Thesis for the Degree of Master of Computer Information System

Nepali Text Part Of Speech Tagging Using Different Deep Learning Algorithms

Tika Ram Khojwar

(2020-2-92-0018)

Nepal College of Information Technology

Faculty of Management Studies

Pokhara University, Nepal

May, 2023

Abstract

POS tagging is an essential and foundational task in numerous natural language processing (NLP) applications. Such as machines translation, text-to-speech conversion, question answering, speech recognition, word sense disambiguation and information retrieval, text summarization, Named entity recognition, sentiment analysis etc. POS tagging entails assigning the correct tag to each token in the corpus, considering its context and the language's syntax. An optimal part-of-speech tagger plays a crucial role in computational linguistics. Its importance cannot be emphasized enough because inaccuracies in tagging can greatly affect the performance of complex natural language processing systems. Developing an efficient POS tagger for morphologically rich languages like Nepali is a challenging task. . This research study will focus on evaluate the performance of different models and algorithms to find the optimal POS tagger. The models will be supervised deep learning models such as RNN, LSTM and BiLSTM. And mBERT. Because Deep Learning oriented methodologies improves the efficiency and effectiveness of POS tagging in terms of accuracy and reduction in false-positive rate. These models will be trained with the available tagged dataset and tested to compare the performance measures of each classification algorithm.

Keywords: POS Tagging, Nepali Text, Recurrent Neural Network, LSTM, BiLSTM, BERT

Table of Contents

List of Figures

Abbreviations/Acronyms

NLP	Natural Language Processing
POS	Part Of Speech
HMM	Hidden Markov Model
SVM	Support Vector Machine
ANN	Artificial Neural Network
LSTM	Long Short-Term Memory
BiLSTM	Bi-directional Long Short-Term Memory

Chapter 1

INTRODUCTION

1.1 Background

Natural Language Processing is a field of artificial intelligence that focuses on the interaction between computers and human language. NLP involves the development of algorithms and techniques to enable computers to understand, interpret, and generate human language in a way that is meaningful and useful.

In NLP, Part-of-Speech tagging is a fundamental task. It involves assigning POS tags (Noun, pronoun, verb, adjective, adverb, preposition etc.) to each word in a sentence of a natural language. The input for the algorithm consists of a sequence of words in a natural language sentence and a predefined set of POS tags. The output is the most suitable POS tag assigned to each word in the sentence. POS tagging provides valuable information about a word and its neighboring words, which proves beneficial for various advanced NLP tasks like speech and natural language processing applications, semantic analysis, machine translation, text-to-speech conversion, question answering, speech recognition, word sense disambiguation and information retrieval, text summarization, Named entity recognition and more.

Nepali is a morphologically rich language. One of the characteristics features of the Nepali language is its rich inflectional system, especially the verbal inflection system. A verb in Nepali can easily display more than 20 inflectional forms while encoding different morphological features, including aspect, mood, tense, gender, number, honorifics, and person.

In Nepali language same words can have different meanings. Without POS tagging both word will be treated as same word having same meaning. But by the means of POS tagging the word can be differentiated as two different words with different meaning. For example

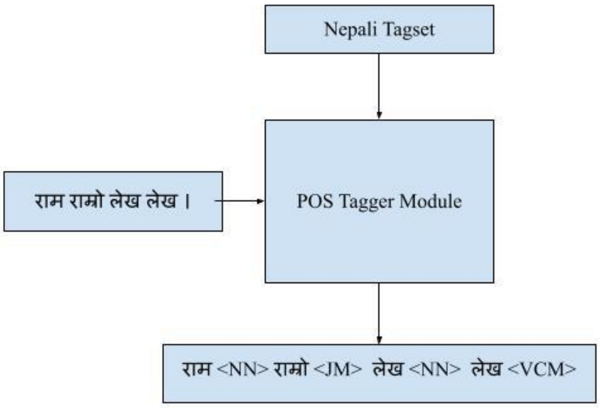

Fig. 1.1: POS tagging example

Here, the word 'लेख' repeated twice in the sentence and have different meanings based on the position of the word. As a result of POS tagging, the sentence can be converted as:

राम <NN> राम्रो <JM> लेख <NN> लेख <VCM>

Each word of the sentence is assigned a part of speech. The first occurrence of the word लेख is tagged as noun whereas the second occurrence is tagged as the command form verb. By this process the words are marked as unique words and the ambiguity can be removed as the application utilizing the POS feature can identify the meaning of the first लेख as article whereas the second लेख as write.

