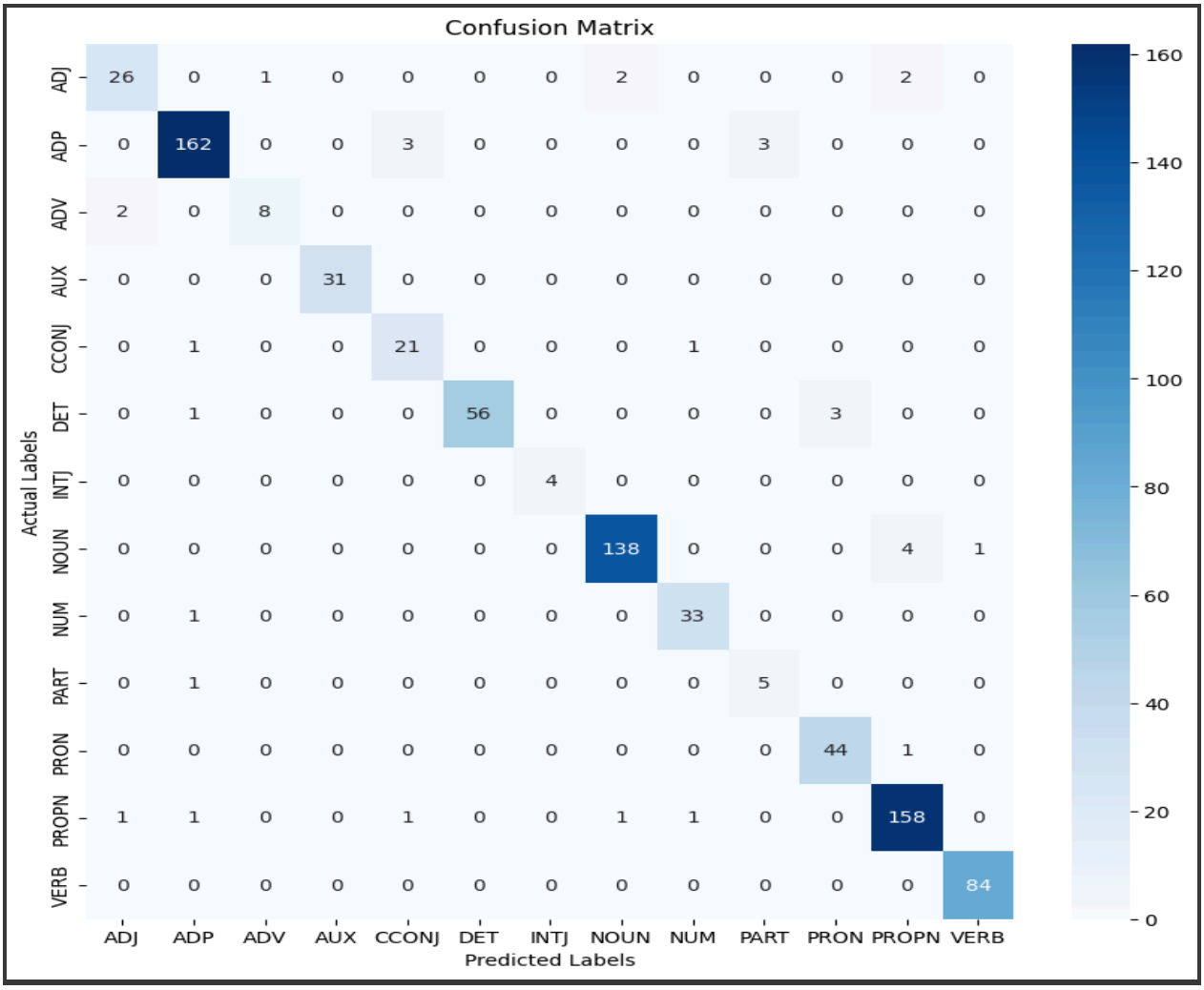
FEED FORWARD NEURAL NETWORK - POS Tagging

Hyperparameters used to train the model: Number of hidden layers,sizes of the hidden layers,activation functions,window size(default s=2 and p=2 used for all the below evaluation metrics),learning rate=.001 for all the models,used a pre trained word2vec model from google which has 300 embedding dimensions,number of epochs used to train all the models are 50.

