Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

A Comparison of Conditional Random Fields and Hidden Markov Model for The Nyishi Part of The Speech Tagging Task

Joyir Siram Murtem[1], Dr. Koj Sambyo[2], Dr. Achyuth Sarkar[3]
[1]*CSE Department, NIT Arunachal Pradesh, India*
[2]*CSE Department, NIT Arunachal Pradesh, India*
[3]*CSE Department, NIT Arunachal Pradesh, India*

Abstract:- Part-of-speech (POS) tagging is used to identify the grammatical function of words in a document. POS refers to word clusters that have common grammatical properties. Nouns, verbs, adjectives, adverbs, pronouns, adverbs, conjunctions, and prepositions make up the majority of POS in English. An estimated three lakh Nyishi people speak the Tani branch of the Sino-Tibetan language, making them one of the most populous ethnic groups in the Indian state of Arunachal Pradesh. Using Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs), this paper discusses POS tagging research for the Nyishi language. In Nyishi, the POS tagging challenge is paired with the word identification difficulty, making it harder to solve than its English equivalent. In this research, the authors have developed a tagset and POS tagging task. Experiments showed that compared to the HMM approach, the proposed technique (CRF) for the Nyishi language achieves higher F-Measure (87%), Precision (89%), and Recall (89%).

Keywords*: Hidden Markov Model, Conditional Random Field, Natural Language Processing (NLP), Tagging, feature extraction.*

1. Introduction

"POS tagging" refers to the method through which a text's words are labeled according to a Part of Speech (POS). Natural language processing (NLP) relies heavily on the process of labeling each word in a text with a tag that represents the lexical category to which the word belongs [1]. The tagged or annotated text may then be used in processes like machine translation, information retrieval, and data extraction [2-3]. POS Tagging is a fundamental component and significant application area in natural language processing. As a result, POS is a hot topic in natural language processing (NLP) tagging, the process of assigning a syntactic label to each word in a document based on the context in which it is found [4]. Automatically labeling words in a phrase using POS tags is called grammatical tagging [5]. Using NLP is as commonplace as using any other piece of modern technology. Machine translation, search, and voice-to-text query response are just some of the possible applications [6]. Using a computer and a number of different technologies and theories, NLP analyses texts mechanically. Natural language processing is also known as computational linguistics [7]. Tagging segments of speech are seen in Figure 1.

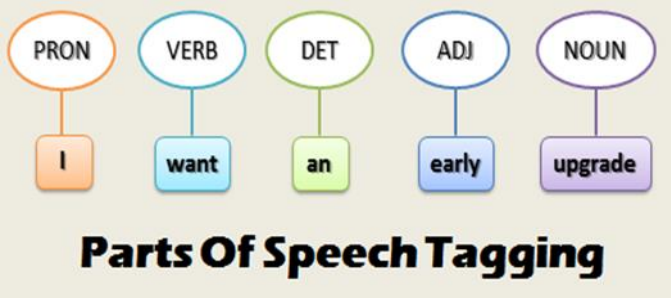

Figure 1. Part of Speech Tagging [8]

The native language of the Nyishi people, who live in the state of Arunachal Pradesh in India, is called "Nyishi," which literally means "the land of people" or "the human beings." It serves as an all-encompassing phrase that includes all of the variants of the Tani language spoken in the region, including a variety of dialects such as Aka Lel, Bangni, and Nishang, amongst others [9]. The Nyishi people, who speak a Tibeto-Burman dialect, make up the biggest scheduled tribal group in Arunachal Pradesh. They are classified as a Tibeto-Burman language group. They have a tremendous impact not just on the culture and economy of the larger Arunachalee society as a whole but also on the culture and economy of the tribal group itself [10]. There is no other grammatical system in the world that can be directly compared to that utilized by Nyishi. There are no gendered third-person pronouns in the Nyishi language, despite the fact that the language possesses gendered third-person nouns. Although there are a few disyllabic nouns in the Nyishi language, the majority of its verbs are one syllable each. A particular identifier serves as a way to tell the difference between the masculine and feminine genders. In addition to this, it makes use of its very own individual set of cardinal numbers. They were able to locate a sizeable number of previously unknown entries in the Nyishi dictionaries and add them to the existing collection. In terms of the English language, this is a matter of relatively little importance. There are a total of 28 distinct letters in the Nyishi alphabet, including 18 different consonants, 7 different vowels, 2 clusters, and 1 glottal [11]. When it comes to the many different parts of speech, the grammar of the Nyishi language follows virtually the same patterns as the grammar of the English language. Figure 2 shows an example of the Nyishi language.

Present Tense	**I read**	**ngo poorywn**
Past Tense	**I read**	**ngo poorypan**
Future Tense	**I will read**	**ngo poorytaywn**

Figure 2. Example of Nyishi Language [12].

Arunachal Pradesh's native language is Tibeto-Burman. Most of them speak Tani, a Tibeto-Burman language. In central Arunachal Pradesh, Nyishi, Adi, Galo, Apatani, Bangni, Tagin, Hills Miri, Bokar, Milang, and others arose as Tani languages. American Standard Version (ASV) study for under-resourced North-East Indian languages continues with Arunachal Pradesh's Adi, Apatani, Galo, and Nyishi. Parts of a speech corpus, as well as the pros and cons of current and publicly available corpora, are examined [13].

