NAME ENTITY RECOGNITION FOR HINDI LANGUAGE

Assignment (2 & 3) for Natural Language Processing (20CSF-392)

Submitted by

MATCH PAVAN KUMAR (20BCS6072)
PRIYANGSHU SARKAR (20BCS6047)

In partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE ENGINEERING (AIML)

Submitted to:
Mr. Siddharth Kumar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
APEX INSTITUE OF TECHNOLOGY
CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413, PUNJAB

February-June 2023

What is NER (Named Entity Recognition): -

It is Identification and classification of named entities in text are tasks involved in named entity recognition (NER), a branch of natural language processing (NLP). Objects or entities with specific names, such as persons, locations, organizations, and other kinds of objects, are known as named entities. These named entities are to be extracted and categorized from unstructured text data using NER.

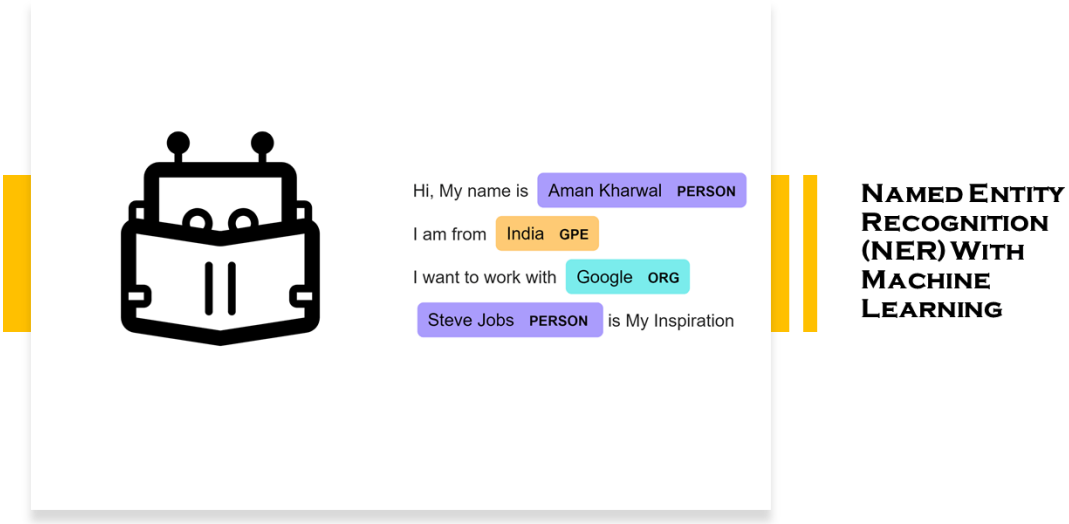


Fig 1: - It is detecting the tag of each word

The process of NER can be divided into several steps:

Preprocessing: This step involves cleaning and tokenizing the text data. Text cleaning involves removing unnecessary characters, such as punctuation marks and special symbols, from the text data. Tokenization involves breaking down the text into individual words or phrases, which are called tokens.

Part-of-Speech (POS) Tagging: This step involves assigning a part-of-speech tag to each token in the text. The part-of-speech tags indicate the grammatical role of each token in the sentence. For example, a token might be tagged as a noun, verb, adjective, or adverb.

Named Entity Recognition: This step involves identifying and classifying the named entities in the text. This is typically done using machine learning algorithms, such as Conditional Random Fields (CRF) or Support Vector Machines (SVM). The machine learning model is trained on a labeled dataset,

where each token in the text is labeled with the appropriate named entity tag (e.g., PERSON, LOCATION, ORGANIZATION, etc.). The model then uses this labeled data to learn patterns in the text that are associated with each named entity tag.

Post-processing: This step involves cleaning up and refining the output of the named entity recognition step. For example, the output might be cleaned up to remove false positives (i.e., tokens that were incorrectly labeled as named entities) or to merge adjacent named entities that refer to the same object.

NER is an important task in NLP because it is used in a wide range of applications, including information extraction, question answering, machine translation, and more. It is particularly important in the field of natural language understanding, where the goal is to enable machines to understand human language in a way that is similar to how humans understand it.

