Computational Linguistics - 1

An Inquiry into the Accuracy of Different Part-of-Speech Tagging Models on Speech Data

Submitted By: Ishaan Romil (2023114011)
Submitted To: Parameswari Krishnamurthy
Date of Submission: 7th May, 2024

Acknowledgement

I express my sincere gratitude to Parameswari ma'am for her invaluable teachings and insights throughout the past semester, which have been instrumental in shaping this report.

Additionally, I extend my thanks to Sankalp sir for his unwavering assistance in navigating doubts and difficulties, contributing significantly to my enhanced understanding of computational linguistics in general, and POS tagging in particular.

Special appreciation goes to my friends and family for their steadfast support.

Index

Introduction

Objectives

The objectives of this report are as follows:

1. To investigate how the quality of speech input (average, good, poor) impacts the performance of POS tagging models (HMM, CRF, LSTM).

2. To examine the effectiveness of POS tagging models (HMM, CRF, LSTM) in handling different languages (French, English, Hindi) and potential language-specific challenges.

3. To compare the strengths and weaknesses of HMM, CRF, and LSTM models in POS tagging tasks across various speech qualities and languages.

4. To identify the reasons behind the potential differences in performance, if any, between the three models for the various qualities and languages.

5. To identify potential strategies for improving POS tagging performance in challenging scenarios, such as low-quality speech or non-native languages.

Hypothesis

1. The performance of POS tagging models will be highest for low-quality speech inputs, followed by average-quality speech, and lowest for high-quality speech.

2. The effectiveness of POS tagging models will differ between languages, with models performing better on English transcripts compared to French transcripts, and will perform the worst on Hindi due to potential language-specific challenges.

3. Among the three models (HMM, CRF, LSTM), LSTM will exhibit the highest overall accuracy in POS tagging tasks, followed by CRF and then HMM.

Null Hypothesis

1. There is no significant difference in the performance of POS tagging models between average, good, and poor speech qualities.

2. There is no significant difference in the effectiveness of POS tagging models between English, French, and Hindi languages.

3. There is no significant difference in the accuracy of POS tagging models between HMM, CRF, and LSTM.

Preliminary Descriptions

Description of POS Tagging Models

Hidden Markov Models

Hidden Markov Model (HMM) is a statistical model used for modeling sequences of observable events with underlying unobservable states. In the context of part-of-speech (POS) tagging, HMMs are used to model the sequence of words in a sentence as observable events, with the POS tags as the unobservable states.

HMM has several key characteristics. The first of which assumes the Markov property, which states that the probability of transitioning to the next state depends only on the current state. Essentially, the POS tag of the next term is only dependent on the POS tag of the current term. This property simplifies the modeling of sequential data by reducing the number of parameters to estimate.

Another key characteristic of HMM is that it uses sequential dependency modeling. The model captures sequential dependencies between POS tags, allowing it to leverage context information from neighboring words in the sequence. This makes it more efficient in situations where the order of the words in a sentence matters.

Lastly, it makes use of emission and transition Probabilities: HMMs represent the model parameters using “horizontal” emission probabilities (probability of observing a word given a POS tag) and “vertical” transition probabilities (probability of transitioning between POS tags). To make use of these probabilities, HMM uses several different algorithms – the most popular of these is the Viterbi algorithm. The Viterbi Algorithm removes edges of paths with lower probability cost and efficiently finds the most likely sequence of hidden states (POS tags) given the observed sequence (sentence).

Conditional Random Fields

Conditional Random Fields (CRF) is a type of probabilistic graphical model used for labeling sequential data.

CRF also has several key characteristics. Firstly, it is a discriminative model that directly models the conditional probability of the label sequence given the input sequence, instead of using joint probability like HMM.

Secondly, instead of sequential dependency modeling, it uses global context modeling. When labeling, CRF takes into account the full input sequence, which enables it to capture intricate relationships and interactions between adjacent labels. When compared to models that just take local context into account such as HMM, models that take into account global context – such as CRF – frequently perform better.

It also has flexible feature representations, which encompasses a broad spectrum of linguistic and contextual features. These characteristics increase the accuracy of the model by capturing pertinent data for the labeling task.

Lastly, CRF models work most efficiently for more complex sentences and datasets due to their high degree of optimization.

Long Short-Term Memory

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture which can be applied to predict POS tags.

One of the key characteristics of LSTM includes its use of the memory cell. LSTM networks incorporate a memory cell, allowing them to maintain information over long sequences. This memory cell can retain information for extended periods, making LSTMs effective for tasks involving long-range dependencies.

They also utilize three types of gates – input gate, forget gate, and output gate – to control the flow of information within the network. These gates regulate the information that is stored, updated, and output by the memory cell, enabling better control over the learning process. The forget gate determines which information to discard from the cell state. It utilizes a sigmoid activation function to output values between 0 and 1, indicating how much of the previous cell state should be retained. The input gate regulates the update of the cell state by selectively incorporating new information. Lastly, the output gate controls the information that is output from the memory cell. It uses a sigmoid activation function to decide which parts of the cell state are revealed as the output, and a tanh activation function to scale the output values.

CRF also makes use of the Backpropagation Through Time (BPTT) algorithm for training, which enables them to learn from sequences of data by backpropagating errors through time. This allows LSTMs to capture temporal dependencies and make predictions based on sequential input data.

Description of Languages

French

French is the 5th most spoken language in the world with around 321 million speakers across five continents. The majority of its speakers are located in Europe, North America, and Africa. French serves as an official language in 28 countries. It serves as one of the 6 official languages and 2 working languages, as recognized by the United Nations.

French originates in the nation of France and is a descendant of Latin – being one of the "Romance languages," alongside languages such as Spanish and Italian. It spread across the world due to the French colonial empire, which had a strong foothold in West and Central Africa, along with dominions in North America, namely Quebec, Haiti, and Louisiana.

English

English is the most-spoken language in the world, with around 18% of the world's population being able to speak the language to some degree. Additionally, around 360 million people are native speakers of English, making it the third-most spoken language in the world by number of native speakers. It remained the primary language of communication in the "Anglosphere", a group of countries including the United States of America, the United Kingdom, Australia, New Zealand, and Canada. Similar to French, it serves as one of the 6 official languages and 2 working languages, as recognized by the United Nations.

English originates from England in Europe. While being a Germanic language, it has strong French influence, with many English words tracing their origin back to Latin and Greek. Similar to French, it spread across the world due to colonization of the Americas, Africa, India, and Oceania. It continues to hold a dominant role in the world due to its status as the official language of the world's preeminent superpower, the United States of America.

Hindi

Hindi is the fourth-most spoken language in the world, ranked by native speakers, with around 345 million people speaking the language natively. It is the most commonly-spoken language in India and is also the fastest-growing language in India.

Hindi originates from the Indo-Gangetic in northern India, where it is still widely-spoken. Hindi can trace its roots back to the ancient language of Sanskrit, an Indo-European language. It is mostly spoken in India and the Indian subcontinent. However, due to colonization and the emigration of Indians to countries in the Caribbean such as Suriname and Trinidad and Tobago, it has become part of the linguistic landscape of parts of the region. It is also spoken in large numbers by the Indian diaspora – the largest diaspora of any nationality in the world.

Description of Training Dataset

UD_French-FQB

The corpus UD_French-FQB is an automatic conversion of the [French QuestionBank v1] (http://alpage.inria.fr/Treebanks/FQB/), a corpus entirely made of questions. The data consists of 2,289 sentences, 23,236, and an average sentence length of 10.15 words per sentence. In terms of token count, it consists of 24,452 tokens.

This dataset is tagged using the Universal Dependencies (UD) tagset. It was used for all 3 part-of-speech models and is the only dataset used to train the models on the French language.

UD_English-PUD

The corpus UD_English-PUD is a part of the Parallel Universal Dependencies (PUD) treebanks created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies (http://universaldependencies.org/conll17/). The sentences in the dataset are taken from Wikipedia and from news articles. It consists of 1,000 sentences in line with the specifications of the shared task. In terms of token count, it consists of 21,718 tokens.

This dataset is tagged using the Universal Dependencies (UD) tagset. It was used for all 3 part-of-speech models and is the only dataset used to train the models on the English language.

Hindi Tourism

This dataset was provided from the Language Technologies Research Center, IIIT-Hyderabad. It consists of 4,516 sentences. In terms of token count, it consists of 86,378 tokens.

The dataset is tagged using the Bureau of Indian Standards (BIS) POS tagset and was used for the HMM and CRF POS models for the Hindi language.

UD_Hindi-PUD

Similar to UD_English-PUD, The corpus UD_Hindi-PUD is a part of the Parallel Universal Dependencies (PUD) treebanks created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies (http://universaldependencies.org/conll17/). The dataset consists of 1,000 sentences and 23,806 tokens.

