Analysis

Hidden Markov Models (HMMs) are statistical models used in natural language processing for tasks like part-of-speech tagging. They operate on the principle of transition probabilities between states. In the context of POS tagging, each word in a sentence is treated as a sequence of states, where each state represents a possible part-of-speech tag. The HMM calculates the probability of transitioning from one state (POS tag) to another based on the observed word sequence. By analyzing a large corpus of text, the HMM learns these transition probabilities, allowing it to predict the most likely sequence of POS tags for a given sentence.

Conditional Random Fields (CRFs) are probabilistic models used for sequence labeling tasks, including POS tagging. Unlike HMMs, which model the transition probabilities between states, CRFs directly model the conditional probability of a label sequence given an input sequence. In POS tagging, the input sequence is the sequence of words in a sentence, and the label sequence is the sequence of POS tags assigned to each word. CRFs consider various features of the input sequence, such as word identities, prefixes, suffixes, and context words, to calculate the likelihood of different label sequences. They use this information to determine the most probable sequence of POS tags for the input sentence.

HMM

English

Precision: 0.591713452574835 Recall: 0.43890675241157556 F1 Score: 0.3811332012839551

CC

CD

DM

DT

JJ

NEG

Confusion matrix:

0.0

PR

0.0

0.0

8.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

1.0

0.0

0.0

0.0

0.0

0.0

0.0

0.0

46.0

0.0

PUNC PU QΤ QTC QT0 RB **RBR** RP T0 VAUX **VBZ** CC 11.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 2.0 CD 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 8.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 DM 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 6.0 0.0 0.0 0.0 $\Theta \cdot \Theta$ 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 4.0 DT 43.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 36.0 0.0 JJ 0.0 18.0 0.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 3.0 0.0 0.0 0.0 0.0 0.0 31.0 0.0 NEG 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 3.0 0.0 0.0 0.0 0.0 $\Theta . \Theta$ NN 0.0 0.0 0.0 0.0 1.0 0.0 46.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 78.0 NNP 0.0 0.0 0.0 0.0 0.0 0.0 11.0 0.0 0.0 0.0

NNP

PR

PRP

PRQ

NN

PSP

Time taken in seconds: 8.01795506477356\

Hindi

Precision: 0.40750128008192527 Recall: 0.24423963133640553 F1 Score: 0.2511683401084323

Confusion matrix:

```
CC
           CCD CCS
                      INTF
                              JJ
                                  NEG
                                        NN
                                             NNP
                                                  PRL
                                                       PRP
                                                             PR0
                                                                   PSP
                                                                         PUNC
    OTC QTO
                     RPD
                                           VAUX
                                                      VNF
OT
                 RB
                          SYM
                                UNK
                                      UT
                                                  VM
                                                             WQ
                                                       0.0
CC
      3.0 0.0
                 0.0
                       0.0
                            0.0
                                  0.0
                                       0.0
                                             0.0
                                                  0.0
                                                             0.0
                                                                   0.0
                                                                          0.0
     0.0
          0.0
                                                             0.0
0.0
                 0.0
                     0.0
                           0.0
                                0.0
                                      0.0
                                             0.0
                                                  0.0
                                                       0.0
                                       0.0
CCD
      1.0 0.0
                0.0
                            0.0
                                  0.0
                                             0.0
                                                  0.0
                                                       0.0
                                                             0.0
                       0.0
                                                                   0.0
                                                                          0.0
     0.0
          0.0
                 0.0
                                             0.0
0.0
                      0.0
                           0.0
                                0.0
                                      0.0
                                                  0.0
                                                        0.0
                                                             0.0
CCS
      3.0 0.0
                 0.0
                       0.0
                            0.0
                                  0.0
                                       0.0
                                             0.0
                                                  0.0
                                                       0.0
                                                             0.0
                                                                   0.0
                                                                          0.0
```

0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 INTF 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 JJ 0.0 0.0 0.0 8.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 3.0 0.0 0.0 4.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 NEG 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.00.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 NN 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.05.0 0.0 0.0 26.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 **NNP** 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 14.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 **PRL** 0.0 **PRP** 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 **PRQ** 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 **PSP** 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 86.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 **PUNC** 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 QΤ 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 8.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 QTC 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 4.0 0.0 6.0 0.0 0.0 0.0 0.0 0.0 QT0 0.0 1.0 0.0 RB 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 **RPD** 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 SYM 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 UNK 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 UT 0.0 VAUX 0.0 VM 0.0 0.0 0.0 0.0 3.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 6.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 **VNF** 0.0 WQ 0.0