Like in any other language, POS labeling is crucial in the Nyishi language. They use a language called Nyishi in Arunachal Pradesh, India. Different Nyishi dialects have a common linguistic ancestor with the Tani language family. POS tagging is used to classify each Nyishi word as a noun, verb, adjective, adverb, and so on, which aids in understanding the language's syntactic and grammatical intricacies. Accurate performance of tools like IR, MT, and sentiment analysis on Nyishi text may be achieved with the use of this tagging. In addition to ensuring the continuity of the Nyishi language and its rich cultural heritage, this aids in its documentation, language revitalization, and protection of literary works. In addition, POS tagging is an invaluable tool for preserving Nyishi and its multiple dialects as cultural objects, studying Nyishi linguistics, and advancing language technology.

The need to analyse Nyishi language data in NLP applications is the focus of this study. It analyzes the performance of Conditional Random Fields (CRF)and Hidden Markov Models (HMM) for POS tagging in Nyishi [14]. The study develops new features for entity recognition, such as the "bag of words" and binary features. The results demonstrate that the CRF-based POS tagging system performs better than HMM using the same characteristics. Challenges include tag design and language comprehension, but this study represents a step forward for Nyishi-specific POS tagging methods. For processing the Nyishi language, we offer the CRF and the HMM models.

This paper aims to investigate A comparison of conditional random fields and hidden Markov model for the Nyishi part of the speech tagging task. Section 1 introduces the topic, while Section 2 reviews various techniques focused on this alteration. In Section 3, a background study is discussed. Section 4 describes the problem formulation. Section 5 outlines the objectives of the proposed research. Section 6 explains the research methodology and various techniques. In Section 7, a detailed discussion of the results of the experiments is presented. Finally, Section 8 provides the conclusion of the study.

2. Literature Of Review

There is a variety of work provided by many writers for POS tagging for the nyishi language using HMM, which is listed below:

Deshmukh et al., (2020) [15] evaluated that POS tagging is often performed as a preliminary task before more advanced NLP operations can be performed. Although Marathi is widely spoken in India, there are presently limited NLP resources available for the language. Accurate POS tagging is essential for many NLP tasks, including sentiment analysis, named entity identification, dependency parsing, etc. In order to facilitate POS tagging in the Marathi language, the study recommends employing a deep learning model with a Bidirectional Long Short-Term Memory (Bi-LSTM) model. The average accuracy of the models was 85% (using deep learning) and 97% (using a more traditional approach) (Bi-LSTM). They have made three contributions: the development of a deep learning model, the creation of a Bi-LSTM model, and a comparison of these two models to those based on machine learning for the same dataset.

KHAN et al., (2019) [16] suggested that POS tagging is difficult because it is a highly inflected and developed language. In several fields of natural language processing, including voice recognition, information extraction, machine translation, and others, POS tagging is used as an initial phase in the analysis of the written text. It's a task that, while the text is being written, analyses each word and determines what syntactic category it belongs in, then applies that tag to all the relevant terms. The previous work is developed on this new project. They previously introduced a POS tagger that uses a CRF and a set of features that are both language-dependent and -independent. They have statistically compared the results of each model on two standard datasets. In this analysis, authors compare three different types of models: the CRF, the support vector machine (SVM), and the bigram HMM (a variation of the n-gram Markov model). The results show that the DRRN models perform best on the BJ dataset, while the CRF-based model outperforms the SVM, RNNs, and n-gram approaches on the CLE dataset.

Priyadarshi et al., (2019) [17] analysed the Indian language's POS tagging it's not known about Maithili. They identified POS taggers and tagged corpora for several other Indian languages (Hindi, Bengali, Tamil, Telugu, Kannada, Punjabi, and Marathi) but not for Maithili. A minority language, Maithili has over 50 million native speakers and is recognized as an official language in India. Several Indian states currently utilize Maithili for official and instructional purposes, requiring the development of Maithili NLP methods and resources crucial. They conducted studies using many different feature sets and found an average accuracy of 82.67%.

Kanakaraddi et al., (2018) [18] focused on one of the primary technologies in NLP is the POS tagger. POS tagging is used in a wide range of application processing, including information retrieval, machine learning information processing, question answering, speech recognition, word sense disambiguation, and so on. A POS tagger, which identifies each word in a sentence with its grammatical function, is an effective component. Language's ambiguities make it challenging to design an effective POS tagger. The purpose of this study is to provide a comprehensive overview of the many various POS taggers developed by researchers and organizations. Each tagger would use its own unique set of tags. Tagging in natural language processing is equivalent to tokenization in programming languages. It can be difficult for POS taggers to select the most suitable tag in any given situation. Several studies have attempted to clarify the difference.

Islam et al., (2017) [19] recommended the ability to communicate effectively through their natural languages is essential for humans. In computer science and computational linguistics, NLP is a significant field of research and practical application that investigates how computers can be trained to comprehend and modify the text or

speech of natural languages. Machine Translation (MT), Word Sense Disambiguation (WSD), Wordnet, Part of Speech Tagging (POST), and Electronic Dictionaries (E-dictionaries) are five of the most significant and one of the key applications of NLP that have been created for North East (NE) natural languages and are widely used in daily life. These applications are very beneficial and provide just an intermediate and required step before the creation of more advanced NLP applications. The basic goal of this research is to examine the current works of these applications produced for NE natural languages in India, focusing on their significance, methodologies, and characteristics.