The dataset is tagged using the Universal Dependencies (UD) tagset. It was used for the LSTM model for the Hindi language as the Hindi tourism dataset was not in a compatible format.

Methodology

Introduction to Variables

In order to test the thesis, three variables were tested and compared. The first variable was the language – which varied between English, French, and Hindi. The second variable was the quality of speech – which varied between "high-proficiency" speech, "intermediate-proficiency" speech, and "low-proficiency" speech. The third variable was the POS model. The POS tagging models used were Hidden Markov Models (HMM), Conditional Random Fields (CRF), and Long Short-Term Memory (LSTM).

It is important to note that the variation in the quality of speech refers not to variations in the quality of speech, but rather variations in the lexical and grammatical complexity.

The analysis section of this report compares the accuracy metrics based on these variables, ceteris paribus.

Analysis Metrics

Three analysis metrics were used to analyze the data – precision, recall, and F1 score.

Precision refers to the accuracy of positive predictions made by a model. It measures the proportion of true positive results among all instances predicted as positive. In other words, precision gauges the model's ability to avoid false positives, indicating how reliable its positive predictions are.

Recall refers to the model's ability to capture all relevant instances of a certain class. It measures the proportion of true positive results among all actual positive instances. A high recall score indicates that the model effectively identifies most of the positive cases within the dataset.

The F1 score is a metric that combines both precision and recall into a single value, providing a comprehensive assessment of a model's performance. It calculates the harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives. A higher F1 score indicates better overall performance, striking a balance between precision and recall.

Interview Methodology

Three interviews were conducted for each of the three selected languages. The interviewees were selected based on their proficiency of the language so as to have a diverse range of speech content to vary the tonality, mood and lexical strength of the data.

All interviews consisted of the same five questions and respondents were told to answer them in around one hundred words, with some respondents far surpassing this requirement whilst others struggled to meet it.

The interview responses to each question were separately recorded and stored for analysis after a conversion by a speech-to-text software.

English

Proficiency of speaker	Time allotted for preparation
Low	1 Minute
Intermediate	3 Minutes
High	10 Minutes

Table 2.1: Interview Preparation Times for English

Due to difficulties in finding non-native speakers of English, the interview methodology was slightly altered for English. To differentiate the proficiencies of speakers, they were given different amounts of preparation time in which they were allowed to jot down any notes, access and search the internet for more advanced vocabulary, and also to think more critically about their responses. The speakers were told that their objective was to speak as fluently as possible – avoiding stutters and ambiguities – whilst attempting to meet the target word count requirements.

The preparation time for the different proficiencies of speakers are attached in Table 2.1.

French and Hindi

Proficiency of speaker	Time allotted for preparation
Low	15 Minutes
Intermediate	15 Minutes
High	15 Minutes

Table 2.2: Interview Preparation Times for French and Hindi

Due to the availability of both native and non-native speakers for both the French language and the Hindi language, a standardized methodology could be followed. Each speaker was allotted 15 minutes to think of their responses after receiving the questions. During this time, they were allowed to access the internet – including translation applications like Google Translate and WordReference – alongside writing any points they might like to say. Unlike English, the speakers were not asked to intentionally obfuscate their answers as it was expected that, due to the varying levels of study and practice, there would be a natural variance in the quality of speech of all three speakers. They were, however, told to try to meet the word count requirements.

The main difference between the Hindi and French speakers was in the "intermediate-proficiency" speakers. For Hindi, the intermediate-proficiency speaker had experience conversationally speaking Hindi, but had not engaged in exploring Hindi literature outside of an academic context – unlike the high-proficiency speaker whose mother was a Sanskrit teacher and who occasionally read Hindi literature. Meanwhile, the French intermediate speaker had engaged in very little French speaking and was not entirely conversationally fluent in the language.

The preparation time for the different proficiencies of speakers are attached in Table 2.2.

Details of French Respondents

High-proficiency speaker

Gender: Female
Age: 18

The high-proficiency speaker for the French language was a native speaker who had lived in France for eight years, and was fluent in the language. Though her mother tongue was a Dravidian language, she grew up in an environment where French was the primary language of communication, being exposed to it through various channels of communication, enhancing her understanding of the morphological and sentence structure of the language.

Intermediate-proficiency speaker

Gender: Male
Age: 19

The intermediate-proficiency speaker for the French language was a non-native speaker who had studied French in school for a number of years as a second language. He had also prepared for the Alliance Française DELF-DALF certification exam, achieving B1 proficiency. Furthermore, he has a background in linguistics which may have aided him in appreciating the syntactical complexity of the language.

Low-proficiency speaker

Gender: Female
Age: 18

The low-proficiency speaker for the French language was a non-native speaker who had studied French in school as a second language throughout high school (9th-12th grade). She has not been tested by the Alliance Française DELF-DALF certification exam.

Details of English Respondents

High-proficiency speaker

Gender: Male
Age: 18

The high-proficiency speaker for the English language was a native speaker who had grown up speaking in English. His mother tongue, however, was a Dravidian language. However, as with the high-proficiency French speaker, he was conditioned from a young age to speak fluent English due to the prevalence of the language in his environment. He grew up in the Middle East, where English serves as a lingua franca and virtually the only means for Indians in the region to communicate with the rest of society. He was also an avid reader and debater, meaning his lexicon was more advanced than the average English speaker.

Intermediate-proficiency speaker

Gender: Female
Age: 18

The intermediate-proficiency speaker for the English language was a native speaker who had grown up speaking in English. Her mother tongue was Cantonese, a Sino-Tibetan language. Due to her childhood and adolescence being spent in the United States, she spoke with a distinct American accent throughout the interview. She engaged with English literature only through academic and scholarly contents and did not engage in many communication-related extracurricular activities.

Low-proficiency speaker

Gender: Female
Age: 18

The low-proficiency speaker for the English language was a native speaker who had grown up speaking in English. Her mother tongue was an Indo-European language – Marathi, which also has a notable influence from Sanskrit. The speaker was conditioned to speak English as she grew up in the United Kingdom. This also had an impact on her accent as she spoke with a distinct British accent throughout the interview. She admitted to not regularly engaging in the exploration of English literature or communication-oriented extracurricular activities such as speech or debate, potentially limited the lexical complexity of her speech

Details of Hindi Respondents

High-proficiency speaker

Gender: Male
Age: 18

The high-proficiency speaker for the Hindi language was a native speaker whose mother tongue was Hindi and who had grown up in a city in the Indo-Gangetic plain, meaning he had interacted with the language on a daily basis for his whole life, being exposed to it through almost every channel of communication such as entertainment, school, and social correspondences. His mother is also a Sanskrit teacher, meaning he had more exposure to Hindi and Sanskrit than most individuals, and was often directed by his mother to read Hindi literature.

Intermediate-proficiency speaker

Gender: Male
Age: 18

The intermediate-proficiency speaker for the Hindi language was also a native speaker whose mother tongue was Hindi and who had grown up in a city in the Indo-Gangetic plain. This means he was also exposed to Hindi from a young age through various channels of communication. However, unlike the high-proficiency speaker, he had no influences which directed him to further his knowledge of Hindi or Sanskrit, and admitted to rarely engaging in the exploration of Hindi literature, which could lead to a lower knowledge of advanced vocabulary and sentence structures.

Low-proficiency speaker

Gender: Male
Age: 18

The low-proficiency speaker for the Hindi language was a non-native speaker whose mother tongue was a Dravidian language and who lived in southern India – where Hindi is less prominent – and the United States – where Hindi is rarely spoken outside of households which belong to the north Indian diaspora – meaning his exposure to the Hindi language was extremely limited. As the speaker was more familiar with a Dravidian language, his concept of morphology and phonology may have been very different which would make concepts like agreement much harder. He studied Hindi in school for two years, making up the vast majority of his exposure to the language.

Speech-to-Text Conversion

A specialized speech-to-text conversion software, known as Speechnotes, was used for the conversion. The software was pre-trained for various languages, ensuring accuracy for the all three languages selected, alongside being pre-trained separately for all three accents of English which were used, ensuring the output was as clear and accurate as possible.

Observations

Qualitative Observations

Language	Proficiency	Observations
French	High	Her responses were characterized by a much faster tone with less interjections due to her fluency.
	Intermediate	His accent was less strong and the speech pattern consisted of more interjections, pauses, and grammatical and pronunciation mistakes.
	Low	The responses to the interview questions were much less fluent than both the intermediate-proficiency and high-proficiency speakers, consisting of a significant number of pauses, interjections, mispronunciations, grammatical mistakes, and incoherent sentences.
English	High	Spoke with extremely eloquent vocabulary, probably wrote down sentences during the preparation phase.
	Intermediate	The speaker had an American accent (as expected) and while she had a few pauses, she mostly spoke clearly going into detail about the questions.
	Low	The speaker had a British accent (as expected). She had few pauses but also gave less information maybe because she had less time to think about her points.
Hindi	High	There was a fair bit of content mostly delivered in Hindi with very little code switching, though examples of examples of code switching were still scene (ie. “kidnap”). Vocabulary was relatively stronger than what is used in everyday speech
	Intermediate	The response contained very few details, potentially because the speaker focused more on trying to speak in proper Hindi during the preparation time and to engage less in code switching.
	Low	While there were a few examples of misalignment and disagreement based on gender and plurality, the speech was comprehensible. The speech was a lot more conversational and there was little lexical complexity.