Named Entity Recognition (NER) has many practical applications in various domains:

Information Extraction: NER is used to extract specific pieces of information from text, such as names, dates, locations, and other entities. For example, NER can be used to extract the names of people and organizations from news articles or social media posts.

Question Answering: NER is used to extract relevant information from text to answer questions. For example, if a user asks a question like "When was Albert Einstein born?" NER can be used to extract the named entity "Albert Einstein" and then search for information related to his birthdate.

Sentiment Analysis: NER can be used to identify the entities that are mentioned in text and then classify the sentiment associated with those entities. For example, if a user tweets about a particular product, NER can be used to extract the product name and then determine whether the sentiment expressed in the tweet is positive, negative, or neutral.

Machine Translation: NER can be used to identify named entities in the source language and then translate them accurately into the target language. This can be particularly useful in translating technical documents or legal texts, where the accurate translation of named entities is critical.

Recommendation Systems: NER can be used to extract relevant information from user-generated content, such as reviews or social media posts, and then use that information to make personalized recommendations. For example, if a user mentions a specific restaurant in their review, NER can be used to extract the restaurant name and then recommend other restaurants that are similar.

Search Engines: NER can be used to improve the accuracy of search results by identifying and categorizing named entities in web pages. For example, if a user searches for "best pizza in New York", NER can be used to extract the named entities "pizza" and "New York" and then return search results that are relevant to those entities.

About Data Cleaning and Data Preprocessing:

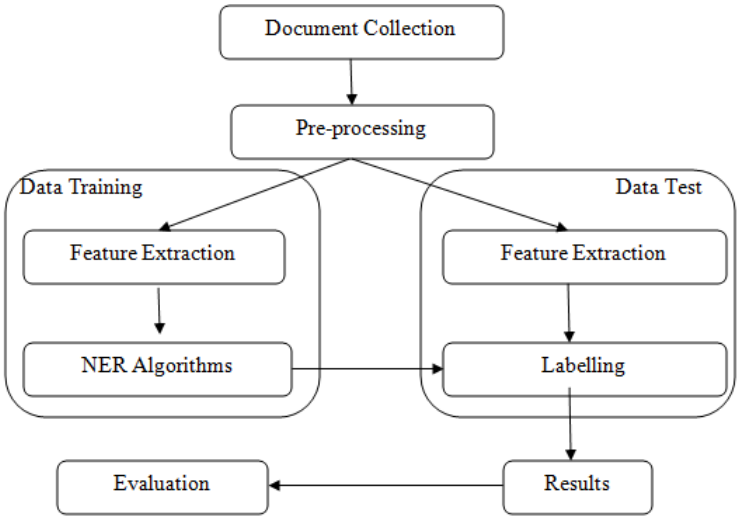


Fig 2: - Flowchart of Data cleaning and data Pre-processing

Data cleaning and preprocessing are important steps in any NER task, including for Hindi language. These steps involve transforming raw text data into a structured format that can be used for training and testing NER models. Here are some common data cleaning and preprocessing steps for NER in Hindi:

Tokenization: This involves breaking the raw text data into individual words or tokens. In Hindi, tokenization can be challenging because words are not always separated by spaces. Therefore, special attention needs to be given to ensure proper tokenization.

Stop-word removal: Stop words are common words in a language such as "the", "a", "an", etc. These words do not carry much meaning and can be removed to reduce noise in the data.

Lemmatization: This involves reducing inflected words to their base or root form, which can help reduce the number of unique words in the data and improve the accuracy of the NER model

Named entity tagging: This involves labeling words in the data that correspond to named entities such as people's names, place names, organization names, etc. This can be a manual or automated process, depending on the availability of labeled data.

Data augmentation: This involves adding synthetic or artificially generated data to the original dataset to increase its size and diversity. This can help improve the performance of the NER model.

Data normalization: This involves converting the data into a standard format, such as lowercasing all the text and removing any special.

Experimental Analysis:

About Dataset: - The data i have took from the mygov portal which is require to take the Hindi language dataset. And it is making the data as a form of .txt format. And then we have to perform a NER in that dataset.

Methodology: -

Initialize a variable "max_num" to store the maximum number found so far. Set it to the first number in the list.