Model 1: 1 Hidden layer,Length of hidden layer is 128,activation function used is the ReLu .

```
Accuracy: 0.9601
Macro Recall: 0.9358
Macro Precision: 0.9249
Macro F1: 0.9287
Micro Recall: 0.9601
Micro Precision: 0.9601
Micro F1: 0.9601
```

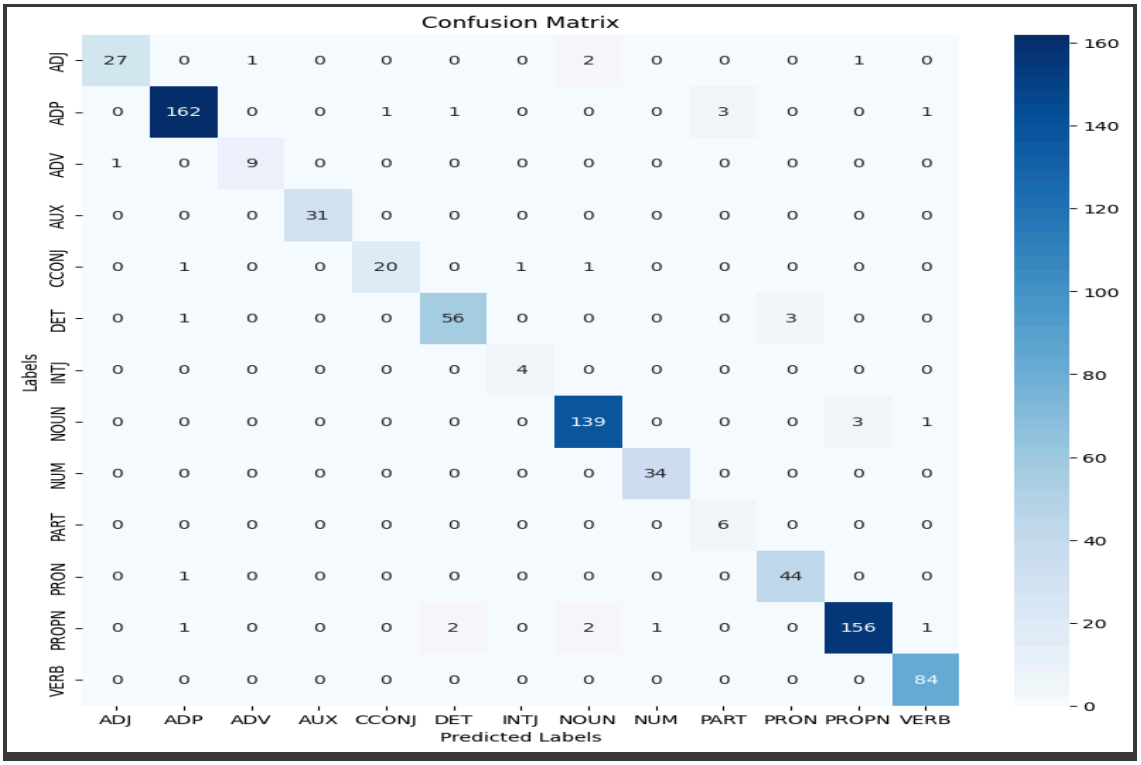


Analysis:

- The confusion matrix's diagonal elements, representing correct predictions, exhibit high values for most POS tags, indicating the model's general effectiveness.
- ADP (adposition) class shows the highest number of correct predictions, with 162 instances classified accurately, followed by NOUN with 138 correct classifications.
- Misclassifications are evident in the non-diagonal elements, such as ADP occasionally being misclassified as DET (determiner), and NOUN being mistaken for ADJ (adjective), NUM (numeral), and VERB.
- VERB class experiences common misclassifications, with predictions often confused with NOUN and ADJ.
- Some classes like CONJ and DET exhibit relatively low correct predictions compared to others, suggesting areas for potential improvement.