Model	French	English	Hindi
HMM	63.3%	78.4%	62.8%
CRF	60.2%	79.2%	61.5%
LSTM	67.2%	73.4%	67.2%

Table 3.1: Language analysis using precision score

LSTM was the most accurate model by-far for Hindi, significantly surpassing both HMM and CRF which both performed similarly. This trend was also seen in French, where LSTM once again outclassed the other two models. However, the difference in precision between HMM and CRF was greater in French than in Hindi. In English, however, HMM and CRF significantly outperformed LSTM, and performed similarly, though CRF was slightly more accurate than HMM.

Model	French	English	Hindi
HMM	59.2%	77.1%	58.5%
CRF	53.7%	75.7%	50.6%
LSTM	62.1%	67.5%	66%

Table 3.2: Language analysis using recall score

LSTM was once again the most accurate model for both French and Hindi, significantly outperforming CRF and edging out HMM by a few percentage points. However, once again, the reverse trend was seen in English, where LSTM was outclassed by both HMM and CRF by a wide margin. However, the recall score was greater for HMM than CRF while the precision score, as seen earlier, was greater for CRF than HMM. The recall scores of CRF were significantly lower than the precision scores

Model	French	English	Hindi
HMM	57.5%	78.7%	58.5%
CRF	52.1%	76.3%	53.2%
LSTM	68%	68.1%	61.25%

Table 3.3: Language analysis using F1 score

As F1 score is a harmonic mean of precision and recall, there was, once again, an observation that LSTM outperformed both HMM and CRF in French and Hindi. Furthermore, similar to precision and recall, the margin between CRF and HMM was quite significant. In English, LSTM once more underperformed compared to both HMM and CRF. Similar to recall, HMM edged out CRF to be the best model for the English language. The LSTM F1 scores were similar for French and English, but there was a significant drop for Hindi. Conversely, both HMM and CRF performed somewhat similarly in Hindi and French, but were stronger by a margin of around 20% for English.

Speech Quality Observations

Model	High	Intermediate	Low
HMM	70.8%	64.1%	69.6%
CRF	70.8%	63%	67.1%
LSTM	73.4%	68.4%	72.7%

Table 3.4: Speech quality analysis using precision score

LSTM performed similarly for all qualities of speech, while there were wide discrepancies in the precision of HMM and CRF for all qualities. Notably, all three models performed the worst for intermediate speech. LSTM showed the greatest relative difference, being around five times greater than the difference in the precision scores for high and low quality speech. A difference on a similar scale occurred for HMM as well. In this regard, CRF was the exception insofar as the difference between the precision scores of intermediate-proficiency and low-proficiency speech was equal to the difference between the precision scores of low-proficiency and high-proficiency speech. Overall, all models performed best for high-proficiency speech.

Model	High	Intermediate	Low
HMM	67.6%	62.8%	64.7%
CRF	62.3%	58%	59.8%
LSTM	65.6%	64.5%	65.4%

Table 3.5: Speech quality analysis using recall score

LSTM once again performed consistently across all qualities of speech, however, once again the relative difference between intermediate-proficiency and low-proficiency speech was around five times greater than the relative difference between high-proficiency and intermediate-proficiency speech. However, this relative difference was not seen in either HMM and CRF, where the relative differences were around equal. Across all three models, the best performance was once again recorded for high-proficiency speech while the worst proficiency was seen in intermediate-proficiency speech. Notably, in high-proficiency speech, HMM performed better than LSTM by the same margin that LSTM outperformed CRF. This performance was not seen in other categories

Model	High	Intermediate	Low
HMM	67%	63.9%	64.5%
CRF	63.2%	59.4%	60.9%
LSTM	66.5%	65.7%	66.6%

Table 3.6: Speech quality analysis using F1 score

As F1 score is simply the harmonic mean between the precision and recall, trends which were seen in both precision and recall were repeated for F1 scores. Notably the fact that LSTM performs consistently across all proficiencies of speech. However, one interesting change is that low-proficiency speech slightly outperforms high-proficiency speech for LSTM – different from both the precision and recall scores. However, this may be because of a rounding error. Furthermore, for all models, high-proficiency speech once again has the highest F1 score while intermediate-proficiency speech has the lowest F1 scores. The significant relative difference seen between intermediate-proficiency and low-proficiency and high-proficiency and intermediate-proficiency speech is no longer as apparent in the case of HMM, though it is still extremely visible in the case of LSTM.

Concluding Observations

Long Short-Term Memory (LSTM) was the best model in 72% of the categories discussed above. This proves that it is the best performing model. Meanwhile, HMM performed the best in 22% of the categories. CRF only performed the best out of the three models in just one out of the eighteen categories discussed – that being for precision in a comparison of how the models performed in the English language. Even there, it was only slightly greater than the score received by HMM.

LSTM managed to be much more consistent no matter what proficiency of speech was used. It was also consistent across various languages, though less so than seen in the comparison between different proficiencies of speech. Examples of this include the F1 score of LSTM being much lower for Hindi, the recall score of LSTM being much lower for French, and the precision score of LSTM being much greater for English than the other two languages.

There was a significant difference in the performance for both HMM and CRF in the English language compared to other languages. This difference was so significant that HMM and CRF were even more accurate than LSTM for these categories. The magnitude of this difference was close to, or even surpassing, 5% in all categories. Meanwhile, the difference in the performance of HMM and CRF for the English language was marginal at best, with CRF outperforming HMM in terms of precision, as mentioned above.

Analysis

After analyzing the observations and coming up with strategies to improve the output of the study, whilst mitigating the prevailing circumstances, a final reflection on the initial hypothesis can be reached based on the results acquired.

Hypothesis Point 1

"The performance of POS tagging models will be highest for low-quality speech inputs, followed by average-quality speech, and lowest for high-quality speech."

This was disproven as the POS tagging models performed best for high-proficiency speech, followed by low-proficiency speech, with intermediate-proficiency speech performing the worst out of all categories. This is due to the fact that all three models work better with more complex datasets and the training data taken – especially in the case of UD_English-PUD and UD_Hindi-PUD – is taken from formal pieces of writing such as Wikipedia and news articles. Thus, it would benefit from more complex speech and would not be hindered by it, as initially assumed.

Meanwhile, low-proficiency speech often consisted of longer and more complex sentences with less lexical complexity. Initially, it was assumed that low-quality speech would include less complex sentences as well, which would have led it to perform worse than intermediate-quality speech. However, this turned out to be false as low-proficiency speech just means that the respondents were less coherent in their thoughts and thus needed more words to express the same thoughts, thereby increasing the complexity.

In terms of intermediate speech data, it is now apparent that intermediate speakers of any language speak in shorter sentences which are more coherent and display more lexical complexity. After deeper analysis, it was found that lexical complexity has less of an impact on the accuracy of the models compared to complexity of the sentences and thus, as the sentences are shorter and more coherent, the models performed the worst on them.

In conclusion, the models performed best for high-proficiency speech, followed by low-proficiency speech, which was then followed lastly by intermediate-proficiency speech.

Hypothesis Point 2

"The effectiveness of POS tagging models will differ between languages, with models performing better on English transcripts compared to French transcripts, and will perform the worst on Hindi due to potential language-specific challenges."

HMM and CRF performed significantly better on English compared to the other two languages, proving this claim of the hypothesis. The rationale behind this is that these models were built by people whose native languages were English and are based on inherent assumptions which are true for English but which may not be true for other languages, including related languages like French. Hindi is very far removed from English, notably in the fact that it is a subject-object-verb (SOV) language compared to the subject-verb-object (SVO) followed by English and French. This contributes to the significant difference in accuracy between English and Hindi.

However, a point to note is that the difference in accuracy between French and Hindi is not apparent. Though it would seem that English borrows so heavily from French and the fact that French is also an SVO language, the model would be more accurate for French than Hindi. However, this is not exactly true. In LSTM, the precision scores for both French and Hindi are equal, while in CRF, the precision scores for French are less than the precision scores for Hindi. Only the precision scores for HMM confirm the hypothesis that the scores for French should imply higher accuracy than Hindi.