Loop through the remaining numbers in the list.

For each number, compare it to the current maximum number "max_num". If it is greater than "max_num", update the value of "max_num" to be equal to the new number.

After all the numbers have been compared, the variable "max_num" will contain the maximum number in the list.

Return "max_num".

Here are the detailed steps:

Input: List of integers

Output: Maximum number in the list

Initialize a variable "max_num" to store the maximum number found so far. Set it to the first number in the list.

max_num = lst[0]

Loop through the remaining numbers in the list.

for num in lst[1:]:

For each number, compare it to the current maximum number "max_num". If it is greater than "max_num", update the value of "max_num" to be equal to the new number.

if num > max_num:

max_num = num

After all the numbers have been compared, the variable "max_num" will contain the maximum number in the list.

Return "max_num".

Implementation with code: -

```
!pip install -U spacy

!python -m spacy download en_core_web_sm

import spacy

nlp = spacy.load("en_core_web_sm")

nlp.pipe_names

doc=nlp("""माननीय प्रधानमंत्री जी महोदय...

निवेदन है की मध्यप्रदेश में नगरीय प्रशासन एवं विकास विभाग के अंतर्गत शहरी आजीविका मिशन में हम लगभग 400 कम्युनिटी ऑर्गेनाइजर की भर्ती एक परीक्षा के द्वारा की गई जबकि हमे संविदा पर 12000 मासिक मानदेय पर रखा गया है

प्रधानमंत्री स्वनिधि योजना एवं शहरी गरीबों हेतु अन्य योजनाओं के बेहतर क्रियान्वयन हेतु इतने कम वेतनमान एवं विना नियमितीकरण के कार्य करना हमारा शोषण है

अतः श्रीमान जी से निवेदन है की हमे नियमित पद प्रदान कर एक सम्मानजनक वेतनमान प्रदान कराने की कृपा करें.

मन की बात कार्यक्रम की 100वीं कड़ी के लिए आपको हार्दिक बधाई।

बेहद व्यस्त रहते हुए भी हर महीने के आखिरी रविवार को आपको उत्सुकतापूर्वक भारत के नागरिकों से जुड़ने, उनके सुझावों को सुनने तथा अपने प्रेरणादायक एवम् नव विचारों
```

के साथ उनका मार्गदर्शन करने तथा राष्ट्र को निरंतर प्रगति एवम् उन्नति के पथ पर ले जाने के लिए आपके समर्पण को सहृदय नमन करता करता हूं ।

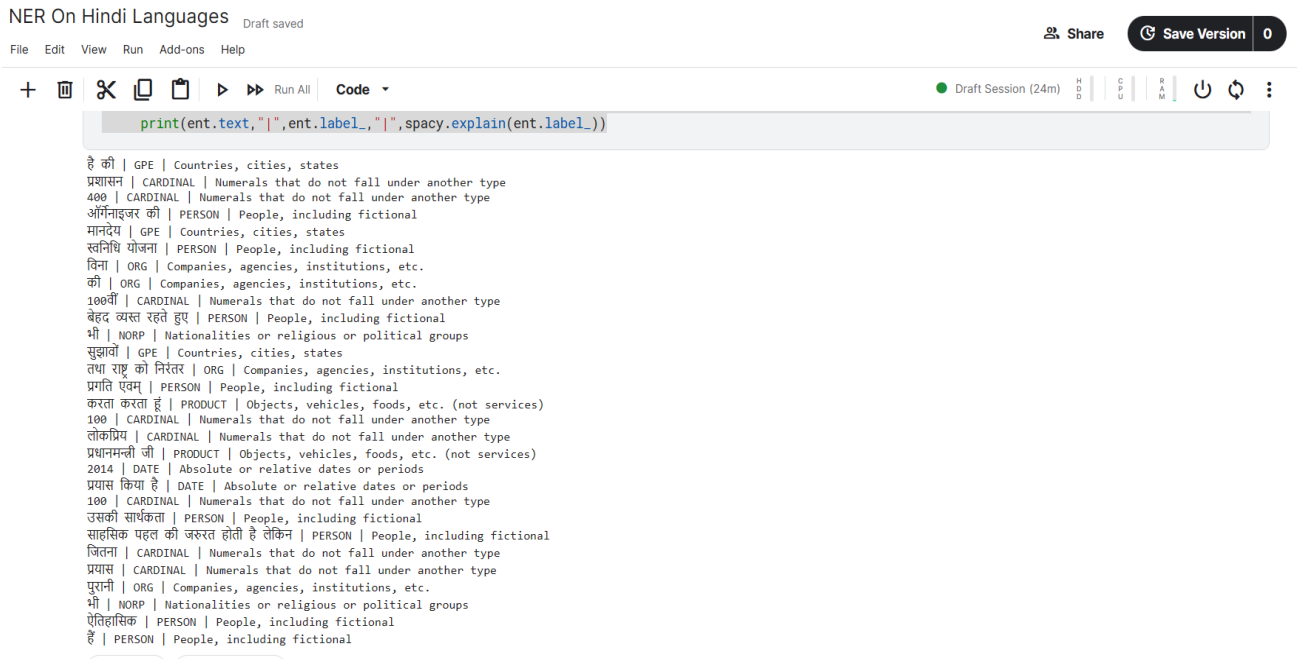
निःसंदेह मन की बात के 100 कड़ियों का ये सफर समाज के प्रत्येक वर्ग की भागीदारी वाला एक प्रेरणादायक एवम् नवविचारों वाला लोकप्रिय कार्यक्रम है ।