Suleiman et al., (2017) [20] presented that one of the most interesting and innovative areas in computer science is NLP. Both rule-based and statistical methods, as well as hybrids of the two, can be used to process natural language. HMM is the most often used model for statistical NLP. In several contexts of NLP, HMM serves as a useful experimental technique. Hidden Element Text classification, named entity identification, and morphological analysis are only some of the models available with a Markov. Results from a comparative study revealed that HMMs can be employed in several stages of NLP, especially in the pre-processing phase, where they excel at tasks including part-of-speech identification, morphological analysis, and syntactic structure.

Stratos et al., (2016) [21] investigated HMMs that are optimized for this task to solve the issue of unsupervised POS tagging. They refer to these HMMs as anchor HMMs, and they assume the relatively innocuous constraint for POS tagging (e.g., "the" occurs exclusively under the determiner tag) that each tag relates to at least one word that can have no other tag. They apply this assumption to create a sustainable estimator for anchor HMMs, which is based on an extension of the non-negative matrix factorization framework. Experiments show that the approach can hold its own against industry-standard methods like clustering and the log-linear model. Words are used to automatically lexicalize hidden states, making the model more understandable.

Kumawat et al., (2015) [22] intended that the linguistic classification of words into their respective POS is known as POS categorization. POS tagging is represented as one of the fundamental mandatory processes in different NLP demonstrations, including word intelligence disambiguation, information recovery, information management, analysis, questioning, and machine interpretation. The challenging goal of developing an efficient and accurate POS Tagger is to classify the difficulties in linguistic and philological elements. According to the study results, POS taggers have only been developed for the Indian languages of Hindi, Punjabi, Bengali, and Dravidian languages. Additionally, generic POS taggers for the languages of Hindi, Telugu, and Bengali have been developed. All predefined POS taggers rely on a wide-ranging Tag-set developed by a wide variety of groups and people. POS taggers and POS tag sets are crucial computational verbal tools for many NLP presentations, and they have seen a lot of growth in Indian languages recently. There is a wide range of author who has given their findings, which are given below in Table 1.

Table 1. Comparison of the different authors and their findings.

Author	**Technique**	**Result**
Deshmukh et al., (2020) [15]	deep learning & Bi-LSTM model	The average accuracy of the models was 85% (using deep learning) and 97% (using a more traditional approach) (Bi-LSTM). They have made three contributions: the development of a deep-learning model.
KHAN et al., (2019) [16]	SVM & HMM	A POS tagger that uses a CRF and a set of features that are both language-dependent and -independent. They have statistically compared the results of each model on two standard datasets.
Priyadarshi et al., (2019) [17]	POS taggers	Maithili NLP methods and resources are crucial. They conducted studies using many different feature sets and found an average

		accuracy of 82.67%.
Kanakaraddi et al., (2018) [18]	POS tagger	A comprehensive overview of the many various POS taggers developed by researchers and organizations.
Islam et al., (2017) [19]	MT, WSD, POST, and E-dictionaries	The current works of these applications are produced for NE natural languages in India.
Suleiman et al., (2017) [20]	Markov technique	A comparative study revealed that HMMs can be employed in several stages of NLP, especially in the pre-processing phase.
Stratos et al., (2016) [21]	HMMs	Industry-standard methods like clustering and the log-linear model.
Kumawat et al., (2015) [22]	POS tagging	POS taggers and POS tag sets are crucial computational verbal tools for many NLP presentations, and they have seen a lot of growth in Indian languages recently.

3. Background Study

Applications in NLP include question-answering, information extraction, and machine translation. One of the recognized languages of the Kingdom is Nyishi. There is an increasing demand for technologies that can handle this Nyishi data because of the quantity of such data already the information from websites. This research evaluated how the CRF and the HMM affected the Nyishi part-of-speech tagging issue. There are three phases to the proposed system: pre-processing, feature extraction, and building model. The authors introduced essential properties like a bag of words in the window and a bag of POS tags in the window to enable the system to recognize multi-word entities.

The author has also worked to enhance the part-of-speech tagger's output tags, which has helped the system better differentiate between name entities and non-entities. Authors have also used the yes/no characteristics: "Is a person," "Is a pre-name," "Is a place," and "Is an organization." The results demonstrated that the conditional random field-based POS system outperformed the HMM-based POS tagging system using the identical feature set [23].

4. Problem Formulation

The CRF model uses a (very) basic feature extraction technique to evaluate tag relations. Tagging is a significant building ingredient and application in NLP. The research will create POS tagging techniques for the Nyishi language, allowing for more complete data identification. There are several problems associated with this tiered tagging approach. Firstly, a well-designed abrasive tag design is crucial for success.

Additionally, it necessitates a complete understanding of the tag set and language. Furthermore, the model can be developed with minimum language or tagging scheme understanding. CRF and HMM are the names of the two distinct models that the author constructed to solve language problems.

5. Research Objectives

- To use the possibility of observation tagger for providing tags to words.
- To provide a sense of the significance of language processes that the new collection of features can represent.
- To enhance the accuracy of CRF-based POS tagging for utilization of sub-label dependency structure.
- To identify words in the Nyishi POS corpus using the CRF technique

6. Research Methodology

The authors were able to create a POS tagging system for the Nyishi language by combining CRF and HMM. Natural language processing relies heavily on POS tagging as a foundational component and technique. Today, NLP is not only useful but essential to everyday life. It is also understood to refer to a method of using computers to analyze and understand human language. When compared to its English counterpart, POS tagging in Nyishi is more challenging because the issue of word identification must also be tackled.

a) Technique Used

This section discusses the techniques used in the proposed methodology.