In terms of recall, French is more accurate than Hindi, as predicted, for both HMM and CRF. However, this is offset by a significant 4% difference favoring Hindi for LSTM, once again disputing the credibility of the assumption that French should be more accurate than Hindi. In the case of F1 scores, however, the pattern reverses. HMM and CRF F1 scores suggest that they are more accurate for Hindi than French while the F1 score for LSTM suggests that the model is 7% more accurate for French than for Hindi.

In conclusion, while all three scores for all three models suggest that the accuracy for English is greater than the accuracy for French and Hindi, the assumption that the accuracy for French would be greater than the accuracy for Hindi remains under dispute.

Hypothesis Point 3

"Among the three models (HMM, CRF, LSTM), LSTM will exhibit the highest overall accuracy in POS tagging tasks, followed by CRF and then HMM."

LSTM, indeed, demonstrated the highest accuracy, being the most accurate model in 72% of comparisons. This is simply due to the neural network nature of the model, which makes it significantly more effective at POS tagging vis-à-vis CRF and HMM. There are various reasons LSTM did not perform better – according to the theoretical approach, LSTM should be better than both models in almost all cases, not just a supermajority. One of these reasons is that the datasets may have been too small. In the case of English, notably, the dataset was only 1,000 sentences or around around 22,000 tokens. This may not be large enough for LSTM to overcome the accuracy of HMM.

Another reason for HMM performing better than expected is the relatively simple nature of speech data compared to textual data, where the models are usually tested. This implies that LSTM may not have underperformed. Rather, HMM had a better performance than usual due to factors related to the dataset.

However, according to the given observations, HMM was more accurate than CRF – by a factor of around 16%. This is contrary to the conventional result where CRF is almost always more accurate than HMM. There could be various reasons for this. One of these reasons is the phenomenon of overfitting. Essentially overfitting means CRF is meant for more complicated sentences while speech data is generally more simple. CRF may try to make connections which would work for more complex data but are not realistic for relatively simplistic speech data.

Another reason that may have caused CRF to underperform compared to HMM may be the fact that speech data is different from the usual textual data these models are trained on. Textual data often has more complex and compound sentences, which CRF takes advantage of to improve its accuracy. However, given that speech data usually has smaller sentence sizes and sentence complexity as previously mentioned, HMM may perform better relative to CRF.

One final reason that could have caused the relatively poor performance of CRF could be the design of the CRF++ Toolkit, which may be designed in a way to prioritize complex sentences.

In conclusion, LSTM was the most accurate model, as predicted. However, it was not as accurate as predicted with HMM being more accurate in various cases. The reasons for this are manifold and relate to the nature of speech data rather than a fault in the model itself. Furthermore, HMM was significantly more accurate than CRF due to the aforementioned reasons and a potential fault in the model.

Strategies for Improving Results

Given that the hypothesis was disproved in various areas, and the fact that the observations often did not align with the theoretical expectations, there are some steps that could be taken to gather a more accurate viewpoint on the hypothesis.

The first step that could be taken is to use a larger dataset. The PUD datasets contain only 1,000 sentences which may not be large enough to train these models – especially LSTM, which is significantly more data-intensive than the other two models. Even the PUD dataset README file suggests using it in combination with the Sequoia dataset to improve the accuracy of results.

Other steps that could be taken is to use the same tagset instead of using both BIS POS and UD as it may have an impact on the results. This could be due to the fact that UD only has 17 tags while BIS POS has a few more tags. Furthermore, there is more ambiguity regarding how to tag corpora based on BIS POS compared to UD, which makes manual annotations more ambiguous and based on the annotator. Furthermore, using two different tagsets for Hindi may invalidate the results found for Hindi. If one tagset was used, especially for all languages, it would lead to more accurate results.

Conclusion

Throughout the report, there has been an in-depth examination of the performance of various POS tagging models on speech-to-text data. After constructing an initial hypothesis, followed by the performance of 9 interviews in 3 languages, multiple conclusions have been reached.

Notably, LSTM is the most accurate POS tagging model. However, its accuracy is highly dependent on the size of the dataset. With a smaller dataset, HMM may be a better and more computationally efficient choice. Furthermore, speech data – which generally has less sentence complexity than textual data – may be ill-suited for CRF, which performs better with more complex and compound sentences.

In terms of speech quality, POS taggers tend to struggle with intermediate-proficiency data due to its smaller sentence size and lower sentence complexity. They perform well for both high-proficiency and low-proficiency speech data. The reason for the good performance with high-proficiency data is that this data consists of more complex speech. Meanwhile, the reason that POS taggers work well with low-proficiency speech data is that this data is often less coherent, increasing words per sentence and thereby increasing the complexity of the sentence. As a rule of thumb, the more complex and/or less coherent the data, the better POS taggers work.

In terms of language, more research is needed to gain a concrete understanding due to the relatively small sizes of the training datasets used. However, preliminary observations suggest that POS tagging models work best for English and perform similarly for both Hindi and French.

In conclusion, the accuracy of POS taggers varies significantly based on the size of the dataset, the type of data, and the complexity and coherence of data, amongst other factors.

Appendix A: French Results

High-Proficiency Speaker

LSTM Results

Precision score: 74.50190180072343%
Recall score: 67.53246753246754%
F1 Score: 68.95959054216199%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	VERB
ADJ	10.0	0.0	6.0	1.0	0.0	2.0	0.0	4.0	0.0	0.0	6.0	0.0	0.0	4.0
ADP	0.0	48.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0
ADV	0.0	0.0	6.0	0.0	0.0	2.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0
AUX	1.0	0.0	0.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CCONJ	0.0	0.0	0.0	0.0	16.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DET	0.0	0.0	0.0	0.0	0.0	46.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
INTJ	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
NOUN	6.0	0.0	2.0	2.0	0.0	8.0	0.0	38.0	0.0	0.0	15.0	0.0	0.0	0.0
NUM	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
PRON	0.0	0.0	2.0	0.0	0.0	2.0	0.0	1.0	0.0	22.0	5.0	1.0	0.0	0.0
PROPN	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	2.0	0.0	0.0	1.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	39.0	0.0	0.0
SCONJ	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
VERB	4.0	2.0	10.0	8.0	0.0	3.0	0.0	5.0	0.0	2.0	2.0	0.0	0.0	26.0

CRF Results

Precision score: 64.34339405370953%
Recall score: 53.76623376623376%
F1 Score: 55.440497107215826%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	VERB
ADJ	13.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	1.0	0.0	9.0	0.0	0.0	5.0
ADP	1.0	43.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0
ADV	2.0	1.0	5.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
AUX	5.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
CCONJ	0.0	0.0	0.0	0.0	16.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DET	5.0	4.0	3.0	0.0	0.0	19.0	0.0	7.0	0.0	3.0	6.0	0.0	0.0	2.0
INTJ	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
NOUN	8.0	0.0	0.0	0.0	0.0	1.0	0.0	41.0	0.0	0.0	15.0	0.0	2.0	4.0
NUM	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
PRON	0.0	0.0	13.0	2.0	0.0	0.0	0.0	1.0	0.0	5.0	2.0	0.0	0.0	9.0
PROPN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	4.0	0.0	0.0	0.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	39.0	0.0	0.0
SCONJ	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
VERB	12.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	1.0	0.0	6.0	0.0	0.0	21.0

HMM Results

Precision score: 64.16140325710269%
Recall score: 62.077922077922075%
F1 Score: 60.48128909074122%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	\
ADJ	3.0	3.0	0.0	0.0	0.0	4.0	0.0	8.0	0.0	0.0	13.0	0.0	
ADP	0.0	46.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	1.0	0.0	
ADV	0.0	1.0	6.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	1.0	0.0	
AUX	0.0	0.0	0.0	0.0	0.0	1.0	0.0	3.0	0.0	0.0	0.0	0.0	
CCONJ	0.0	0.0	0.0	0.0	16.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
DET	0.0	6.0	0.0	0.0	0.0	40.0	0.0	1.0	0.0	2.0	0.0	0.0	
INTJ	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
NOUN	1.0	0.0	0.0	1.0	0.0	8.0	0.0	51.0	0.0	0.0	7.0	0.0	
NUM	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
PRON	0.0	7.0	1.0	0.0	0.0	8.0	0.0	0.0	0.0	9.0	0.0	0.0	
PROPN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	3.0	0.0	
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	39.0	
SCONJ	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	
VERB	1.0	4.0	0.0	2.0	0.0	6.0	0.0	11.0	0.0	5.0	3.0	0.0	

	SCONJ	VERB
ADJ	0.0	0.0
ADP	0.0	2.0
ADV	0.0	0.0
AUX	0.0	3.0
CCONJ	0.0	0.0
DET	0.0	0.0
INTJ	0.0	0.0
NOUN	0.0	2.0
NUM	0.0	0.0
PRON	0.0	7.0
PROPN	0.0	0.0
PUNCT	0.0	0.0
SCONJ	0.0	0.0
VERB	0.0	26.0