1.2 Statement of Problems

Nepali is morphologically rich language. Several POS tagging model for Nepali language have been done in the past, but satisfactory results have not been obtained. There is also a constraint in automatically tagging "Unknown" words with a high false positive rate. Rule based and statistical techniques do not show significant results as they do not take care of context and sequence. Very few implementations of deep learning approaches can be found in context of morphologically rich languages like Nepali. This research study will focus on evaluate the performance of different models and algorithms to find the optimal POS tagger. The models will be supervised deep learning models such as RNN, LSTM and BiLSTM. And pre-trained language model BERT. Because Deep Learning oriented methodologies improves the efficiency and effectiveness of POS tagging in terms of accuracy and reduction in false-positive rate. These models will be trained with the available tagged dataset and tested to compare the performance measures of each classification algorithm.

1.3 Objectives of the Study

The best model depends on various factors, including the availability of training data, Language, computational resources, and the specific requirements of the application. So, experiment with different models and compare their performance on the specific task or dataset to determine the most suitable model.

The main objective of this paper is:

- Train and compare between different deep learning algorithms such as RNN, LSTM, BiLSTM and mBERT for nepali text POS tagging.
- Find out which algorithm is best suited for the process of POS tagging for Nepali text.

1.4 Significance of Study

The main significance of this research is to assign the correct tag to the word of text. Only correct assignment of the tag gives correct sense of the words. Which proves beneficial for various advanced NLP tasks like speech and natural language processing applications, semantic analysis, machine translation, text-to-speech conversion, question answering, speech recognition, word sense disambiguation and information retrieval, text summarization, Named entity recognition.

Chapter 2

LITERATURE REVIEW

There are only few researches have been done in the field of POS tagging for Nepali language. Some of them used statistical model (HMM) for identifying the tags while some used supervised machine learning model and some used supervised deep learning model to train the model.

Ingroj Shrestha, Shreeya Singh Dhakal (2021):

This article applied three deep learning models: BiLSTM, BiGRU, and BiLSTM-CRF for fine-grain POS tagging for the Nepali language. It uses Nepali National Corpus (NNC). It has 17 million manually and semi-manually words tagged with 112 POS-tags. Results show that deep learning models could capture fine-grained morphological features like gender, person, number, and honorifics that are encoded within words in highly inflectional languages like Nepali with a large enough dataset. This study trained all the models using two embedding: pre-trained multi-lingual BERT and randomly initialized Bare embedding. It found that training a randomly initialized Bare embedding is better than the ones trained using large pre-trained multi-lingual BERT embedding for downstream tasks in Nepali like POS tagging. Among the tested models, the BiLSTM-CRF with the Bare embedding performed the best and achieved a new state-of-the-art F1 score of 98.51% for fine-grained Nepali POS tagging. This research contributes to the advancement of NLP techniques tailored specifically for the Nepali language.

Sarbin Sayami, Tej Bahadur Shahi and Subarna Shakya (2019):

This paper addresses the implementation and comparison of various deep learning-based POS taggers for Nepali text. The examined approaches include RNN, LSTM, GRU, and BiLSTM. These models are trained and evaluated using Nepali English parallel corpus annotated with 43 POS tag and contains nearly 88000 words which is collected from m Madan Puraskar Pustakalaya. The design of this Nepali POS Tag-set is inspired by the PENN Treebank POS Tag-set. The data set is divided into three sections i.e. training, development and testing. The accuracy obtained for simple RNN, LSTM, GRU and

Bidirectional LSTM are 96.84%, 96.48%, 96.86% and 97.27% respectively. Therefore, Bi-directional LSTM outperformed all other three variants of RNN