Model 2: 2 Hidden layer,Length of hidden layer is 512,activation function used in both the layers are ReLu .

```
Accuracy: 0.9651
Macro Recall: 0.9530
Macro F1: 0.9362
Micro Recall: 0.9651
Micro F1: 0.9651
```

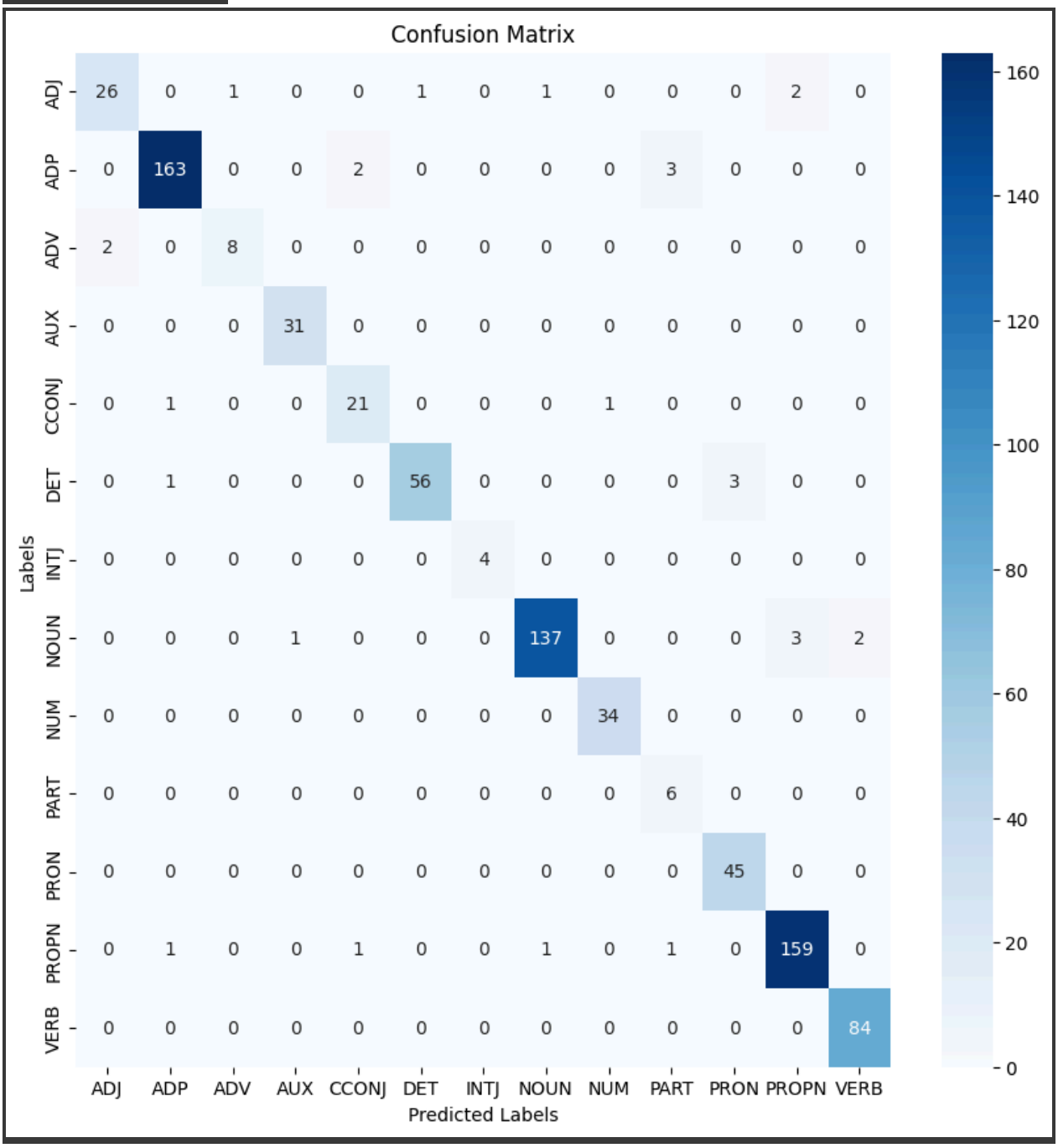


Analysis:

- The majority of predictions lie on the diagonal, indicating correct classifications. This is evident in the darker colors along the diagonal, especially for ADP, NOUN, VERB, and PROPN (proper noun), which suggests high accuracy for these POS tags.
- The number of correct predictions for each POS tag is relatively high, as indicated by the darker squares on the diagonal, with ADP and NOUN having the most correct classifications at 162 and 139, respectively.
- Misclassifications are relatively low for most classes but still present. For instance, NOUN has been misclassified as NUM (numeral) and ADJ (adjective) a few times.
- There are some cases where VERB is confused with NOUN and vice versa, which is a common challenge in POS tagging due to the syntactic roles these words can play.