Intermediate-Proficiency Speaker

LSTM Results

Precision score: 76.48624374682298%
Recall score: 68.53002070393374%
F1 Score: 70.52875942242555%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	VERB	X
ADJ	11.0	0.0	7.0	1.0	0.0	4.0	0.0	4.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
ADP	0.0	60.0	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
ADV	4.0	0.0	11.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	2.0	0.0
AUX	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
CCONJ	0.0	0.0	0.0	0.0	22.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DET	1.0	0.0	1.0	0.0	0.0	64.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
INTJ	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
NOUN	5.0	0.0	6.0	0.0	0.0	8.0	0.0	49.0	0.0	0.0	17.0	0.0	0.0	4.0	0.0
NUM	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
PRON	0.0	0.0	3.0	0.0	0.0	6.0	0.0	1.0	0.0	26.0	5.0	1.0	0.0	1.0	0.0
PROPN	3.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	3.0	0.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	46.0	0.0	0.0	0.0
SCONJ	0.0	1.0	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
VERB	9.0	2.0	6.0	17.0	0.0	4.0	0.0	1.0	0.0	2.0	2.0	0.0	0.0	36.0	0.0
X	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

CRF Results

Precision score: 61.82634685118183%
Recall score: 51.345755693581786%
F1 Score: 52.76679682184145%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	VERB	X
ADJ	13.0	0.0	0.0	0.0	0.0	1.0	0.0	2.0	0.0	0.0	7.0	0.0	0.0	6.0	0.0
ADP	3.0	55.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0
ADV	5.0	0.0	7.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	2.0	0.0	0.0	2.0	0.0
AUX	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
CCONJ	0.0	0.0	1.0	0.0	21.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DET	8.0	4.0	5.0	0.0	0.0	21.0	0.0	10.0	1.0	3.0	9.0	0.0	1.0	5.0	5.0
INTJ	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
NOUN	19.0	0.0	0.0	0.0	0.0	1.0	0.0	47.0	0.0	0.0	20.0	0.0	0.0	2.0	0.0
NUM	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
PRON	5.0	0.0	18.0	0.0	0.0	0.0	0.0	1.0	0.0	4.0	2.0	0.0	0.0	12.0	0.0
PROPN	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	46.0	0.0	0.0	0.0
SCONJ	1.0	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
VERB	20.0	0.0	1.0	6.0	0.0	1.0	0.0	16.0	0.0	4.0	6.0	0.0	0.0	24.0	1.0
X	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

HMM Results

Precision score: 62.50504062540768%
Recall score: 61.28364389233955%
F1 Score: 59.318369566930585%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT
ADJ	4.0	3.0	0.0	2.0	0.0	1.0	0.0	8.0	0.0	0.0	6.0	0.0
ADP	2.0	56.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0
ADV	1.0	2.0	7.0	0.0	0.0	2.0	0.0	3.0	0.0	0.0	2.0	0.0
AUX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CCONJ	0.0	0.0	0.0	0.0	21.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
DET	0.0	7.0	0.0	2.0	0.0	52.0	0.0	7.0	0.0	3.0	0.0	0.0
INTJ	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NOUN	3.0	2.0	0.0	0.0	0.0	6.0	0.0	66.0	0.0	1.0	8.0	0.0
NUM	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
PRON	0.0	10.0	3.0	0.0	0.0	9.0	0.0	3.0	0.0	8.0	1.0	0.0
PROPN	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	10.0	0.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	46.0
SCONJ	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
VERB	0.0	5.0	0.0	4.0	0.0	19.0	0.0	19.0	0.0	5.0	0.0	0.0
X	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

	SCONJ	VERB	X
ADJ	0.0	2.0	0.0
ADP	0.0	0.0	0.0
ADV	0.0	0.0	0.0
AUX	0.0	1.0	0.0
CCONJ	0.0	0.0	0.0
DET	0.0	1.0	0.0
INTJ	0.0	0.0	0.0
NOUN	0.0	2.0	0.0
NUM	0.0	0.0	0.0
PRON	0.0	7.0	0.0
PROPN	0.0	0.0	0.0
PUNCT	0.0	0.0	0.0
SCONJ	0.0	0.0	0.0
VERB	0.0	26.0	0.0
X	0.0	0.0	0.0

Low-Proficiency Speaker

LSTM Results

Precision score: 72.33373925020243%
Recall score: 62.022471910112365%
F1 Score: 64.75331645697646%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	VERB	X
ADJ	10.0	0.0	17.0	2.0	0.0	2.0	0.0	1.0	0.0	0.0	5.0	0.0	0.0	5.0	0.0
ADP	0.0	52.0	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
ADV	2.0	0.0	11.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0
AUX	1.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
CCONJ	0.0	0.0	0.0	0.0	17.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DET	0.0	0.0	1.0	0.0	0.0	55.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
INTJ	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	0.0
NOUN	8.0	0.0	6.0	0.0	0.0	6.0	0.0	33.0	1.0	0.0	12.0	0.0	0.0	3.0	0.0
NUM	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	1.0	0.0
PRON	2.0	0.0	6.0	0.0	0.0	2.0	0.0	0.0	0.0	24.0	4.0	3.0	1.0	5.0	0.0
PROPN	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
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	47.0	0.0	0.0	0.0
SCONJ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	3.0	0.0	0.0
VERB	7.0	0.0	12.0	16.0	0.0	3.0	0.0	3.0	1.0	1.0	4.0	0.0	0.0	23.0	0.0
X	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

CRF Results

Precision score: 58.54494852776412%
Recall score: 46.741573033707866%
F1 Score: 48.84350806758618%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	VERB	X
ADJ	16.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	1.0	16.0	0.0	0.0	7.0	0.0
ADP	3.0	49.0	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
ADV	6.0	2.0	8.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0
AUX	3.0	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
CCONJ	0.0	0.0	0.0	0.0	17.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DET	14.0	2.0	2.0	0.0	0.0	15.0	0.0	8.0	2.0	1.0	9.0	0.0	0.0	4.0	3.0
INTJ	1.0	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
NOUN	18.0	1.0	0.0	1.0	0.0	0.0	0.0	28.0	2.0	2.0	12.0	0.0	1.0	4.0	0.0
NUM	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
PRON	5.0	0.0	13.0	0.0	0.0	0.0	0.0	2.0	0.0	6.0	5.0	0.0	0.0	16.0	0.0
PROPN	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	1.0	0.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	48.0	0.0	0.0	0.0
SCONJ	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0	0.0
VERB	10.0	0.0	1.0	1.0	0.0	1.0	0.0	25.0	0.0	1.0	9.0	0.0	0.0	20.0	2.0
X	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

HMM Results

Precision score: 61.821390909504444%
Recall score: 53.033707865168545%
F1 Score: 52.82712438615672%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	\
ADJ	6.0	1.0	0.0	0.0	0.0	2.0	0.0	8.0	0.0	1.0	16.0	0.0	
ADP	0.0	46.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	
ADV	0.0	5.0	8.0	0.0	0.0	2.0	0.0	2.0	0.0	1.0	1.0	0.0	
AUX	0.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	
CCONJ	0.0	0.0	0.0	0.0	17.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
DET	0.0	10.0	1.0	0.0	0.0	41.0	0.0	4.0	0.0	1.0	1.0	0.0	
INTJ	2.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
NOUN	3.0	2.0	0.0	2.0	0.0	8.0	0.0	44.0	0.0	0.0	4.0	0.0	
NUM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	
PRON	1.0	6.0	1.0	0.0	0.0	24.0	0.0	1.0	0.0	6.0	2.0	0.0	
PROPN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	45.0	
SCONJ	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
VERB	1.0	3.0	0.0	2.0	0.0	9.0	0.0	25.0	0.0	0.0	2.0	0.0	
X	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	

	SCONJ	VERB	X
ADJ	0.0	3.0	0.0
ADP	0.0	0.0	0.0
ADV	0.0	0.0	0.0
AUX	0.0	0.0	0.0
CCONJ	0.0	0.0	0.0
DET	0.0	0.0	0.0
INTJ	0.0	0.0	0.0
NOUN	0.0	1.0	1.0
NUM	0.0	0.0	1.0
PRON	1.0	3.0	0.0
PROPN	0.0	0.0	0.0
PUNCT	0.0	0.0	0.0
SCONJ	2.0	0.0	0.0
VERB	0.0	21.0	1.0
X	0.0	0.0	0.0