Greeshma Prabha, Jyothsna P V, Shahina kk, Premjith B, Soman K P (2018):

This paper proposed a deep learning based POS tagger for Nepali text which is built using Recurrent Neural Network (RNN), Long Short Term Memory Networks (LSTM), Gated Recurrent Unit (GRU) and their bidirectional variants. It uses POS Tagged Nepali Corpus generated by translating 4325 English sentences from the PENN Treebank corpus tagged with 43 POS tags. The results demonstrate that the proposed model outperforms existing state-of-the-art POS taggers with an accuracy rate exceeding 99%. This research contributes to the field by showcasing the effectiveness of deep learning techniques in improving POS tagging for Nepali text.

Ashish Pradhan, Archit Yajnik (2021):

This article presents a comprehensive study and comparing two techniques, HMM and GRNN, for POS Tagging in Nepali text. The POS taggers aim to address the issue of ambiguity in Nepali text. Evaluation of the taggers is performed using corpora from TDIL (Technology Development for Indian Languages) which contains a total of 424716 tagged words with 39 tags, tags follow the guidelines of ILCI (Indian Languages Corpora Initiative), BSI (Bureau of Indian Standard), with implementation carried out using Python and Java programming languages, along with the NLTK Toolkit library. The achieved accuracy rates are as follows: 100% for known words (without ambiguity), 58.29% for ambiguous words (HMM), 60.45% for ambiguous words (GRNN), and 85.36% for non-ambiguous unknown words (GRNN). Although the GRNN tagger achieves the accuracy as high as the HMM Tagger, it fails or becomes unstable when the training data set is greater than 7000 words, while the HMM Tagger is trained with more than 400000 words with corresponding tags. A total of 4000 words are used for testing on both HMM and GRNN taggers.

Archit Yajnik (2018):

This article focuses on POS tagging for Nepali text using the GRNN. Because GRNN is less expensive as compared to standard algorithms viz. Back propagation, Radial basis function, support vector machine etc. And also neural network is usually much faster to train than the traditional multilayer perceptron network. The corpus has total 7873 Nepali words with 41 tags. Out of which 5373 samples are used for training and the remaining 2500 samples for testing. The results show that 96.13% of words are correctly tagged on the training set, while 74.38% are accurately tagged on the testing set using GRNN. To compare the performance, the traditional Viterbi algorithm based on HMM is also evaluated. The Viterbi algorithm achieves classification accuracies of 97.2% and 40% on the training and testing datasets, respectively. The study concludes that the GRNN-based POS tagger demonstrates more consistency compared to the traditional Viterbi decoding technique. The GRNN approach yields a higher accuracy on the testing dataset, suggesting its potential for improved POS tagging in Nepali text compared to the Viterbi algorithm.

Archit Yajnik (2018):

The article that introduces POS tagging for Nepali text using three Artificial Neural Network (ANN) techniques. A novel algorithm is proposed, extracting features from the marginal probability of the Hidden Markov Model. These features are used as input vectors for Radial Basis Function (RBF) network, General Regression Neural Networks (GRNN), and Feed forward neural network. The training database contains 42100 words whereas the testing set consists of 6000 words with 41 tags. The GRNN-based POS tagging technique outperforms the others, achieving 100% accuracy for training and 98.32% accuracy for testing. This research contributes to Nepali POS tagging by presenting a novel algorithm and highlighting the effectiveness of the GRNN approach.

Archit Yajnik (2017):

This article focuses on POS tagging for Nepali text using the HMM and Viterbi algorithm. The study reveals that the Viterbi algorithm outperforms HMM in terms of computational efficiency and accuracy. Database is generated from NELRALEC Tagset with 41 tags. A report on Nepali Computational Grammar is made available by Prajwal Rupakheti et al. Database contains 45000 Nepali words with corresponding Tag, out of which 15005 samples are randomly collected for testing purpose. The Viterbi algorithm achieves an accuracy rate of 95.43%. The article also provides a detailed discussion of error analysis, specifically examining the instances where mismatches occurred during the POS tagging process.