Model 3: 1 Hidden layer,Length of hidden layer is 64,activation function used is tanh.

```
Accuracy: 0.9626
Macro Recall: 0.9573
Macro F1: 0.9372
Micro Recall: 0.9626
Micro F1: 0.9626
```



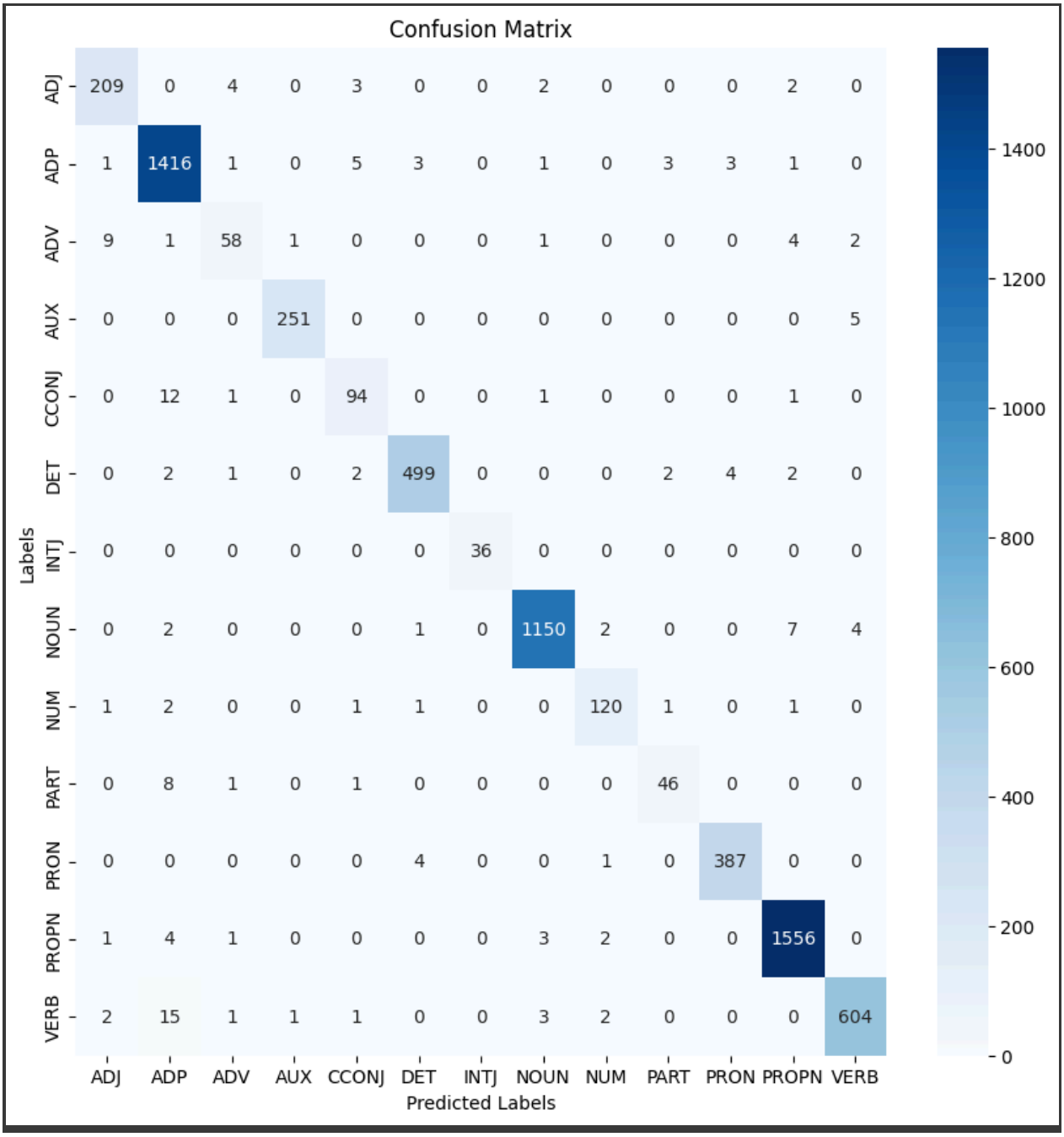
Analysis:

- The ADP tag has the highest number of correct predictions (163), similar to the previous model, followed closely by NOUN (137), VERB (159), and PROPN (84). These classes seem to be well-learned by the model.
- Misclassifications are relatively fewer, but certain patterns of confusion can be observed. For example, NOUN is sometimes confused with NUM, ADJ, and VERB, which is similar to the confusion observed in the previous models.
- The lower number of hidden units in this model (64 compared to 512 in the best-performing model) doesn't seem to have significantly impacted its ability to classify the majority of tags correctly.
- In conclusion, this third model shows robust performance, similar to my best model, and serves as a good candidate for POS tagging tasks, especially if computational efficiency is a concern.