(i) Hidden Markov Model

HMMs are a kind of calculative Markov model in which the modeled system is believed to be a Markov process with hidden (or invisibly distributed) states. The probability of state changes is the sole parameter in a standard Markov model and is presented explicitly. On the other hand, with an HMM, the states directly are hidden, but the state-dependent output is shown. Computer-assisted processing of human language is the focus of natural language processing, a multi-disciplinary discipline that focuses inspiration on linguistics, AI, and computer science. The probabilities of components of speech in single words, pairs of words, triples, and longer sequences of words can be recognized and identified using HMM. If they consider an article and a verb, for instance, the probabilities of the next word being a verb are substantially lower than the possibilities of it being a preposition, another article, or a noun [24]. The HMM is a popular statistical approach to POS tagging. The affix tree technique, which can deal with the Out-of-Vocabulary (OOV) word issue and affixation, has been used to enhance HMM for the Indonesian language. The difficulty is that the affix tree does not include any information for processing clitics. Therefore, this research suggests morphological analysis for POS tagging in Indonesian. They use a hybrid HMM approach for better results in POS tagging. It has been shown that the probabilistic-based technique known as HMM can provide very precise results when used for classification [25].

A Hidden Markov Model (HMM) may be utilized to find the best possible solution when a classification issue has a natural representation of states as a sequence. The HMM produces a string of output symbols as its result. The state sequence that perfectly suited a given observation series remains a secret when using an HMM. They have used an HMM. The system is divided into three distinct parts to automate the process of tagging POS in text written in natural languages. Figure 3 shows the three parts of the HMM-based tagger. The system needs context for the challenge of POS disambiguation. The data can be encoded in several different ways, depending on the source of the knowledge. This model is what languages refer to as a "linguistic model." The parameters □□ (□, A, B) of the model describe the language model for HMM. The goal is to use corpora to estimate the HMM's model parameters □ □ (□, A, B). The HMM's model parameters are calculated using the labeled data as a recommendation during supervised learning. The model parameters are re-estimated using unlabelled data during semi-supervised learning. The Baum-Welch method is used to re-estimate the model parameters. Taggers based on HMM models (both bigram and trigram) would be implemented [26].

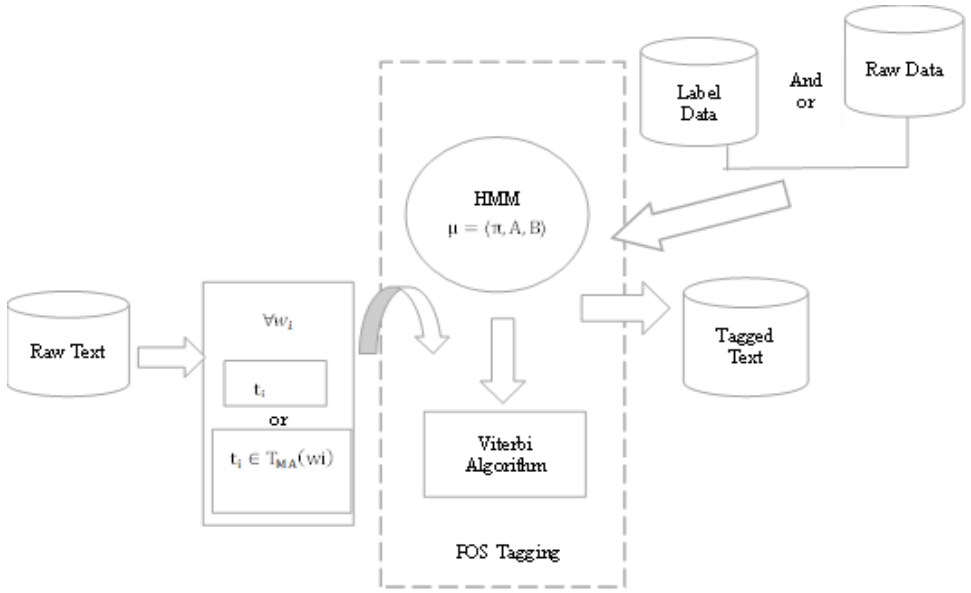

Figure 3. HMM-based POS tagging [26].

(ii) Conditional Random Fields

The conditional probability of values on specified output nodes given values on other designated input nodes is computed using CRFs, which are undirected graphical models. In this study, they present the results of applying the statistical CRF method to the extensive process of POS tagging in Bengali. Twenty-six POS tags specific to Indian languages were used in the development of the POS tagger. The equation for the conditional probability of a state sequence $S =< s_1, s_2, .., s_r >$Given a series of observations $O =< o_1, o_2, .., o_r >$is as follows:

$$P_\Lambda(s|o) = \frac{1}{z_0} exp\left(\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(s_t \quad {}_1s_t, o, t)\right)$$

For each feature function $f_k(s_t \quad {}_1s_t, o, t)$, the training-dependent weight λ_k It must be determined. Typically, the values of the feature functions are binary, although they can range anywhere between $-\propto \cdots . +\propto$. [27]. Since generative models have difficulty (or are unable) to describe unlimited overlapping and non-independent features, discriminative models like CRFs are beneficial for many kinds of NLP challenges [28]. CRFs are a kind of conditional-probabilistic model, and they have adapted the strategy for POS tagging. CRFs have been shown to achieve state-of-the-art results among the several sequential models used for segmentation and labeling tasks. Considering the ability to effectively train on many multiple, overlapping, and independent features is essential for CRFs. Furthermore, CRF deals with the HMM dependency on data independence and the maximum entropy model's label bias problem [29].

b) Proposed Methodology

In this section, various steps of the proposed methodology are discussed below. Using HMM and CRF, the planned technique of POS tagging for the Nyishi language is depicted in Figure 4. The three-step process recommended for Nyishi part-of-speech tagging is as follows: Steps include (1) initial possession, (2) feature extraction, and (3) model construction.