Abhijit Paul, Bipul Syam Purkayastha, Sunita Sakar (2015):

This paper discusses HMM based POS tagging for the Nepali language. The study evaluates the HMM tagger using corpora from Technology Development for Indian Languages (TDIL) which contains around 1,50,839 tagged words and tagset consists of 42 tags including generic attributes and language specific attribute values. It has been followed the guidelines of ILCI (Indian Languages Corpora Initiative), BSI (Bureau of Indian Standard). The implementation is done using Python and the NLTK library. The HMM-based tagger achieves an accuracy of over 96% for known words but the system is not performing well for the text with unknown words yet. Overall, the paper provides insights into the effectiveness of HMM for Nepali POS tagging and highlights areas for future improvement.

Tej Bahadur Shahi, Tank Nath Dhamala, Bikash Balami (2013):

This paper focuses on SVM based POS tagger for Nepali language which uses the dictionary as a primary resources. This dictionary is collected from the FinalNepaliCorpus which contains only 11147 unique words. The POS tagging approaches like rule-based and HMM cannot handle many features that would generally be required for modeling a morphologically rich language like Nepali. SVM is efficient, portable, scalable and trainable. So, this paper proposes a SVM based tagger. The SVM tagger constructs feature vectors for each word in the input and classifies them into one of two classes using a One Vs Rest approach. The SVM algorithm achieves an accuracy rate of 96.48% for known words, 90.06% for unknown words and 93.27% in overall. That means SVM tagger demonstrates strong performance for known words. In comparison to rule-based and Hidden Markov Model (HMM) approaches, the SVM-based tagger exhibits a slightly higher overall accuracy.

Chapter 3

RESEARCH METHODOLOGY

In this chapter, we present the proposed method for determining the POS tags of the provided text. The following figure provides an overview of the tagging process.

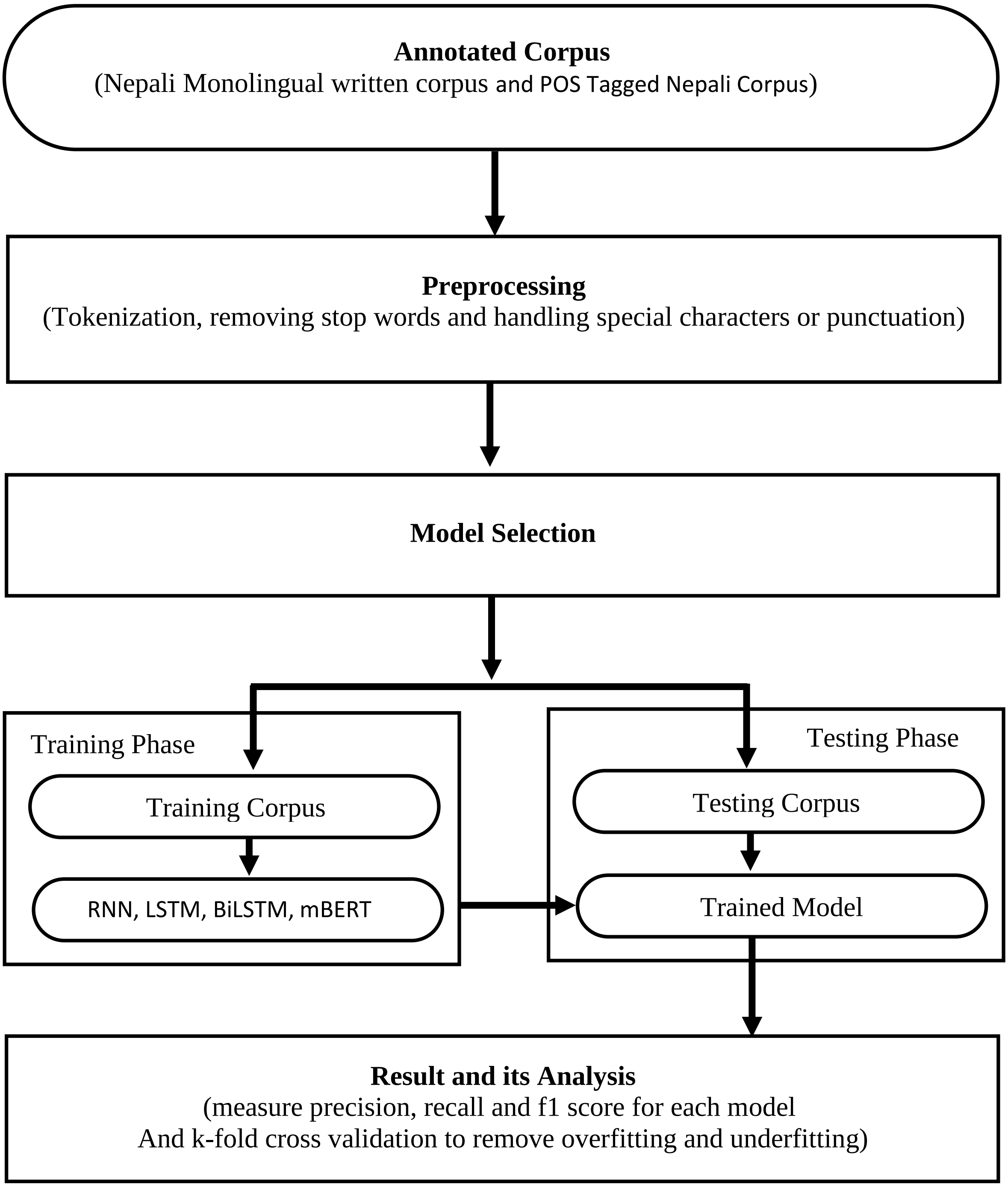
Annotated Corpus
(Nepali Monolingual written corpus and POS Tagged Nepali Corpus)
Preprocessing
(Tokenization, removing stop words and handling special characters or punctuation)
Model Selection
Training Phase
Training Corpus
RNN, LSTM, BiLSTM, mBERT
Testing Phase
Testing Corpus
Trained Model
Result and its Analysis
(measure precision, recall and f1 score for each model
And k-fold cross validation to remove overfitting and underfitting)