Appendix B: English Results

High-Proficiency Speaker

LSTM Results

Precision score: 76.77138872385837%
Recall score: 63.687150837988824%
F1 Score: 65.95367724643293%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	NOUN	PART	PRON	PROPN	PUNCT	SCONJ	VERB
ADJ	13.0	0.0	0.0	0.0	2.0	0.0	1.0	0.0	0.0	8.0	0.0	0.0	5.0
ADP	0.0	44.0	2.0	0.0	1.0	0.0	1.0	6.0	0.0	0.0	0.0	1.0	0.0
ADV	1.0	0.0	8.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0
AUX	0.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CCONJ	0.0	0.0	0.0	0.0	14.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DET	0.0	0.0	0.0	0.0	0.0	38.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NOUN	14.0	0.0	3.0	0.0	4.0	0.0	28.0	0.0	0.0	29.0	0.0	0.0	10.0
PART	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
PRON	0.0	0.0	1.0	0.0	3.0	0.0	0.0	0.0	13.0	4.0	0.0	0.0	0.0
PROPN	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	6.0	0.0	0.0	2.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.0	0.0	0.0
SCONJ	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
VERB	3.0	0.0	2.0	1.0	4.0	0.0	2.0	0.0	0.0	7.0	1.0	0.0	20.0

CRF Results

Precision score: 84.36781898881281%
Recall score: 78.2122905027933%
F1 Score: 80.17726699324726%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	NOUN	PART	PRON	PROPN	PUNCT	SCONJ	VERB
ADJ	18.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	5.0	0.0	0.0	1.0
ADP	0.0	45.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	2.0	0.0	0.0	4.0
ADV	1.0	0.0	4.0	0.0	0.0	0.0	4.0	1.0	0.0	2.0	0.0	0.0	0.0
AUX	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
CCONJ	0.0	0.0	0.0	0.0	14.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DET	0.0	0.0	0.0	0.0	0.0	34.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0
NOUN	4.0	0.0	0.0	1.0	0.0	0.0	77.0	0.0	0.0	3.0	0.0	0.0	1.0
PART	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
PRON	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	15.0	3.0	0.0	0.0	2.0
PROPN	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	4.0	1.0	0.0	0.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.0	0.0	0.0
SCONJ	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
VERB	2.0	0.0	1.0	0.0	0.0	0.0	6.0	0.0	0.0	3.0	0.0	0.0	26.0

HMM Results

Precision score: 81.26841339081146%
Recall score: 78.49162011173185%
F1 Score: 78.29456296171163%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	NOUN	PART	PRON	PROPN	PUNCT	\
ADJ	19.0	0.0	0.0	0.0	0.0	8.0	2.0	0.0	0.0	1.0	0.0	
ADP	0.0	51.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	
ADV	0.0	1.0	5.0	0.0	0.0	0.0	2.0	1.0	0.0	2.0	0.0	
AUX	0.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
CCONJ	0.0	0.0	0.0	0.0	14.0	0.0	0.0	0.0	0.0	0.0	0.0	
DET	0.0	0.0	0.0	0.0	0.0	38.0	0.0	0.0	0.0	0.0	0.0	
NOUN	3.0	0.0	0.0	1.0	0.0	2.0	75.0	0.0	0.0	6.0	1.0	
PART	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	
PRON	0.0	1.0	0.0	0.0	0.0	2.0	1.0	0.0	13.0	0.0	0.0	
PROPN	1.0	0.0	0.0	0.0	0.0	1.0	5.0	0.0	0.0	4.0	0.0	
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.0	
SCONJ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
VERB	1.0	5.0	0.0	0.0	0.0	5.0	10.0	0.0	0.0	0.0	0.0	

	SCONJ	VERB
ADJ	0.0	1.0
ADP	0.0	0.0
ADV	0.0	2.0
AUX	0.0	0.0
CCONJ	0.0	0.0
DET	0.0	0.0
NOUN	0.0	1.0
PART	0.0	0.0
PRON	2.0	0.0
PROPN	0.0	0.0
PUNCT	0.0	0.0
SCONJ	0.0	1.0
VERB	0.0	19.0

Intermediate-Proficiency Speaker

LSTM Results

Precision score: 69.84251833183728%
Recall score: 67.71739130434783%
F1 Score: 67.50846231751757%

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	VERB
ADJ	33.0	0.0	11.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	7.0	0.0	0.0	5.0
ADP	0.0	76.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0	1.0	0.0
ADV	6.0	0.0	78.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	1.0	4.0	0.0	0.0	5.0
AUX	1.0	0.0	2.0	32.0	0.0	0.0	0.0	3.0	0.0	3.0	0.0	2.0	0.0	0.0	6.0
CCONJ	0.0	0.0	0.0	0.0	48.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DET	0.0	0.0	0.0	0.0	0.0	48.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	4.0	0.0
INTJ	0.0	30.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
NOUN	10.0	0.0	1.0	0.0	1.0	0.0	0.0	73.0	0.0	1.0	0.0	19.0	3.0	0.0	11.0
NUM	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
PART	0.0	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	22.0	0.0	0.0	0.0	0.0	0.0
PRON	9.0	0.0	9.0	0.0	9.0	3.0	0.0	5.0	0.0	0.0	62.0	11.0	5.0	1.0	4.0
PROPN	2.0	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	5.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	79.0	0.0	0.0
SCONJ	0.0	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	9.0	0.0
VERB	3.0	5.0	3.0	19.0	1.0	1.0	0.0	15.0	0.0	0.0	0.0	11.0	1.0	0.0	59.0

CRF Results

Precision score: 74.15991340120146%
Recall score: 72.5%
F1 Score: 71.87309339567098%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	VERB
ADJ	32.0	0.0	2.0	0.0	0.0	0.0	0.0	15.0	1.0	0.0	0.0	2.0	0.0	1.0	6.0
ADP	0.0	80.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0
ADV	3.0	0.0	69.0	1.0	0.0	0.0	0.0	9.0	0.0	0.0	5.0	6.0	0.0	0.0	5.0
AUX	0.0	1.0	2.0	8.0	0.0	0.0	0.0	9.0	0.0	0.0	0.0	4.0	0.0	0.0	25.0
CCONJ	0.0	0.0	0.0	0.0	36.0	0.0	0.0	0.0	0.0	0.0	0.0	9.0	0.0	0.0	3.0
DET	0.0	0.0	0.0	0.0	0.0	41.0	0.0	0.0	0.0	0.0	5.0	2.0	0.0	0.0	4.0
INTJ	0.0	28.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	3.0	0.0	0.0	3.0
NOUN	3.0	0.0	0.0	0.0	0.0	0.0	0.0	102.0	1.0	0.0	0.0	5.0	0.0	0.0	6.0
NUM	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
PART	0.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	23.0	0.0	0.0	0.0	0.0	0.0
PRON	0.0	0.0	0.0	0.0	0.0	1.0	0.0	13.0	0.0	0.0	99.0	2.0	0.0	1.0	2.0
PROPN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	3.0	0.0	0.0	2.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	79.0	0.0	0.0
SCONJ	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	1.0	0.0	10.0	0.0
VERB	4.0	3.0	0.0	9.0	0.0	0.0	0.0	17.0	0.0	0.0	0.0	3.0	0.0	0.0	84.0

HMM Results

Precision score: 73.05991284218935%
Recall score: 73.15217391304348%
F1 Score: 72.65805088448502%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	\
ADJ	34.0	2.0	1.0	0.0	0.0	2.0	0.0	4.0	0.0	0.0	0.0	
ADP	0.0	75.0	1.0	0.0	0.0	0.0	0.0	2.0	0.0	4.0	0.0	
ADV	4.0	1.0	70.0	1.0	1.0	2.0	0.0	4.0	0.0	0.0	6.0	
AUX	1.0	3.0	1.0	30.0	0.0	4.0	0.0	3.0	0.0	0.0	1.0	
CCONJ	0.0	0.0	0.0	0.0	36.0	1.0	0.0	0.0	0.0	0.0	0.0	
DET	0.0	0.0	0.0	0.0	0.0	46.0	0.0	0.0	0.0	0.0	5.0	
INTJ	4.0	17.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
NOUN	2.0	1.0	0.0	1.0	0.0	3.0	0.0	98.0	0.0	0.0	0.0	
NUM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
PART	0.0	1.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	21.0	0.0	
PRON	0.0	1.0	1.0	0.0	0.0	1.0	0.0	7.0	0.0	0.0	96.0	
PROPN	0.0	0.0	0.0	0.0	0.0	2.0	0.0	2.0	0.0	0.0	0.0	
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
SCONJ	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	
VERB	2.0	5.0	0.0	17.0	1.0	5.0	0.0	17.0	0.0	0.0	0.0	