- Dataset

The Nyishi POS tagging method has been included in the suggested system. Researchers also need to resolve the problem of identifying words, making it significantly more complex than its English equivalent to solving the Nyishi POS tagging problem. Thirty-six item tag sets have been created, each of which includes 28 distinct letters, 18 distinct consonants, 7 distinct vowels, 2 clusters, and 1 glottal for the Nyishi POS system.

Additionally, more than 25000 entries from the Nyishi-to-English Dictionary have been collected, and this was done both manually and automatically.

- **Pre-processing**

In this section, the author follows some steps:

• Segmentation: Segmentation can be used to detect the limitations between the languages and determine where one language ends and another begins when a text is written in more than one language. It increased their rate of accuracy from 95% to 97 %, which was beneficial to the research during POS. The CRF and HMM tools are used for text segmentation in the research.

• Data Cleaning: Data cleaning is the process of converting the pattern of certain words into the correct format so that the system can process data more effectively and with more effectiveness.

• Normalization: The process of turning a word into its basic form is referred to as normalization. The word order form of a letter is changed to its normal form during the process of normalization. This allows the base form to be more easily understood. This process is important. After all, the model could be trained and then fail to understand anything because different people write and speak the same words in different ways. Furthermore, a human guidebook may be unable to recognize it. As a result, all gazetteers, listings, and databases have been updated.

- **Extraction Feature**

The Authors have used some features for each letter and word:

Word itself: The letters and words are recognized as a feature in the system that authors have presented.

• Bag of words: This bag contains the word in its entire duration. The word that comes before the word itself, as well as the word that comes after the word itself. Therefore, it presents both the classifiers and the word embedding representation of the phrase on the screen.

• Bag of POS Tag: Each tag has unique classifiers and word embeddings that correlate to the POS tag on the screen. The property is very important because the POS tag differentiates between the entity and the non-entity. To improve the accuracy of the tags, the POS Tagger is trained using the manually tagged corpus. The Nyishi portion of the dataset was used to hone the POS corpus's training. The organization has manually tagged the POS tagger training. Tag combination, which can be initially found in the Computational Linguistics and Psycholinguistics Research Center (CLIPS).

• Is a Pre-Name binary feature: In order to maintain consistency with how the pre-name whitelist defines this feature value, it is only appropriate to assign it to one of the pre-names on the whitelist, or else to set it to zero. It has 123 distinct pre-name phrases that are not duplicated anywhere. It contains the people's nationalities, titles (such as president or professor), and certain adjectives that come before human entities.

• Is a Pre-Location Binary Feature: This feature value is defined by the pre-location (Pre-Loc) whitelist; if the word is on the pre-location whitelist, then it should be allocated to one; if it is not on the pre-location whitelist, then it should be assigned to zero. It contains the names of 45 pre-locations in their entirety. Some of the names and adjectives that come before the location names are included on the pre-loc whitelist.

• Is an Organization Binary Feature: This feature value is defined by our organization's Guidebook; if the term is in the organization guidebook, then it should be allocated to one; if the word is not in the organization guidebook, then it should be assigned to zero. It contains the full names of 424 well-known organizations from all around the world. The names of governmental organizations and non-governmental organizations are both included in the organization guidebook. The list of government organizations includes ministries, agencies, educational institutions, and armed forces. The names of clubs, political parties, corporations, sports teams, churches, mosques, and charitable organizations are all included in the list of non-governmental organizations (NGOs).

➢ **Building Model**

In this research, the authors investigated the CRF and HMM algorithms and see how they operate. During the process of developing a Nyishi POST model, they utilized the advantage of the screen's Classifier and word embedding characteristics.

An exclusionary associative probabilistic graphical model is referred to as a CRF. It would classify as an algorithm for labeling sequences. The conditional probability is increased to its maximum during training. They used the CRF tool. The CRF package includes a template file, as well as training and testing files to experiment. The template file was used to provide information on the classifier and word embedding capabilities. Each phrase will have its own row in the Training file, and each feature will have its own column; the identification will be in the very last column. The only difference between the training and test files is an empty value in the last column of the test file.

A stochastic method for POS tagging, known as HMM, is also accessible. Applications of HMMs include learning algorithms and bioinformatics, musical score following, partial discharges, temporal pattern recognition, handwriting, voice, and gesture recognition. Learning HMMs that are especially well-suited for the challenge is how they handle POS tagging. Figure 4 shows the architecture of the proposed methodology.

