PPSU
P P SAVANI UNIVERSITY
School of Engineering

Natural Language Processing Project

on

Hindi ,Gujarati & English Part Of Speech Tagging

Academic Year: 2023-24

Student's Full Name	Jay Chauhan, Manan Patel, Vedant Rajpurohit
Enrollment No	20SE02ML009 , 20SE02ML034, 20SE02ML036
Branch	Artificial Intelligence and Machine learning
Semester	7TH

Supervised by

Mr.Ravirajsinh Chauhan & Ms. Megha S. Patel
P. P. Savani School of Engineering

PPSU
P P SAVANI UNIVERSITY
School of Engineering

CERTIFICATE

This is to certify that **Mr. Jay Chauhan**, Enrollment No**. 20SE02ML009** from the Department of ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING, has successfully completed the Natural Language processing Project on the **Hindi, Gujarati & English Part Of Speech Tagging** during Academic Year 2023-24.

Date:

_____	_____
Mr.Ravirajsinh Chauhan	**Ms. Megha S. Patel**

PPSU
P P SAVANI UNIVERSITY
School of Engineering

CERTIFICATE

This is to certify that **Mr. Manan Patel,** Enrollment No**. 20SE02ML034** from the Department of ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING, has successfully completed the Natural Language processing Project on the **Hindi ,Gujarati & English Part Of Speech Tagging** during Academic Year 2023-24.

Date:

Mr.Ravirajsinh Chauhan	**Ms. Megha S. Patel**

PPSU
P P SAVANI UNIVERSITY
School of Engineering

CERTIFICATE

This is to certify that **Mr. Vedant Rajpurohit,** Enrollment No**. 20SE02ML036** from the Department of ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING, has successfully completed the Natural Language processing Project on the **Hindi ,Gujarati & English Part Of Speech Tagging** during Academic Year 2023-24.

Date:

_____	_____
Mr.Ravirajsinh Chauhan	**Ms. Megha S. Patel**

ACKNOWLEDGEMENT

I am very thrilled and honored to offer my great gratitude and heartfelt appreciation to all those who have played a crucial role in the successful completion of this NLP project report.

First and foremost, I want to offer my sincerest thanks to my excellent mentors, **Mr.Ravirajsinh Chauhan & Ms. Megha S. Patel,** for their unyielding leadership, invaluable insights, and unwavering support over the duration of this project. Their extensive knowledge in the field of Natural Language Processing has been important in influencing the course of this project. Their kind and insightful coaching has not only improved the content of my report but has also provided me with an amazing learning experience that will certainly leave an enduring impression on my academic career.

Moreover, I wish to extend my gratitude to my fellow classmates and colleagues who offered constructive input and help during the project's evolution. Their collaborative approach and readiness to engage in meaningful debates considerably raised the caliber of this job.

JAY CHAUHAN 20SE02ML009

MANAN PATEL 20SE02ML034

VEDANT RAJPUROHIT 20SE02ML036

ABSTRACT

Part-of-speech (POS) tagging is a fundamental task in natural language processing (NLP) that involves assigning grammatical labels to words in a given text. This project focuses on developing POS taggers for two languages: English and Hindi. Accurate POS tagging is crucial for various NLP applications, including machine translation, sentiment analysis, and information extraction. By investigating the challenges and techniques specific to each language, this project aims to improve the accuracy and efficiency of POS tagging systems.

The project begins with an overview of POS tagging and its importance in NLP. The distinct characteristics of English and Hindi languages are explored, highlighting their diverse grammatical structures, morphological variations, and contextual nuances. Different linguistic features, such as word context, word morphology, and syntactic dependencies, are considered for POS tagging in each language.

For English POS tagging, the project examines traditional rule-based approaches, statistical models, and machine learning techniques. NLTK (Natural Language Toolkit) is a comprehensive Python library for natural language processing (NLP) tasks. It offers a wide range of tools, algorithms, and resources that facilitate the analysis, manipulation, and understanding of human language. NLTK provides functionalities for tasks such as tokenization, stemming, lemmatization, POS tagging, named entity recognition, and sentiment analysis In the case of Hindi, which is a morphologically rich language, the project investigates the challenges posed by its extensive inflectional and derivational morphology. Specific techniques, such as morphological analysis, stemming, and lemmatization, are employed to handle morphological variations and improve POS tagging accuracy. The project also explores the use of linguistic resources, such as Hindi WorldNet and morphological analyzers, to aid the POS tagging process.