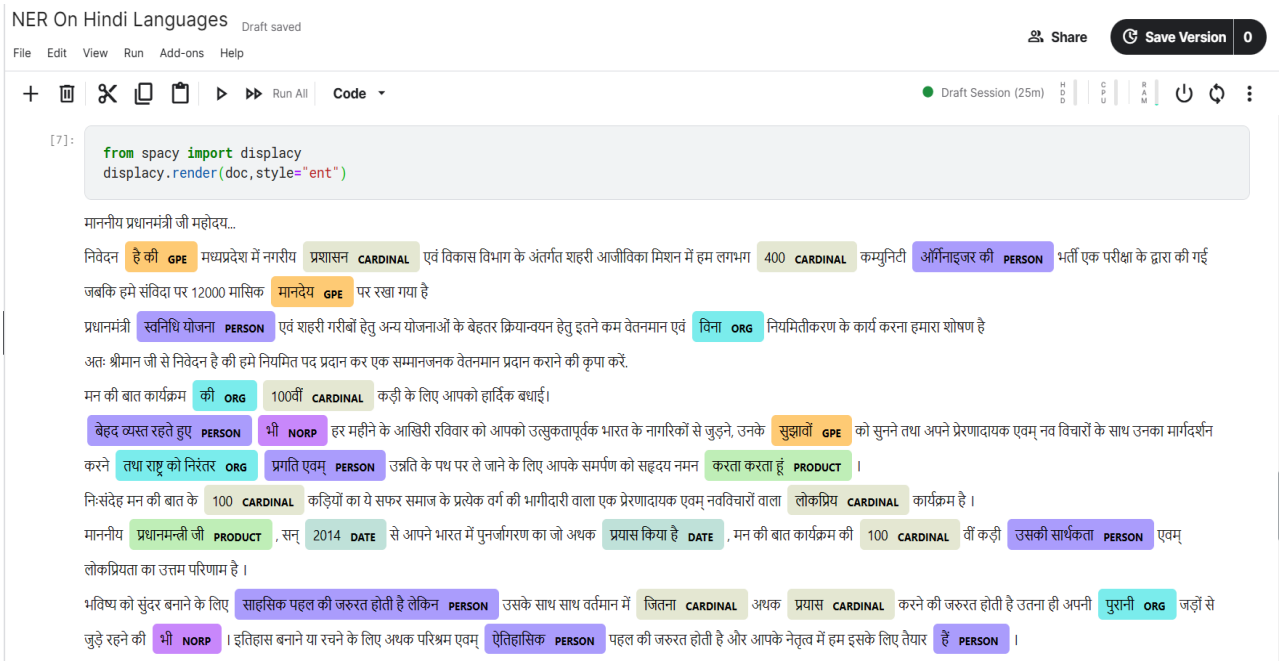
माननीय प्रधानमन्त्री जी, सन् 2014 से आपने भारत में पुनर्जागरण का जो अथक प्रयास किया है, मन की बात कार्यक्रम की 100 वीं कड़ी उसकी सार्थकता एवम् लोकप्रियता का उत्तम परिणाम है ।