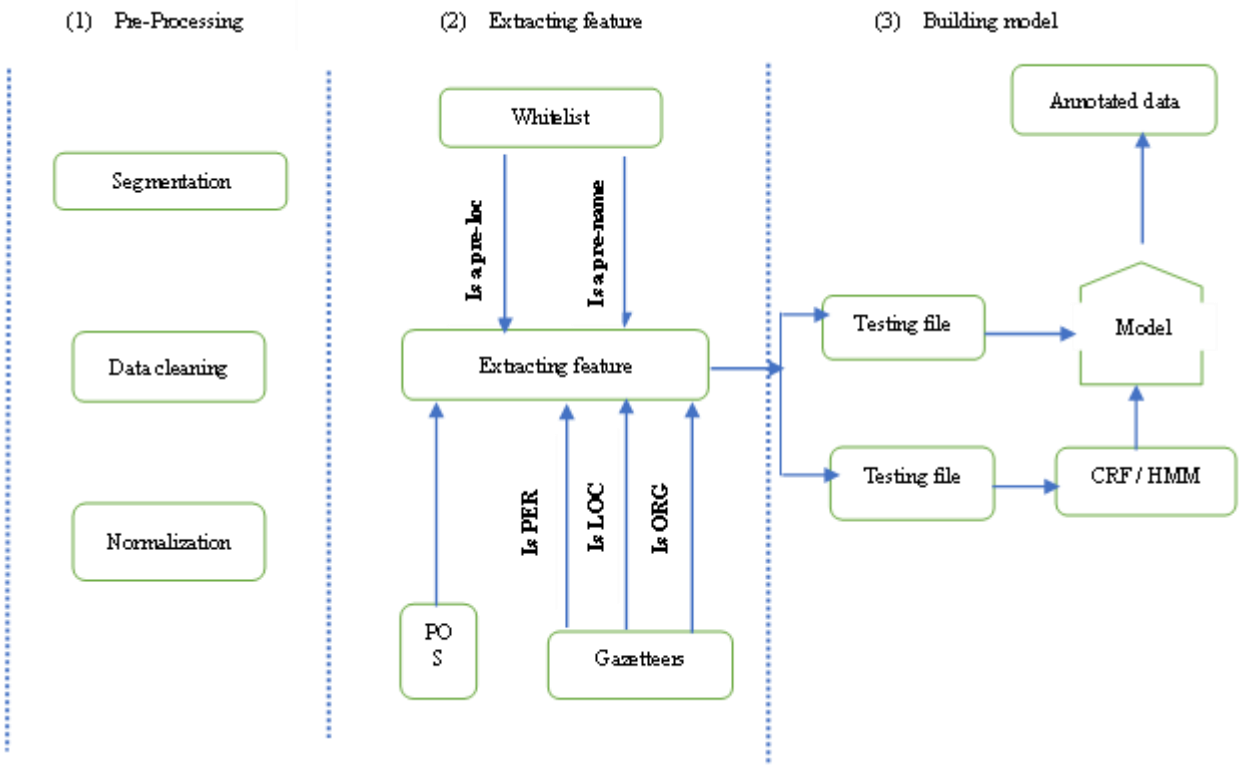

Figure 4. The architecture of Proposed Methodology

7. Experiment And Result

The following three parameters are standard evaluation metrics for classification applications:

$$Precision = \left(\frac{true\ positive}{true\ positive + false\ positive}\right) \quad (1)$$

$$Recall = \left(\frac{true\ positive}{true\ positive + false\ negative}\right) \quad (2)$$

$$F - Measure = 2 * \frac{precision * recall}{precision + recall} \quad (3)$$

After that, the average values for F-measure, Precision, and Recall are determined for each tag.

Models based on HMM and CRF were trained and tested using a similar set of data with a similar architecture. Table 2 shows the F-measure, precision, and recall value, which was found by the above-given formula.

Table 2. F-measure, Precision, and Recall of proposed HMM and CRF for the Nyishi language

Proposed	**Tagset**	**Techniques**	**F-measure**	**Recall**	**Precision**
Nyishi	36	HMM CRF	64.7% 87%	94% 96%	58.2% 89%

Table 3 gives the findings of a comparison between the proposed HMM and CRF model for POS tagging in Nyishi and other Indian languages that have employed the CRF and HMM model for POS tagging.

Table 3. Comparison with different current CRF and HMM POS taggers for other Indian languages.

Authors	**Languages**	**Techniques**	**F-measure**	**Recall**	**Precision**
Warjr et al., (2021) [30]	Khasi corpus	CRF	91%	92%	92%
Ayogu I. et al., (2017) [31]	Yoruba	HMM	46.2%	79%	95%
Khan W. et al., (2019)[29]	Urdu	CRF	86.9%	85.42%	90.2%
Cling D. et al., (2020) [32]	Myanmar	HMM	84%	92%	94%
Joshi N. et al., (2013) [33]	Hindi	HMM	92%	92%	92%
Proposed	**Nyishi**	**HMM/CRF**	**64.7%/87%**	**94%/96%**	**58.2%/89%**

In this paper, researchers used two methods, HMM and CRF, in which the proposed technique (CRF) has the highest F-measure (87%), precision (89%), and recall (89%) than the HMM method. The figure indicates a comparison analysis of the highest proposed method with other methods. In the graph, CRF has the highest precision value than the other method and has the lowest F-measure and recall than the other method. Figure 5 depicts the comparative analysis of our method with others.