Best Model: The best model from the evaluation metrics is the model number 2.

Evaluation metrics on Test set: (for s=p=2)

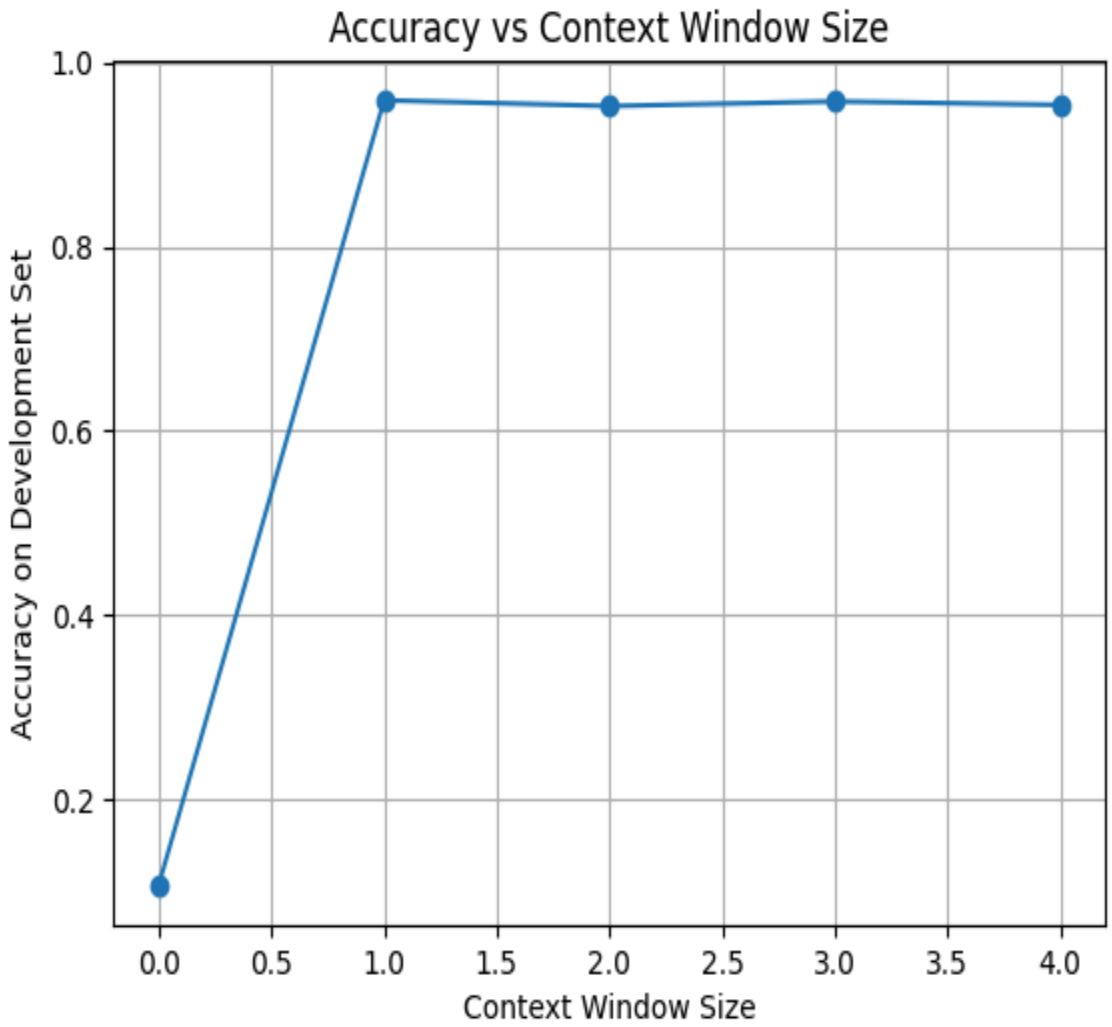
```
Accuracy: 0.9775
Macro Recall: 0.9414
Macro Precision: 0.9511
Macro F1: 0.9455
Micro Recall: 0.9775
Micro Precision: 0.9775
Micro F1: 0.9775
```