Time taken in seconds: 22.796072483062744

CRF

English

	СС	CD	DM		DT		JJ	N	EG		NN	NNP	F	PR	PF	RP	PR0	PS
PUNC	QT		QT0) VAU						-	. 0
CC	11.0	0.0	0.0			0						0.0						Θ.
0.0	0.0	0.0	0.0	1.	. 0	0.0	0	. 0	Θ.	0	0.0				. 0			
CD	0.0	0.0	0.0		0.0	0	. 0	0	. 0		2.0	1.0	0	. 0	0	. 0	0.0	0.
0.0	6.0	0.0	0.0		. 0						0.0	0.0		0	. 0	Θ.	0	
MC	0.0	0.0	0.0		6.0						1.0	0.0	6	. 0	0	. 0	0.0	0.
3.0	0.0	0.0	0.0	Θ.	. 0	0.0	0	. 0	Θ.	0	0.0	0.0		0	. 0	Θ.	0	
DT	0.0	0.0									0.0	1.0						Θ.
8.0	0.0	0.0									0.0							
JJ	2.0	0.0	0.0								12.0							Θ.
0.0	0.0	0.0									0.0							
NEG	0.0										0.0							Θ.
0.0	0.0	0.0									0.0				. 0			
NN	5.0	0.0	0.0								.02.0	1.0						1.
1.0	1.0	0.0									0.0							
	3.0				0.0						16.0							Θ.
1.0	10.0	0.0									0.0							
PR	0.0										1.0							Θ.
0.0	0.0	0.0	0.0								0.0							
PRP	0.0											1.0						0.
9.0	0.0	0.0	0.0								0.0							٠.
PRQ	0.0																0.0	0.
0.0	0.0	0.0									0.0							٠.
PSP	0.0	0.0									2.0							57.
0.0	0.0	0.0									0.0							0
PUNC	0.0	0.0	0.0								0.0							0.
52.0	0.0	0.0	0.0								0.0							٠.
QT	0.0										0.0							0.
0.0	0.0										0.0							0.
QTC		0.0										1.0						Θ.
1.0											0.0							٠.
QT0											0.0							Θ.
0.0		0.0									0.0							٥.
RB		0.0									0.0							Θ.
0.0	0.0	0.0									0.0							٠.
RBR		0.0										0.0						Θ.
0.0	0.0	0.0									0.0				. 0			0.
RP											0.0							1.
0.0	0.0										0.0							Τ.
TO		0.0										0.0						0.
0.0	0.0										0.0							0.
VAUX		0.0										0.0						1.
0.0	0.0	0.0										0.0			. 0			Τ.
VBZ		0.0										0.0						0.
0.0	0.0	0.0									0.0							0.
																		Θ.
VM	⊥.⊍	0.0	0.0		0.0	2	. 0	0	. 0		9.0	1.0	U	. U	U	. U	0.0	U

Hindi

Precision: 0.6881771240205123 Recall: 0.7027649769585254 F1 Score: 0.6863571656255918

Confusion matrix:

CC CCD CCS INTF JJ NEG NN NNP PRL PRP PR₀ **PSP** QTC RPD SYM UNK UT **PUNC** QΤ QTO RB VAUX VM VNF WO 0.0 CC 2.0 3.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.00.0 0.0 0.0 0.0 0.0 0.00.0 0.0 0.0 0.0 0.0 0.0 CCD 0.0 1.0 0.0 CCS 1.0 3.0 0.00.00.0 0.0 0.0 0.0 0.0 0.0 0.00.0 0.00.00.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 JJ 0.0 0.0 0.0 9.0 0.0 4.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 $\Theta \cdot \Theta$ NEG 0.0 0.0 4.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0 0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 $\Theta \cdot \Theta$ NN 0.0 0.0 0.0 0.0 0.0 87.0 5.0 0.0 0.0 0.0 4.0 1.0 0.0 0.0 0.0 0.0 3.0 $\Theta \cdot \Theta$ 0.0 1.0 0.0 0.0 0.0 12.0 0.0 2.0 0.0 8.0 0.0 0.0 0.0 0.0 NNP 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 3.0 **PRL** 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 $\Theta \cdot \Theta$ **PRP** 0.0 0.0 0.0 0.0 0.0 1.0 1.0 2.0 11.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.01.0 0.0 0.0 PR₀ 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 **PSP** $\Theta \cdot \Theta$ 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 86.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 **PUNC** 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 23.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0QΤ 0.0 0.0 0.0 0.0 0.00.0 0.0 0.0 0.0 6.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 QTC 0.0 0.0 0.0 0.0 0.0 0.0 4.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 4.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 QT0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 $\Theta \cdot \Theta$ RB 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 **RPD** 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 1.0 0.0 $\Theta \cdot \Theta$ $\Theta \cdot \Theta$ 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 SYM 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

13.0 0.0 0.0 0.0 0.0 0.0 $0.0 \quad 0.0$ 0.0 0.00.0 0.0 0.0 UNK 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 UT 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 **VAUX** 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.00.00.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 30.0 10.0 0.0 VM 0.0 0.0 0.0 4.0 0.0 4.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 3.0 30.0 0.0 VNF 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 2.0 0.0 0.0 0.0 0.0 WQ 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

Observations

- 1. The CRF model performs better than the HMM model in both languages.
- 2. The CRF model performs better in Hindi than in English.
- 3. The HMM model performs better in English than in Hindi.
- (1) is because the CRF model is able to take into account the context of the words and the features of the words, which the HMM model is not able to do. The CRF model is able to learn the features of the words and the context in which they appear, which helps it in making better predictions. The HMM model, on the other hand, is not able to learn the features of the words and the context in which they appear, which makes it less accurate in making predictions.