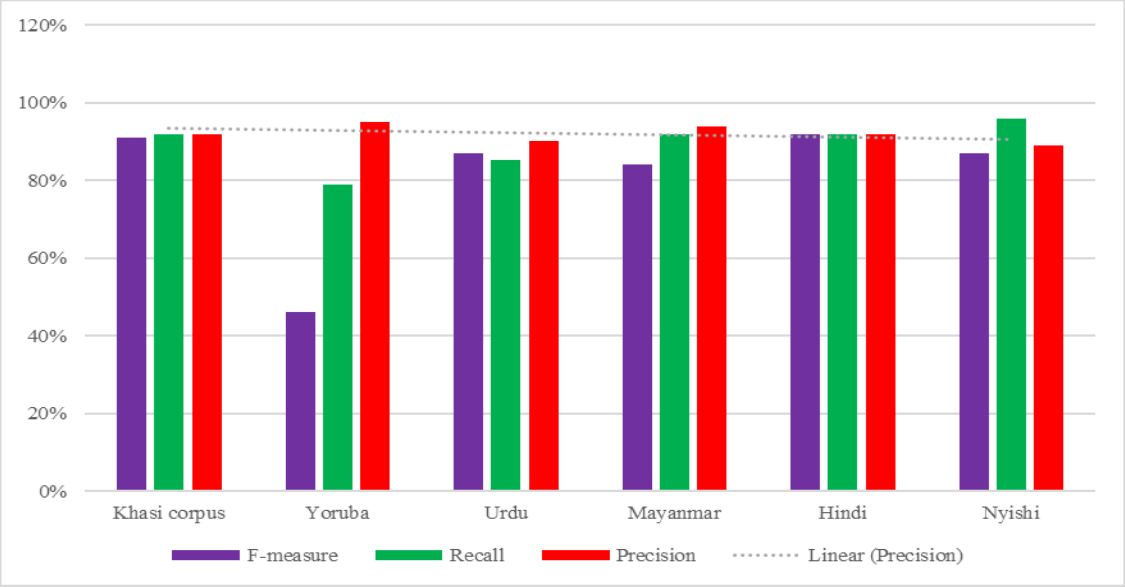

Figure 5: Comparative analysis of the proposed method with others

8. Conclusion

A POS Tagger classifies every word in a phrase into one of four categories: noun, adjective, verb, or adverb. There are several problems associated with this tiered tagging approach. Firstly, a well-designed abrasive tag design is crucial for success. Additionally, it necessitates a complete understanding of the tag set and language. In this paper, the authors investigate the use of the CRF and HMM methods for POS tagging in the Nyishi language. A custom-designed Nyishi POS task with around 36 tag sets to train a CRF and HMM model. The experimental results show the proposed technique (CRF) has the highest F-measure (87%), Precision (89%), and Recall (89%) than the HMM method. In the future, more Nyishi data must be collected to make sure that each word is properly tagged. Additionally, researchers would do experiments on different models using the planned Nyishi corpus.

9. Acknowledgement

At the opening of my research paper, I would like to express my profound gratitude to everyone who has assisted me in this quest. I would like to express my heartfelt gratitude to our research supervisors Dr.Koj Sambyo and Dr.Achyuth Sarkar for providing us with the opportunity to create this research paper on the topic 'A research of A comparison of conditional random fields and hidden Markov model for the Nyishi part of the speech tagging task', which allowed me to conduct an extensive study and learn about many new things. I also express my heartfelt thanks to my parents and family, who have always supported me morally and financially. Finally, my thanks go to all my friends who provided excellent advice and direction for completing my research paper. Finally, I Would like to thank everyone who has already been recognized.

Conflicts of interest

There is no conflict of interest

Refrences

[1] Y. Gu et al., "Domain-specific language model pretraining for biomedical natural language processing," ACM Trans. Comput. Healthcare, vol. 3, no. 1, pp. 1-23, 2022 [doi:10.1145/3458754].
[2] S. Thavareesan and S. Mahesan, "Word embedding-based part of speech tagging in Tamil texts" in 15th International conference on Industrial and information systems (ICIIS). IEEE. IEEE, 2020, pp. 478-482 [doi:10.1109/ICIIS51140.2020.9342640].
[3] M. Silfverberg et al., 'Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy.' Unknown Host Publication, 2014.
[4] T. Gung ¨ or, "Part-of-speech tagging" in Handbook of Natural Language Processing, 2nd ed. Chapman & Hall/CRC, 2010, pp. 205-235.
[5] M. Janssen, "NeoTag: A POS tagger for grammatical neologism detection" in LREC, 2012, pp. 2118-2124.
[6] I. Guy, "Searching by talking: Analysis of voice queries on mobile web search" in Proc. 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2016, pp. 35-44 [doi:10.1145/2911451.2911525].
[7] Chiche and B. Yitagesu, "Part of speech tagging: A systematic review of deep learning and machine learning approaches," J. Big Data, vol. 9, no. 1, pp. 1-25, 2022.
[8] Thinkinfi.com,extract-custom-keywords-using-nltk-pos-tagger-in-python,URL: https://thinkinfi.com/extract-custom-keywords-using-nltk-pos-tagger-in-python/,accessedat 26/09/23
[9] M. Dey, "Negation in Nyishi," The NEHU J., vol. 15, no. 2, pp. 79-100, 2017.
[10] R. Tarh et al., "Changing cultural practices among the Nyishis of Arunachal Pradesh: A contextual study," Int. J. Res. Anal. Rev. (IJRAR), vol. 5, no. 2, 2018.
[11] D. N. Prabhu Khorjuvenkar et al., "Parts of speech tagging for Konkani language" in Second International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2018, pp. 605-607 [doi:10.1109/ICCMC.2018.8487620].
[12] J. Siram et al., "Part-of-speech (POS) tagging for the Nyishi language" in Advances in Information Communication Technology and Computing. Singapore: Springer, 2022, pp. 191-199 [doi:10.1007/978-981-19-0619-0_17].
[13] U. Bhattacharjee and K. Sarmah, "Development of a speech corpus for speaker verification research in multilingual environment," Int. J. Soft Comput. Eng. (IJSCE), vol. 2, no. 6, pp. 443-446, 2012.