Overall, this project contributes to the field of NLP by developing and evaluating POS taggers for English and Hindi. By addressing the specific challenges and characteristics of each language, the project aims to improve the accuracy and efficiency of POS tagging systems, enabling better NLP applications.

CHAPTER 1: INTRODUCTION OF PROJECT

Part-of-speech (POS) tagging is a fundamental concept in natural language processing (NLP) that involves assigning grammatical categories, or "tags," to each word in a sentence. These tags represent the syntactic role and grammatical function of each word within the context of a sentence. POS tagging is a crucial preprocessing step for various NLP tasks, as it provides valuable information about the structure, semantics, and relationships between words in a text.

In NLP projects, POS tagging serves as a building block for many advanced language understanding applications. Whether you're developing a Chabot, a sentiment analysis tool, a machine translation system, or any other language-related application, understanding the POS of words can significantly enhance the accuracy and performance of your model. This introduction will outline the importance of POS tagging in NLP projects and its role in various applications.

OBJECTIVES:

- High Accuracy: One of the primary objectives of a POS tagging project is to achieve a high level of accuracy in predicting the correct part-of-speech tags for words in a given sentence. The accuracy of your model's predictions will significantly impact the quality of downstream NLP tasks that rely on accurate POS information.

- Generalization: It's important for your POS tagging model to generalize well to unseen data. The model should perform well not only on the training data but also on new, previously unseen sentences. This ensures that your model is robust and can handle various language styles, domains, and contexts.

- Multi-lingual Support: If your project involves multilingual text, an objective might be to develop a POS tagging system that can accurately

tag parts of speech in multiple languages. This can be particularly challenging due to the grammatical differences across languages.

- Efficiency and Speed: Depending on the application, you might want your POS tagging model to be efficient and able to process text quickly. Achieving a good balance between accuracy and speed is important, especially for real-time or large-scale applications**.**

FUNCTIONALITIES OF PROJECT:

- Word Categorization: POS tagging categorizes each word in a sentence into a specific part of speech, such as noun, verb, adjective, adverb, pronoun, preposition, conjunction, and more. This categorization provides insight into the syntactic and grammatical structure of the sentence.

- Syntactic Structure Analysis: By assigning POS tags to words, POS tagging helps in analyzing the sentence's syntactic structure. It reveals how words relate to one another and how they contribute to the overall grammatical structure of the sentence.

- Grammar Correction: POS tagging can be employed in grammar correction tools to identify and rectify grammatical errors by analyzing the correct placement of words within a sentence.

- Information Extraction: In tasks like named entity recognition, POS tagging helps in identifying named entities based on the patterns of specific POS tags, such as proper nouns.

- Linguistic Analysis: Linguists use POS tagging to study linguistic phenomena, such as word order, morphological patterns, and syntactic rules, across various languages.

CHAPTER 1: INTRODUCTION OF PROJECT

1) Transformation-Based Error-Driven Learning and Natural

Language Processing:

Corpus-based methods are often able to succeed while ignoring the true complexities of language, banking on the fact that complex linguistic phenomena can often be indirectly observed through simple epiphenomena. For example, one could accurately assign a part-of-speech tag to the word race in (1-3) without any reference to phrase structure or constituent movement: One would only have to realize that, usually, a word one or two words to the right of a modal is a verb and not a noun. An exception to this generalization arises when the word is also one word to the right of a determiner.

2) Part-of-Speech (POS) Tagging Using Deep Learning-Based

Approaches on the Designed Khasi POS Corpus:

In the present designed Khasi POS corpus, each word is tagged manually using the designed tagset. Methods of deep learning have been used to experiment with our designed Khasi POS corpus. The POS tagger based on BiLSTM, combinations of BiLSTM with CRF, and character-based embedding with BiLSTM are presented. The main challenges of understanding and handling Natural Language toward Computational linguistics to encounter are anticipated. In the presently designed corpus, we have tried to solve the problems of ambiguities of words concerning their context usage, and also the orthography problems that arise in the designed POS corpus. The designed Khasi corpus size is around 96,100 tokens and consists of 6,616 distinct words. Initially, while running the first few sets of data of around 41,000 tokens in our experiment the taggers are found to yield considerably accurate results. When the Khasi corpus size has been increased to 96,100 tokens, we see an increase in accuracy rate and the analyses are more pertinent. As results, accuracy of 96.81% is achieved.