	PROPN	PUNCT	SCONJ	VERB
ADJ	2.0	3.0	0.0	12.0
ADP	0.0	0.0	0.0	4.0
ADV	1.0	0.0	2.0	0.0
AUX	0.0	0.0	0.0	6.0
CCONJ	0.0	0.0	0.0	0.0
DET	0.0	0.0	1.0	1.0
INTJ	0.0	0.0	0.0	13.0
NOUN	6.0	3.0	0.0	5.0
NUM	0.0	0.0	0.0	0.0
PART	0.0	0.0	0.0	0.0
PRON	2.0	1.0	4.0	5.0
PROPN	4.0	0.0	0.0	2.0
PUNCT	0.0	79.0	0.0	0.0
SCONJ	0.0	0.0	11.0	0.0
VERB	0.0	0.0	0.0	72.0

Low-Proficiency Speaker

LSTM Results

Precision score: 73.65607647913428%
Recall score: 71.02526002971769%
F1 Score: 71.21046266018745%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	VERB
ADJ	22.0	0.0	7.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	1.0	0.0	0.0	4.0
ADP	0.0	73.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	3.0	0.0
ADV	3.0	1.0	32.0	0.0	0.0	2.0	0.0	2.0	1.0	0.0	0.0	1.0	1.0	0.0	3.0
AUX	0.0	0.0	4.0	18.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	4.0
CCONJ	0.0	0.0	0.0	0.0	38.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DET	0.0	0.0	0.0	0.0	0.0	49.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
INTJ	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
NOUN	17.0	0.0	0.0	0.0	0.0	0.0	0.0	76.0	0.0	0.0	0.0	6.0	0.0	0.0	4.0
NUM	1.0	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
PART	0.0	4.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	9.0	0.0	0.0	0.0	0.0	0.0
PRON	6.0	0.0	3.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	49.0	10.0	7.0	3.0	8.0
PROPN	1.0	0.0	1.0	0.0	2.0	0.0	0.0	4.0	0.0	0.0	1.0	5.0	1.0	0.0	6.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	44.0	0.0	0.0
SCONJ	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	10.0	0.0
VERB	6.0	1.0	2.0	19.0	3.0	0.0	0.0	12.0	0.0	0.0	0.0	3.0	0.0	0.0	51.0

CRF Results

Precision score: 79.1243623735546%
Recall score: 76.3744279346211%
F1 Score: 77.11611953880544%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	VERB
ADJ	23.0	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	3.0
ADP	0.0	74.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	1.0	1.0
ADV	2.0	1.0	30.0	0.0	0.0	2.0	0.0	7.0	0.0	0.0	0.0	1.0	0.0	0.0	3.0
AUX	1.0	0.0	1.0	8.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	5.0	0.0	0.0	13.0
CCONJ	0.0	0.0	1.0	0.0	33.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0
DET	1.0	1.0	0.0	0.0	0.0	45.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
INTJ	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
NOUN	9.0	0.0	0.0	0.0	0.0	0.0	0.0	82.0	2.0	0.0	0.0	2.0	0.0	0.0	5.0
NUM	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	0.0
PART	0.0	2.0	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
PRON	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	73.0	4.0	0.0	7.0	2.0
PROPN	0.0	0.0	1.0	0.0	0.0	0.0	0.0	5.0	1.0	0.0	0.0	9.0	1.0	0.0	1.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	44.0	0.0	0.0
SCONJ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	11.0	0.0
VERB	2.0	2.0	0.0	11.0	0.0	1.0	0.0	14.0	0.0	0.0	0.0	2.0	0.0	0.0	65.0

HMM Results

Precision score: 80.80208241784189%
Recall score: 79.64338781575037%
F1 Score: 79.81267846957925%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	\
ADJ	24.0	1.0	2.0	1.0	0.0	3.0	0.0	4.0	0.0	0.0	0.0	
ADP	0.0	74.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
ADV	0.0	2.0	41.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	
AUX	1.0	0.0	1.0	14.0	0.0	4.0	0.0	1.0	0.0	0.0	1.0	
CCONJ	0.0	1.0	0.0	0.0	33.0	0.0	0.0	0.0	0.0	0.0	0.0	
DET	0.0	1.0	0.0	0.0	0.0	47.0	0.0	1.0	0.0	0.0	2.0	
INTJ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
NOUN	4.0	3.0	0.0	1.0	0.0	3.0	0.0	83.0	0.0	0.0	0.0	
NUM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	
PART	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.0	0.0	
PRON	0.0	1.0	0.0	0.0	0.0	2.0	0.0	3.0	0.0	0.0	77.0	
PROPN	1.0	1.0	0.0	0.0	0.0	2.0	0.0	3.0	0.0	0.0	0.0	
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
SCONJ	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
VERB	2.0	2.0	0.0	19.0	0.0	0.0	0.0	12.0	0.0	0.0	0.0	

	PROPN	PUNCT	SCONJ	VERB
ADJ	0.0	0.0	0.0	3.0
ADP	0.0	0.0	3.0	0.0
ADV	1.0	0.0	0.0	0.0
AUX	0.0	0.0	0.0	9.0
CCONJ	0.0	0.0	0.0	0.0
DET	0.0	0.0	0.0	0.0
INTJ	1.0	0.0	0.0	0.0
NOUN	4.0	1.0	0.0	4.0
NUM	0.0	0.0	0.0	0.0
PART	0.0	0.0	0.0	0.0
PRON	1.0	0.0	5.0	0.0
PROPN	11.0	1.0	0.0	2.0
PUNCT	0.0	44.0	0.0	0.0
SCONJ	0.0	0.0	12.0	0.0
VERB	1.0	2.0	0.0	59.0

Appendix C: Hindi Results

High-Proficiency Speaker

LSTM Results

Precision score: 70.45961861469883%
Recall score: 65.50802139037432%
F1 Score: 64.63852163161778%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTF	INTJ	NOUN	PART	PRON	PROPN	PUNCT	RP	VERB
ADJ	12.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	4.0	1.0	2.0	0.0	0.0	0.0	0.0
ADP	0.0	41.0	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
ADV	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	1.0
AUX	4.0	1.0	0.0	23.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	6.0
CCONJ	0.0	5.0	0.0	0.0	12.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0
DET	0.0	0.0	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
INTF	1.0	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	0.0
INTJ	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
NOUN	14.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	41.0	0.0	11.0	0.0	1.0	0.0	0.0
PART	1.0	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
PRON	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0	60.0	0.0	1.0	0.0	2.0
PROPN	2.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	1.0	1.0	0.0	1.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	23.0	0.0	0.0
RP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0
VERB	14.0	0.0	0.0	6.0	0.0	0.0	0.0	0.0	2.0	1.0	2.0	0.0	0.0	0.0	23.0

CRF Results

Precision score: 63.77917930949236%
Recall score: 54.81283422459893%
F1 Score: 54.299555183073565%
Confusion matrix:

	ADV	CC	CCD	INJ	INTF	JJ	NEG	NN	NNP	PRP	PSP	PUNC	QT	RB	RP	VAUX	VM
ADV	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
CC	0.0	0.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0
CCD	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
INJ	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
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
JJ	0.0	0.0	0.0	0.0	0.0	14.0	0.0	3.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0
NEG	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	0.0	0.0	0.0	0.0
NN	0.0	0.0	0.0	0.0	0.0	2.0	0.0	56.0	4.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
NNP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
PRP	0.0	0.0	0.0	0.0	0.0	11.0	0.0	16.0	14.0	15.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0
PSP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	38.0	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	23.0	0.0	0.0	0.0	0.0	0.0
QT	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
RB	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
RP	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
VAUX	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	0.0	0.0	17.0	16.0
VM	0.0	0.0	0.0	0.0	0.0	3.0	0.0	10.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	31.0

HMM Results

Precision score: 66.88622949815235%
Recall score: 62.299465240641716%
F1 Score: 62.15632863242904%
Confusion matrix:

	ADV	CC	CCD	INJ	INTF	JJ	NEG	NN	NNP	PRP	PSP	PUNC	QT	\
ADV	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	
CC	0.0	0.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	4.0	0.0	0.0	
CCD	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	
INJ	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	
INTF	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
JJ	0.0	0.0	0.0	0.0	0.0	16.0	0.0	3.0	0.0	1.0	0.0	0.0	0.0	
NEG	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	
NN	0.0	0.0	0.0	0.0	0.0	2.0	0.0	59.0	0.0	0.0	3.0	0.0	0.0	
NNP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	1.0	0.0	3.0	0.0	0.0	
PRP	0.0	0.0	0.0	0.0	0.0	7.0	0.0	8.0	10.0	29.0	3.0	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	38.0	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	23.0	0.0	
QT	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	
RB	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	
RP	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	
VAUX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0	0.0	
VM	0.0	0.0	0.0	0.0	0.0	1.0	0.0	9.0	0.0	1.0	1.0	0.0	0.0	