[14] I. I. Ayogu et al., "A comparative study of hidden Markov model and conditional random fields on a Yorùba part-of-speech tagging task" in International Conference on Computing Networking and Informatics (ICCNI). IEEE, 2017, pp. 1-6 [doi:10.1109/ICCNI.2017.8123784].
[15] R. D. Dhumal Deshmukh and A. Kiwelekar, "Deep learning techniques for part of speech tagging by natural language processing" in 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, 2020, pp. 76-81 [doi:10.1109/ICIMIA48430.2020.9074941].
[16] W. Khan et al., "Part of speech tagging in Urdu: Comparison of a machine and deep learning approaches," IEEE Access, vol. 7, pp. 38918-38936, 2019 [doi:10.1109/ACCESS.2019.2897327].
[17] Priyadarshi and S. K. Saha, "Towards the first Maithili part of speech tagger: Resource creation and system development," Comput. Speech Lang., vol. 62, p. 101054, 2020 [doi:10.1016/j.csl.2019.101054].
[18] S. G. Kanakaraddi and S. S. Nandyal, "Survey on parts of speech tagger techniques" in International Conference on Current Trends towards Converging Technologies (ICCTCT). IEEE, 2018, pp. 1-6 [doi:10.1109/ICCTCT.2018.8550884].
[19] Islam et al., "A study on various applications of NLP developed for North-East languages," Int. J. Comput. Sci. Eng., vol. 9, no. 6, pp. 386-378, 2017.
[20] D. Suleiman et al., "The use of hidden Markov model in natural Arabic language processing: A survey," Procedia Comput. Sci., vol. 113, pp. 240-247, 2017 [doi:10.1016/j.procs.2017.08.363].
[21] K. Stratos et al., "Unsupervised part-of-speech tagging with anchor hidden markov models," Trans. Assoc. Comp. Linguist., vol. 4, pp. 245-257, 2016 [doi:10.1162/tacl_a_00096].
[22] D. Kumawat and V. Jain, "POS tagging approaches: A comparison," Int. J. Comput. Appl., vol. 118, no. 6, 32-38, 2015 [doi:10.5120/20752-3148].
[23] M. Muhammad et al., "A comparison between conditional random field and structured support vector machine for Arabic named entity recognition," J. Comput. Sci., vol. 16, no. 1, pp. 117-125, 2020 [doi:10.3844/jcssp.2020.117.125].
[24] P. Alva and V. Hegde, "Hidden markov model for pos tagging in word sense disambiguation" in International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). IEEE, 2016, pp. 279-284 [doi:10.1109/CSITSS.2016.7779371].
[25] U. Afini and C. Supriyanto, "Morphology analysis for Hidden Markov Model based Indonesian part-of-speech tagger" in 1st International Conference on Informatics and Computational Sciences (ICICoS). IEEE, pp. 237-240.
[26] S. Dandapat, Part-of-Speech Tagging for Bengali. Department of Computer Science and Engineering Indian Institute of Technology Kharagpur, 2009.
[27] Ekbal et al., "Bengali part of speech tagging using the conditional random field" in Proc. Seventh International Symposium on Natural Language Processing (SNLP2007), 2007, pp. 131-136.
[28] S.H. Na, "Conditional random fields for Korean morpheme segmentation and POS tagging," ACM Trans. Asian Low Resour. Lang. Inf. Process., vol. 14, no. 3, pp. 1-16, 2015 [doi:10.1145/2700051].
[29] W. Khan et al., "Urdu part of speech tagging using conditional random fields," Lang. Resour. Eval., vol. 53, no. 3, pp. 331-362, December 2019 [doi:10.1007/s10579-018-9439-6].
[30] S. Warjri et al., "Part-of-speech (pos) tagging using conditional random field (crf) model for Khasi corpora," Int. J. Speech Technol., vol. 24, no. 4, pp. 853-864, 2021 [doi:10.1007/s10772-021-09860-w].
[31] I. Ayogu et al., "A comparative study of hidden Markov model and conditional random fields on a Yorùba part-of-speech tagging task" in International Conference on Computing Networking and Informatics (ICCNI). IEEE, 2017, pp. 1-6 [doi:10.1109/ICCNI.2017.8123784].
[32] D. L. Cing and K. M. Soe, "Improving accuracy of part-of-speech (POS) tagging using hidden markov model and morphological analysis for Myanmar Language," Int. J. Electr. Comput. Eng., vol. 10, no. 2, p. 2023,April 2020 [doi:10.11591/ijece.v10i2.pp2023-2030].
[33] N. Joshi et al., "HMM based POS tagger for Hindi" in Proceeding of 2013 international conference on artificial intelligence, soft computing (AISC-2013), 2013, pp. 341-349 [doi:10.5121/csit.2013.3639].