3) Part-of-Speech Tagging with Rule-Based Data Preprocessing and Transformer:

POS tagging for a word depends not only on the word itself but also on its position, its surrounding words, and their POS tags. POS tagging can be an upstream task for other NLP tasks, further improving their performance. Therefore, it is important to improve the accuracy of POS tagging. In POS tagging, bidirectional Long Short-Term Memory (Bi-LSTM) is commonly used and achieves good performance. However, Bi-LSTM is not as powerful as Transformer in leveraging contextual information, since Bi-LSTM simply concatenates the contextual information from left-to-right and right-to-left. In this study, we propose a novel approach for POS tagging to improve the accuracy. For each token, all possible POS tags are obtained without considering context, and then rules are applied to prune out these possible POS tags, which we call rule-based data preprocessing. In this way, the number of possible POS tags of most tokens can be reduced to one, and they are considered to be correctly tagged. Finally, POS tags of the remaining tokens are masked, and a model based on Transformer is used to only predict the masked POS tags, which enables it to leverage bidirectional contexts. Our experimental result shows that our approach leads to better performance than other methods using Bi-LSTM.

4) Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields:

We have trained a CRF on Gujarati which gives an accuracy of around 92%. From the experiments we observed that if the language specific rules can be formulated in to features for CRF then the accuracy can be reached to very high extents. The CRF learns from both tagged that is 600 sentences and also untagged data, which is 5,000 sentences. From the errors we conclude that as the training data increases, the less number of unknown words will be encountered in the test corpus, which will increase the accuracy. We can also use some machine readable resources like dictionaries, morphs etc. whenever they are built.

5) A Review on Part-Of-Speech Tagging on Gujarati Language

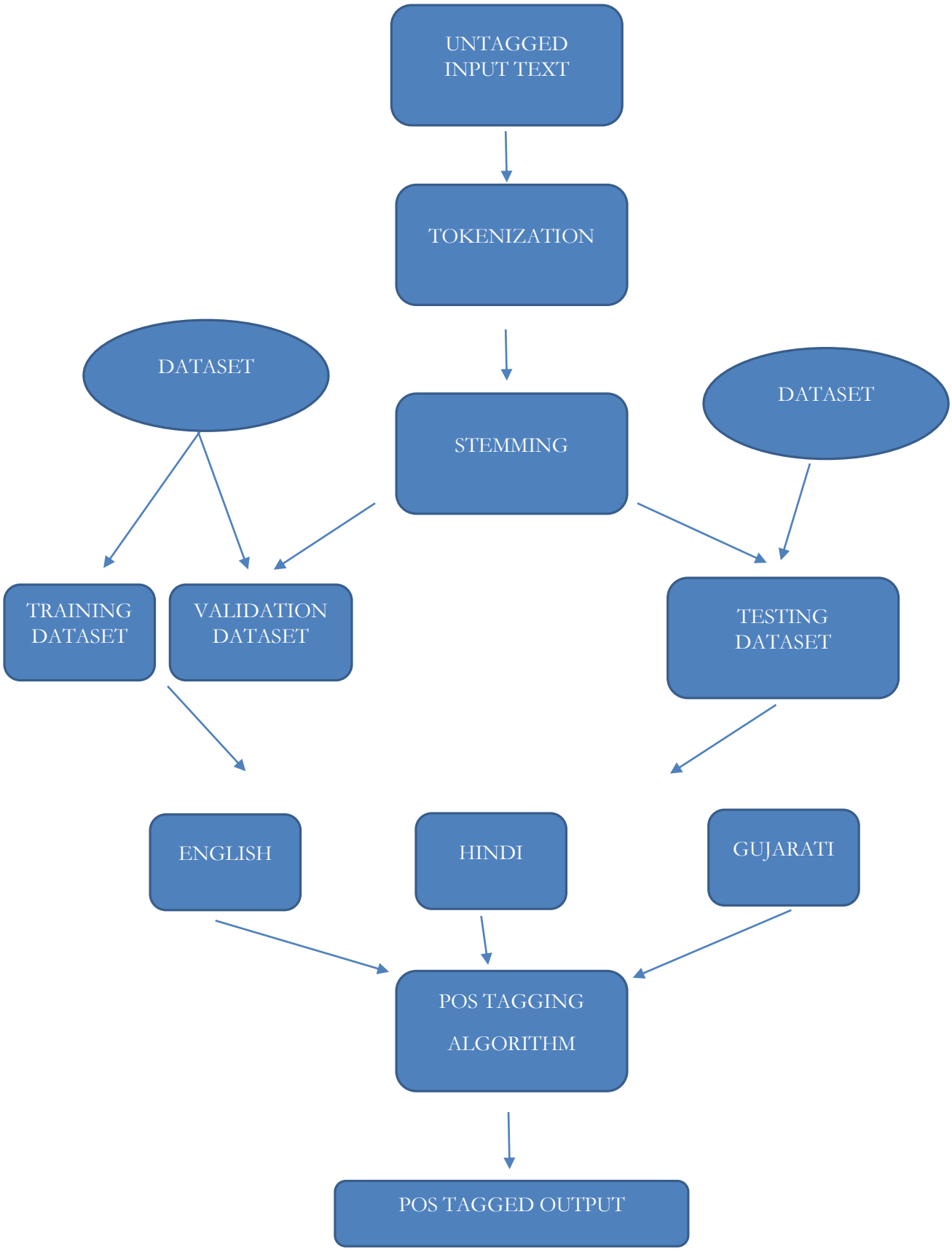
Algorithm:

i. Input the text using file upload button or manually enter by user.
ii. Tokenize the input text word by word.
iii. Normalized the tokenized words. I.e. separate out the punctuation marks and the symbols from the text
iv. Search the number tag by using Regular Expression. For Example: - ૨૦૧૨, ૧-૨, ૧૨મી etc.
v. Search the date tag by using regular expression. For Example: - ૧૭/૧૦/૧૯૧૭ etc.
vi. Search the time tag by using regular expression. For Example: - ૧૭: ૧૦, ૧૦: ૧૦: ૧૦ etc.
vii. Search for the abbreviation using regular expression. For Example: - એ. આર. કે etc.
viii. Search in database for different input words and tag the word according to corresponding tag.
ix. Then different rules are applied to tag the unknown words.
x. Display the tagged data to the user.

Overall comparison of different approaches:

Sr. No	Approach	Corpus Size	Reference	Accuracy
1.	Stochastic approach	351 words	"part of speech tagging using statistical approach for Guajarati text"	92.87%
2.	Stochastic approach - CRF	5000 words	"Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields"	89.90%
3.	Stochastic approach	5000 Statement	"a statistical Chunker for Indian language Guajarati"	96%
4.	Rule-based and Hybrid approach	8,525,649 Words	"Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati"	90.7%
5.	Stochastic approach - CRF	10000 words	"Improve accuracy of Parts of Speech tagger for Gujarati language"	92%

CHAPTER 3: DESIGN AND PLANNING



1) **Tokenization**: Tokenization in natural language processing (NLP) refers to the process of breaking down a text or a sequence of characters into smaller units, usually words, phrases, symbols, or other meaningful elements known as tokens. These tokens are the building blocks that can be easily analyzed and processed by computers. The goal of tokenization is to segment the input text into meaningful units that can be individually processed, manipulated, and understood by algorithms.

For example, consider the sentence: "Google is a powerful company." Tokenizing this sentence might result in the following tokens:

- "Google"
- "is"
- "a"
- "powerful"
- "company"
-

2) **Stemming**: Stemming in natural language processing (NLP) is the process of reducing a word to its base or root form, by removing suffixes or prefixes. The purpose of stemming is to normalize words so that variations of the same word, which might have different inflections or grammatical forms, can be treated as the same word. This helps in improving text analysis and information retrieval tasks.

Original words:

- jumping
- jumps
- jumped

After stemming:

- jump
- jump
- jump

3) **Different POS Tagging Algorithm**:

- Rule-Based POS Tagging: This approach uses handcrafted rules and patterns to assign POS tags to words based on their context within a sentence. These rules might consider word suffixes, prefixes,

capitalization, and surrounding words. It's relatively simple but may not handle all cases accurately.

- Probabilistic Models:
    - Hidden Markov Models (HMMs): HMMs are widely used for POS tagging. In an HMM, the states represent POS tags, and the observations correspond to words. The transition probabilities model the sequence of POS tags, and the emission probabilities model the likelihood of a word given a POS tag.
    - Maximum Entropy Markov Models (MEMMs): MEMMs are an extension of HMMs that allow for richer feature representations and handle state transitions and emissions independently, potentially leading to better performance.

4) **Sample output**:

Example: The cat is sitting on mat.

- 'DT': Determiner (e.g., 'The', 'the')
- 'NN': Noun (e.g., 'cat', 'mat')
- 'VBZ': Verb, 3rd person singular present (e.g., 'is')
- 'VBG': Verb, gerund or present participle (e.g., 'sitting')
- 'IN': Preposition or subordinating conjunction (e.g., 'on')
- '.': Punctuation mark (e.g., '.')