भविष्य को सुंदर बनाने के लिए साहसिक पहल की जरुरत होती है लेकिन उसके साथ साथ वर्तमान में जितना अथक प्रयास करने की जरुरत होती है उतना ही अपनी पुरानी जड़ों से जुड़े रहने की भी । इतिहास बनाने या रचने के लिए अथक परिश्रम एवम् ऐतिहासिक पहल की जरुरत होती है और आपके नेतृत्व में हम इसके लिए तैयार हैं।""")

```
for ent in doc.ents:
    print(ent.text,"|",ent.label_,"|",spacy.explain(ent.label_))
from spacy import displacy
displacy.render(doc,style="ent")
```

Results/Outputs:




NER On Hindi Languages
Draft saved
File Edit View Run Add-ons Help
Share
Save Version 0
Run All
Code
Draft Session (25m)
[7]:
from spacy import displacy
displacy.render(doc,style="ent")
माननीय प्रधानमंत्री जी महोदय...
निवेदन है की GPE मध्यप्रदेश में नगरीय प्रशासन CARDINAL एवं विकास विभाग के अंतर्गत शहरी आजीविका मिशन में हम लगभग 400 CARDINAL कम्युनिटी ऑर्गेनाइजर की PERSON भर्ती एक परीक्षा के द्वारा की गई जबकि हमे संविदा पर 12000 मासिक मानदेय GPE पर रखा गया है
प्रधानमंत्री स्वनिधि योजना PERSON एवं शहरी गरीबों हेतु अन्य योजनाओं के बेहतर क्रियान्वयन हेतु इतने कम वेतनमान एवं बिना ORG नियमितीकरण के कार्य करना हमारा शोषण है
अतः श्रीमान जी से निवेदन है की हमे नियमित पद प्रदान कर एक सम्मानजनक वेतनमान प्रदान कराने की कृपा करें.
मन की बात कार्यक्रम की ORG 100वीं CARDINAL कड़ी के लिए आपको हार्दिक बधाई।
बेहद व्यस्त रहते हुए PERSON भी NORP हर महीने के आखिरी रविवार को आपको उत्सुकतापूर्वक भारत के नागरिकों से जुड़ने, उनके सुझावों GPE को सुनने तथा अपने प्रेरणादायक एवम् नव विचारों के साथ उनका मार्गदर्शन करने तथा राष्ट्र को निरंतर ORG प्रगति एवम् PERSON उन्नति के पथ पर ले जाने के लिए आपके समर्पण को सहृदय नमन करता करता हूं PRODUCT ।
निःसंदेह मन की बात के 100 CARDINAL कड़ियों का ये सफर समाज के प्रत्येक वर्ग की भागीदारी वाला एक प्रेरणादायक एवम् नवविचारों वाला लोकप्रिय CARDINAL कार्यक्रम है ।
माननीय प्रधानमन्त्री जी PRODUCT , सन् 2014 DATE से आपने भारत में पुनर्जागरण का जो अथक प्रयास किया है DATE , मन की बात कार्यक्रम की 100 CARDINAL वीं कड़ी उसकी सार्थकता PERSON एवम् लोकप्रियता का उत्तम परिणाम है ।
भविष्य को सुंदर बनाने के लिए साहसिक पहल की जरुरत होती है लेकिन PERSON उसके साथ साथ वर्तमान में जितना CARDINAL अथक प्रयास CARDINAL करने की जरुरत होती है उतना ही अपनी पुरानी ORG जड़ों से जुड़े रहने की भी NORP । इतिहास बनाने या रचने के लिए अथक परिश्रम एवम् ऐतिहासिक PERSON पहल की जरुरत होती है और आपके नेतृत्व में हम इसके लिए तैयार हैं PERSON ।