Fig 3.1: Proposed methodology for POS tagging

3.1 Dataset Collection

This research study will implement two corpus one is Nepali Monolingual written corpus. It consists of two main parts: the core corpus (core sample) and the general corpus. The core sample (CS) is a compilation of Nepali written texts from 15 diverse genres, with each text containing 2000 words. These texts were published between 1990 and 1992. On the other hand, the general corpus (GC) comprises written texts gathered from various sources, including the internet, newspapers, books, publishers, and authors, opportunistically collected without a specific sampling criteria. The corpus has total 2,202,000 words. It is a morphologically annotated corpus. A parts-of-speech tagset has been produced within the project: the Nelralec Tagset.

The other is POS Tagged Nepali Corpus which is available in Center for Language Engineering (CLE) website. It contains 4325 Sentences with 100720 annotated words. The POS tagset has 43 POS tags (null included) which is influenced by the PENN Treebank tagset.

3.2 Preprocessing Dataset

Only good dataset can give good output. To make good dataset, we need to transform the text into something meaningful that the algorithm can use. Preprocessing includes the tokenization, removing less useful parts such as removing stop words, empty line etc, stemming or lemmatization, and handling special characters or punctuation. Text data are converted to label encoded form to represent numerically so that models can perform well.

3.3 Model Selection

There are several models have been widely used and achieved good performance in POS tagging tasks. Different models have their own different features and specific task. There is no single "best" model for POS tagging, as the effectiveness of a model can vary depending on factors such as the dataset, language, and specific requirements of the task. According to previous research Deep Learning algorithms based models gives better accuracy in testing dataset and can deal with ambiguous, unknown words as compare to rule based, statistical and machine learning algorithms for POS tagging. Let’s see some probable DL model for labeled dataset.

3.3.1 RNN

First selected model is Recurrent Neural Network (RNN). It is a type of neural network specifically designed to handle sequential data. It has feedback connections that enable it to maintain and utilize information from previous steps in the sequence. RNNs are effective in capturing dependencies over time and are commonly used in tasks such as natural language processing, speech recognition, and time series analysis. They can learn patterns and make predictions based on the context of the sequence. Overall, RNNs are powerful tools for modeling and processing sequential data.