Analysis:

- The confusion matrix shows a large number of correct predictions, as indicated by the high values along the diagonal, particularly for ADP (1416 correct predictions), NOUN (1150), and VERB (1556). This suggests that the model is highly accurate for these parts of speech.
- The matrix also shows some misclassifications, but they are relatively low compared to the correct predictions. Common misclassifications include ADJ being mistaken as NOUN and ADV (adverb) being mistaken as ADP.
- The AUX (auxiliary verb) class has a notable number of misclassifications as VERB, which could be due to the linguistic similarity between these categories.

Graph for context_window ∈ {0...4} vs dev set accuracy:



Analysis:

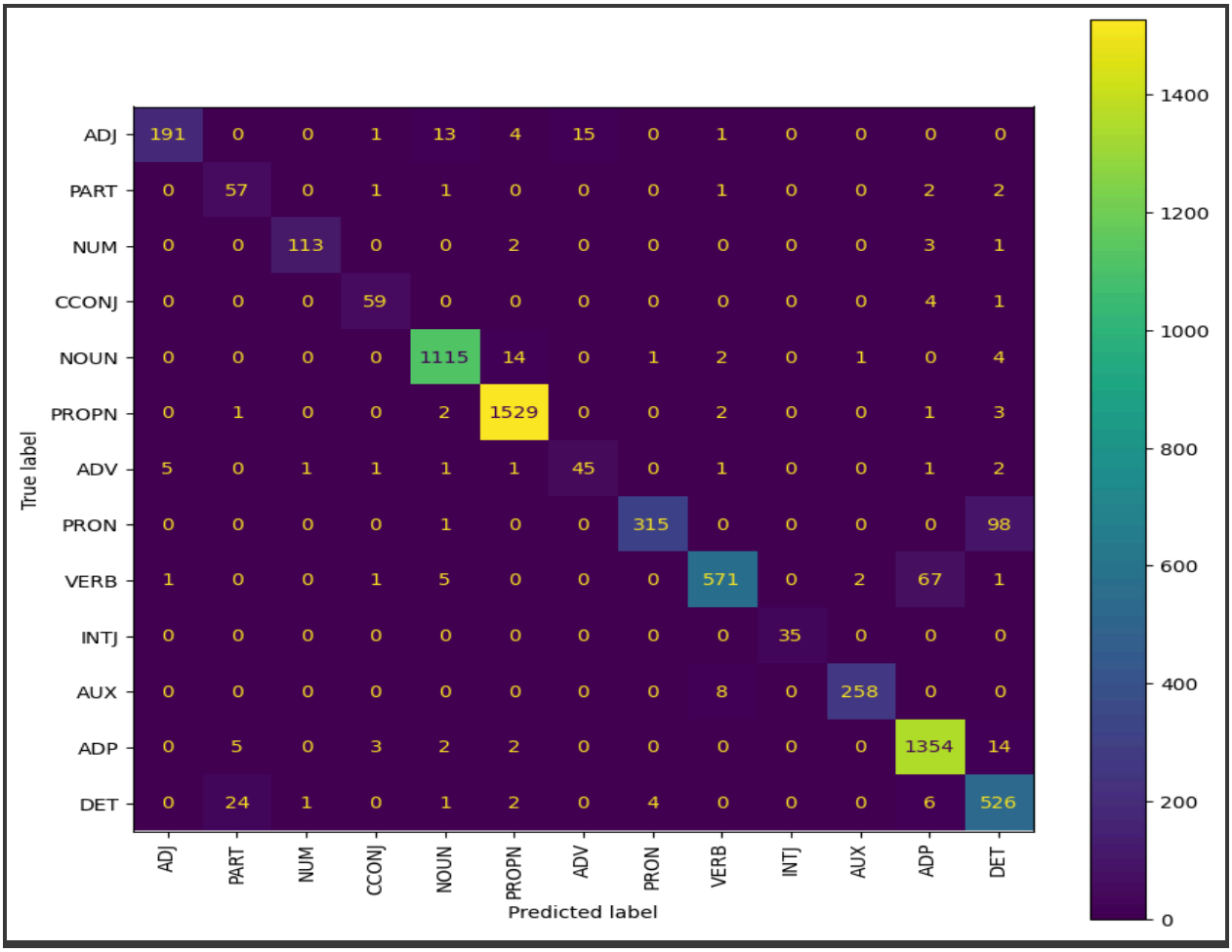
- The graph illustrates a significant rise in accuracy as the context window size expands from 0 to 1. This underscores the vital role of incorporating context, or surrounding words, in POS tagging, as words often rely heavily on their context for accurate classification.
- Once the context window size surpasses 1, the accuracy levels off. This suggests that enlarging the context beyond a single surrounding word on each side does not notably enhance model performance within this dataset and model architecture.
- Considering the plateau effect, a context window size of 1 emerges as the optimal choice for this task given the present model setup and dataset characteristics. This size strikes a balance between computational efficiency and performance.
- Larger context windows demand increased computational resources for each prediction, as more input features are processed by the neural network. Given that accuracy does not improve with larger windows, opting for a smaller window size proves to be computationally efficient.
- The substantial accuracy leap from a context window size of 0 to 1 indicates that unigrams (single words) alone are inadequate for capturing the necessary information for POS tagging. However, bigrams, or a window size encompassing one word before and after the target, seem to capture essential information that aids in classification.