	RB	RP	VAUX	VM
ADV	0.0	0.0	0.0	0.0
CC	0.0	0.0	0.0	0.0
CCD	0.0	0.0	0.0	0.0
INJ	0.0	0.0	0.0	0.0
INTF	0.0	0.0	0.0	0.0
JJ	0.0	0.0	0.0	0.0
NEG	0.0	0.0	0.0	0.0
NN	0.0	0.0	0.0	2.0
NNP	0.0	0.0	0.0	1.0
PRP	0.0	0.0	1.0	2.0
PSP	0.0	0.0	0.0	0.0
PUNC	0.0	0.0	0.0	0.0
QT	0.0	0.0	0.0	0.0
RB	0.0	0.0	0.0	0.0
RP	0.0	0.0	0.0	0.0
VAUX	0.0	0.0	22.0	11.0
VM	0.0	0.0	2.0	34.0

Intermediate-Proficiency Speaker

LSTM Results

Precision score: 59.01318377989235%
Recall score: 57.399103139013455%
F1 Score: 55.41662443618599%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DEM	DET	INTF	...	PRON	PROPN	PUNCT	QC	QF	RP	VADV	VERB
ADJ	7.0	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
ADP	0.0	25.0	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
ADV	0.0	0.0	1.0	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
AUX	2.0	1.0	0.0	13.0	0.0	0.0	0.0	0.0	...	1.0	0.0	3.0	0.0	0.0	0.0	0.0	4.0
CCONJ	1.0	3.0	0.0	0.0	5.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
DEM	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
DET	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
INTF	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NOUN	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	3.0	2.0	0.0	0.0	0.0	0.0	0.0	3.0
PART	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
PRON	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	30.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
PROPN	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	16.0	0.0	0.0	0.0	0.0	0.0
QC	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
QF	2.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RP	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
VADV	0.0	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
VERB	0.0	0.0	0.0	7.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0

CRF Results

Precision score: 52.95616818966131%
Recall score: 50.224215246636774%
F1 Score: 48.49266876892379%
Confusion matrix:

	CC	DEM	INTF	JJ	NEG	NN	NNP	NST	PRP	PSP	PUNC	QC	QF	QT	RB	RP	VAUX	VM	VRB
CC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DEM	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
INTF	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
JJ	0.0	0.0	0.0	1.0	0.0	5.0	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
NEG	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	0.0
NN	0.0	0.0	0.0	1.0	0.0	28.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0
NNP	0.0	0.0	0.0	0.0	0.0	1.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0
NST	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
PRP	0.0	0.0	0.0	4.0	0.0	11.0	9.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
PSP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	24.0	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	16.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
QC	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
QF	0.0	0.0	0.0	2.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
QT	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
RB	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	1.0	0.0	0.0	0.0	0.0
RP	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	1.0	0.0
VAUX	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.0	6.0	0.0
VM	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	4.0	14.0	0.0
VRB	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	0.0	0.0	0.0	0.0

HMM Results

Precision score: 56.84968489212514%
Recall score: 53.81165919282511%
F1 Score: 52.724029725747116%
Confusion matrix:

	CC	DEM	INTF	JJ	NEG	NN	NNP	NST	PRP	PSP	PUNC	QC	QF	\
CC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	
DEM	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	
INTF	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	
JJ	0.0	0.0	0.0	3.0	0.0	2.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	
NEG	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	
NN	0.0	0.0	0.0	2.0	0.0	26.0	1.0	0.0	0.0	3.0	1.0	0.0	0.0	
NNP	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	1.0	0.0	0.0	0.0	
NST	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	
PRP	0.0	0.0	0.0	1.0	0.0	4.0	8.0	0.0	11.0	3.0	0.0	0.0	0.0	
PSP	0.0	0.0	0.0	0.0	0.0	1.0	0.0	2.0	0.0	24.0	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	16.0	0.0	0.0	
QC	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	
QF	0.0	0.0	0.0	2.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
QT	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	
RB	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	
RP	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	
VAUX	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	
VM	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	
VRB	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	

	QT	RB	RP	VAUX	VM	VRB
CC	0.0	0.0	0.0	0.0	0.0	0.0
DEM	0.0	0.0	0.0	0.0	0.0	0.0
INTF	0.0	0.0	0.0	0.0	0.0	0.0
JJ	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
NN	0.0	0.0	0.0	2.0	3.0	0.0
NNP	0.0	0.0	0.0	0.0	2.0	0.0
NST	0.0	0.0	0.0	0.0	0.0	0.0
PRP	0.0	0.0	0.0	0.0	2.0	0.0
PSP	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
QC	0.0	0.0	0.0	0.0	0.0	0.0
QF	0.0	0.0	0.0	0.0	0.0	0.0
QT	0.0	0.0	0.0	0.0	0.0	0.0
RB	0.0	1.0	0.0	0.0	0.0	0.0
RP	0.0	0.0	0.0	0.0	1.0	0.0
VAUX	0.0	0.0	0.0	16.0	6.0	0.0
VM	0.0	0.0	0.0	5.0	14.0	0.0
VRB	0.0	0.0	0.0	0.0	1.0	0.0

Low-Proficiency Speaker

LSTM Results

Precision score: 72.2182026943345%
Recall score: 63.27543424317618%
F1 Score: 63.7480283223264%
Confusion matrix:

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTF	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	RP	SYM	VERB
ADJ	16.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
ADP	2.0	42.0	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
ADV	1.0	1.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
AUX	2.0	0.0	0.0	29.0	0.0	0.0	0.0	0.0	2.0	0.0	1.0	0.0	2.0	0.0	0.0	16.0
CCONJ	0.0	1.0	0.0	0.0	14.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0
DET	2.0	0.0	0.0	0.0	0.0	12.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
INTF	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
INTJ	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
NOUN	30.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	41.0	0.0	2.0	1.0	4.0	0.0	0.0	1.0
NUM	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
PRON	2.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0	43.0	0.0	3.0	0.0	0.0	0.0
PROPN	13.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	3.0	2.0	1.0	0.0	0.0	4.0
PUNCT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	30.0	0.0	0.0	0.0
RP	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
SYM	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
VERB	9.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	21.0

CRF Results

Precision score: 63.71684588448121%
Recall score: 56.32754342431762%
F1 Score: 56.69454645477462%
Confusion matrix:

	CCD	INJ	INTF	JJ	NN	NNP	PRP	PSP	PUNC	QT	QTC	QTF	RB	RP	SYM	VAUX	VM
CCD	13.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
INJ	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
INTF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0
JJ	0.0	0.0	0.0	13.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
NN	0.0	0.0	0.0	6.0	63.0	5.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	5.0
NNP	0.0	0.0	0.0	1.0	12.0	5.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0
PRP	0.0	0.0	0.0	5.0	13.0	9.0	11.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
PSP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.0	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	30.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
QT	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	9.0	1.0	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	2.0	0.0	0.0	0.0	0.0	0.0	0.0
QTF	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
RB	0.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0
RP	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
SYM	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
VAUX	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	24.0	26.0
VM	0.0	0.0	0.0	2.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	22.0

HMM Results

Precision score: 66.22697642858932%
Recall score: 61.53846153846154%
F1 Score: 60.99071368316152%
Confusion matrix:

	CCD	INJ	INTF	JJ	NN	NNP	PRP	PSP	PUNC	QT	QTC	QTF	RB	V
CCD	13.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	
INJ	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	
INTF	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	
JJ	0.0	0.0	0.0	9.0	12.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	
NN	0.0	0.0	0.0	2.0	67.0	2.0	1.0	8.0	0.0	0.0	0.0	0.0	0.0	
NNP	0.0	0.0	0.0	1.0	14.0	3.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	
PRP	0.0	0.0	0.0	3.0	6.0	5.0	22.0	4.0	1.0	0.0	0.0	0.0	0.0	
PSP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	41.0	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	30.0	0.0	0.0	0.0	0.0	
QT	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	9.0	2.0	0.0	
QTC	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
QTF	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
RB	0.0	0.0	1.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	2.0	
RP	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	
SYM	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	
VAUX	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	
VM	0.0	0.0	0.0	0.0	2.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	

	RP	SYM	VAUX	VM
CCD	0.0	0.0	0.0	0.0
INJ	0.0	0.0	0.0	0.0
INTF	0.0	0.0	0.0	0.0
JJ	0.0	0.0	0.0	0.0
NN	0.0	0.0	0.0	3.0
NNP	0.0	0.0	0.0	2.0
PRP	0.0	0.0	0.0	0.0
PSP	0.0	0.0	0.0	0.0
PUNC	0.0	0.0	0.0	0.0
QT	0.0	0.0	0.0	0.0
QTC	0.0	0.0	0.0	0.0
QTF	0.0	0.0	0.0	0.0
RB	0.0	0.0	0.0	0.0
RP	0.0	0.0	1.0	0.0
SYM	0.0	0.0	0.0	0.0
VAUX	0.0	0.0	31.0	20.0
VM	0.0	0.0	3.0	27.0