3.3.2 LSTM

Second selected model is Long Short-Term Memory (LSTM) which is a type of RNN architecture. It can deal with capturing long-term dependencies in sequential data. Because LSTMs utilize a memory cell, along with input, forget, and output gates, to selectively retain or discard information based on context. This helps them overcome the vanishing gradient problem and effectively learn from sequences of varying lengths. LSTMs have been successfully applied to tasks such as natural language processing, speech recognition, machine translation, and time series analysis. They are known for their ability to capture long-term dependencies and have become popular for modeling sequential data.

3.3.3 BiLSTM

Third selected model is Bidirectional Long Short-Term Memory (BiLSTM) is a variant of the LSTM model that processes input sequences in both forward and backward directions. By capturing context from both past and future elements, Bi-LSTM can model long-range dependencies and excel in tasks that require complete context understanding. It is widely used in NLP tasks, enabling improved performance and accuracy in analyzing sequential data with bidirectional context.

3.3.4 mBERT

Multilingual Bidirectional Encoder Representations from Transformers (mBERT) is a variation of the BERT model that is trained on multiple languages simultaneously. It is designed to handle multilingual and cross-lingual NLP tasks. mBERT is trained on a large corpus of text data from different languages, allowing it to learn shared representations that capture similarities and transferable knowledge across languages. It learns deep contextual representations of words and can effectively handle tasks such as text classification, part-of-speech tagging, named entity recognition, and machine translation across multiple languages.

One of the key advantages of mBERT is that it can perform well on a wide range of languages without requiring separate models for each language. It can leverage shared representations to transfer knowledge from high-resource languages to low-resource languages, improving performance on languages with limited training data.

3.4 POS Tagging

After model selection training phase will start and then train the selected models RNN, LSTM, BiLSTM, and BERT by using Training dataset that is split form the original dataset. After the models are trained, they will test with the test dataset in the testing phase. The trained POS tagger will fed with the sentences as input and the corresponding POS tags as the expected outputs.

3.4 Results and its analysis

In this step, we will check whether the predicted tag and the expected tags are same. We will make a count on how many of the words will be correctly tagged and how many will be falsely tagged. Based on these we will calculate the accuracy of our model. The precision, recall and f1 score will also be measured for each model. To remove the issue of overfitting and underfitting, cross validation technique will be also used, where the dataset will divide in the three folds and in each iteration two of the folds will be taken as training set and the remaining one will be taken as the testing set.

These parameters are used to compare the performance of the implemented models.

3.5 Validation Criteria

Once a model is developed, it is very important to check the performance of the model. To measure the performance of a predictor, there are commonly used performance metrics such as confusion matrix. In classification problems, the primary source of performance measurements is confusion matrix.

3.5.1 Confusion Matrix

Confusion Matrix is a performance evaluation metric which provides a summary of the predictions made by a classification model, highlighting the correct and incorrect classifications across different classes. It is typically represented as a table with rows and columns corresponding to the predicted and actual classes, respectively. It helps in assessing the model's accuracy and identifying common types of errors.

Actual class

Predicted class

	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Fig. 3.5.1: Confusion Matrix

3.5.2 Accuracy

The overall accuracy of the model, calculated as
Accuracy = (TP + TN) / (TP + TN + FP + FN).

3.5.3 Recall

The proportion of actual positive instances correctly identified by the model, calculated as
Recall (Sensitivity or True Positive Rate) = TP / (TP + FN).

3.5.4 Precision

The ability of the model to correctly identify positive instances, calculated as
Precision = TP / (TP + FP).

3.5.5 F1 Score

A combined metric that considers both precision and recall, calculated as
F1 Score = 2 * (Precision * Recall) / (Precision + Recall).