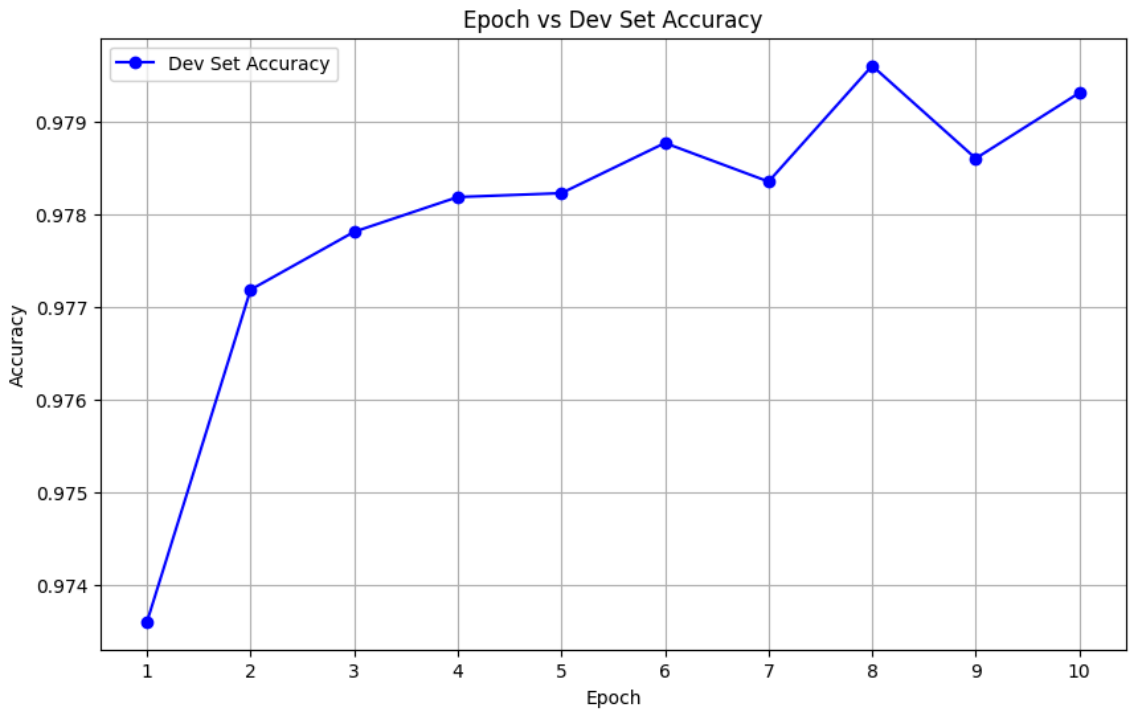
LSTM - POS Tagging

Hyperparameters used to train the model: Number of hidden layers,sizes of the hidden layers/dimensions,activation functions,bidirectionality,learning rate=.001 for all the models,used a pre trained word2vec model from google which has 300 embedding dimensions,number of epochs used to train all the models are 10.

