(A very close game | sports) = P(a| sports) X P(very | sports) X P (close| sports) X P (game|sports) = 2 / 11 X 1 / 11 X 0 / 11 X 2 / 11 = 0

The probability 0 means, the result could not be found. In this case we have solve this issue with a special term Laplace Smoothing

$$\theta_i = \frac{x_i + \alpha}{N + \alpha d}$$

θ - Probability of current word
α is always 1 (α>0)
x– is the word count
i – is the count (no. of words from 1 – last word)
N - Total number of words
d - Number of Unique words / Distinct words

In the Simple Language,

$$P(Word) = \frac{word\ count + 1}{Total\ no.of\ words + No.of\ Unique\ words}$$

Now let's calculate,

$$P(a\ |sports) = \frac{2+1}{11+14}$$

$$P(very\ |sports) = \frac{1+1}{11+14}$$

$$P(close\ |sports) = \frac{0+1}{11+14}$$

$$P(game\ |sports) = \frac{2+1}{11+14}$$

After finding all the probability for the two categories,
P(A | Sports) = P(A| Sports) X P(very| Sports) X P(close| Sports) X P(game| Sports)
$= 4.61X10^{-5} = 0.0000461$
P(A | Non Sports) = P(A| Non Sports) X P(very| Non Sports) X P(close| Non Sports) X P(game| Non Sports)
$= 1.43 \times 10^{-5} = 0.0000143$
So,
P(A | Sports) = P (a very close game | sports) is high,
Hence conclude as the given sentence is belongs to sports category.
The above method is used in the proposed problem to give the optimum feasibility to assign the POS Tag, to a particular word.

F. POS Tagging for Women health documents
By using the naïve bayes classification algorithm, explained above, the POS tagging model has implemented for the proposed women health related documents.

10. Evaluation and Findings
The F- test evaluation method is used to evaluate the results of the POS Tagging model.

$$Precision = \frac{No.of\ Retrived\ relevant\ documents}{Total\ no.of\ retrived\ documents}$$

$$Recall = \frac{No.of\ Retrived\ relevant\ documents}{Total\ no.of\ relevant\ documents}$$

$$F\ Score = \frac{2\ X\ Precision\ X\ Recall}{(Precision + Recall)}$$

Words	Precision	Recall	F-Score
6549	0.8	0.9	0.8

Table 2. *Evaluation*

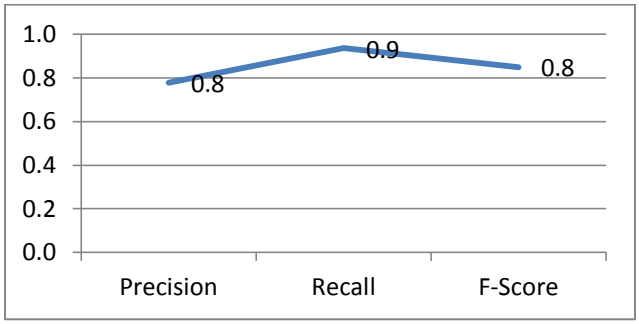

Figure 6. *Evaluation Chart*

11. CONCLUSION

The POS Tagging for the women health related documents is implemented and tested for 53 documents by using Naïve bayes classification method. The evaluation is done with Precision, Recall and F-Score evaluation method. It gave good result. It is very much useful for the Natural Language Processing Applications in the specific domain.
In future by using this POS Tagging method, the Named Entity recognition and Information Extraction model will be developed.

12. REFERENCES

[1] https://en.wikipedia.org/wiki/Part-of-speech_tagging
[2] https://ta.wikipedia.org/wiki/%E0%AE%A8%E0%AE%A9%E0%AF%8D%E0%AE%A9%E0%AF%82%E0%AE%B2%E0%AF%8D
[3] Rajendran S,Complexity of Tamil in POS tagging, Language in India, Jan 2007.
[4] http://www.ldcil.org/Download/POSANIL2011/9Hierarchal%20POS%20tagging%20for%20Tamil%20language%20using%20Machine%20learning%20approach.pdf
[5] Dhanalakshmi. V., M Kumar, A., Soman K. P., and S., R., "POS Tagger and Chunker for Tamil Language", Proceedings of the 8th Tamil Internet Conference. Cologne, Germany, 2009.
[6] POS TAGGING FOR CLASSICAL TAMIL TEXTS, R.Akilan,E.R.Naganathan, International Journal of Business Intelligent volume: 1 No: 01 January – June 2012
[7] Advanced Tamil POS Tagger for Language Learners, M. Rajasekar, Dr. A. Udhayakumar, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-10, August 2019
[8] https://monkeylearn.com/text-classification/
[9] https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html
[10] https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[11] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
[12] https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/
[13] https://en.wikipedia.org/wiki/Fuzzy_clustering