3.5.5 K-fold Cross Validation

K-fold cross-validation is a technique used for model evaluation and performance estimation in machine learning. It involves dividing the dataset into k equal-sized folds and iteratively training and testing the model k times. In each iteration, a different fold is used as the testing set while the remaining folds are combined as the training set. The model's performance is evaluated on each iteration, and the performance metrics are averaged to provide an overall estimate of the model's performance. K-fold cross-validation allows for better utilization of the data, reduces the risk of overfitting or underfitting, and provides insights into the model's generalization performance. Stratified k-fold cross-validation can be used to preserve the class distribution in each fold, especially for imbalanced datasets. Overall, k-fold cross-validation is a widely used technique for reliable model evaluation and selection.

Chapter 4

EXPECTED OUTCOME

The key expectations from this research paper are listed below:

i. This paper will provide the best supervised classification model that could have a higher accuracy rate to assign the tag for Nepali text.

ii. This research will improve the word sense disambiguation (WSD).

iii. This research will also enhance the accuracy of unknown text

iv. This research paper will beneficial for various advanced NLP tasks like speech and natural language processing applications, semantic analysis, machine translation, text-to-speech conversion, question answering, speech recognition, and information retrieval, text summarization, Named entity recognition (NER) etc.

APPENDIX A: GANTT CHART

The project will be five months long and the works are divided accordingly. The planned schedule for the project are illustrated in Gantt Chart below:

Months Tasks	May	Jun	July	Aug	Sep
Identify Research Area	■				
Literature Review	■				
Identify necessary technologies	■	■	■	■	
Design Methodology	■				
Proposal Defense	■				
Datasets related work		■			
Empirical Analysis		■	■	■	
Appraisal of research and make required changes		■	■	■	
Mid-term Defense			■		
Final Defense				■	
Documentations	■	■	■	■	■

REFERENCES

[1] I. Shrestha, S. S. Dhakal, “Fine-grained part-of-speech tagging in Nepali text,” *Procedia Computer Science*, vol. 189, PP 300-311, 2021

[2] A. Pradhan, A. Yajnik, “Probabilistic and Neural Network Based POS Tagging of Ambiguous Nepali text: A Comparative Study,” *ISEEIE 2021: 2021 International Symposium on Electrical, Electronics and Information Engineering*, PP. 249–253, feb 2021

[3] A. Yajnik, “Part of Speech Tagging Using Statistical Approach for Nepali Text,” *International Scholarly and Scientific Research & Innovation*, vol. 11, no. 1, 2017

[4] S. Sayami, T. B. Shahi and S. Shakya, “Nepali POS Tagging using Deep Learning Approaches,” *NU. International Journal of Science*, Dec 2019.

[5] A. Paul, B. S. Purkayastha, S. Sakar, “Hidden Markov Model Based Part of Speech Tagging for Nepali Language,” *International Symposium on Advanced Computing and Communication (lSACC)*, Sep 2015.

[6] A. Yajnik, “GENERAL REGRESSION NEURAL NETWORK BASED POS TAGGING FOR NEPALI TEXT,” 4th *International Conference on Natural Language*, pp. 35–40, 2018.

[7] Tej Bahadur Shahi, Tank Nath Dhamala, Bikash Balami, “Support Vector Machines based Part of Speech Tagging for Nepali Text,” *International Journal of Computer Applications* (0975 – 8887), Volume 70– No.24, May 2013.

[8] A. Yajnik, “ANN Based POS Tagging For Nepali Text,” *International Journal on Natural Language Computing (IJNLC)*, Vol.7, No.3, June 2018.

[9] Greeshma Prabha, P. V. Jyothsna, K. K. Shahina, B. Premjith, K. P. Soman, “A Deep Learning Approach for Part-of-Speech Tagging in Nepali Language,” *International*

Conference on Advances in Computing, Communications and Informatics (ICACCI), Sep 2018.

[10] M. Jayaweera, N. G. J Dias, "HIDDEN MARKOV MODEL BASED PART OF SPEECH TAGGER FOR SINHALA LANGUAGE," *International Journal on Natural Language Computing (IJNLC),* Vol. 3, No.3, June 2014.

[11] P. Sinha, N. M. Veyie, B. S. Purkayastha, "Enhancing the Performance of Part of Speech tagging of Nepali language through Hybrid approach," *International Journal of Emerging Technology and Advanced Engineering,* Vol. 5, Issue 5, May 2015.