Model 1: 2 LSTM layers,Size of hidden layer is 128,activation function used is the ReLu .

```
Accuracy: 0.9794
Macro Recall: 0.8165
Macro F1: 0.8318
Micro Recall: 0.9794
Micro F1: 0.9794
```





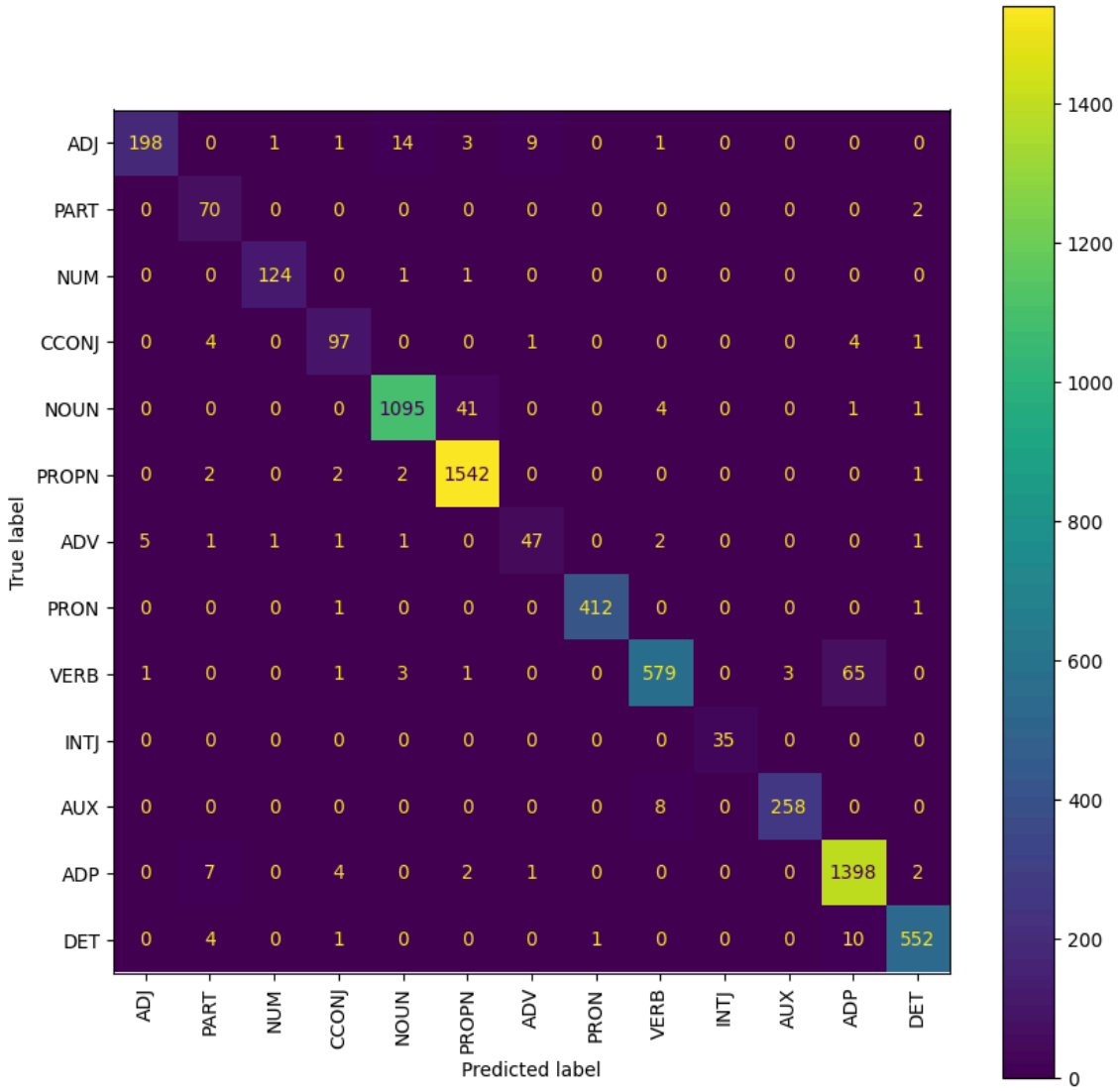
Analysis:

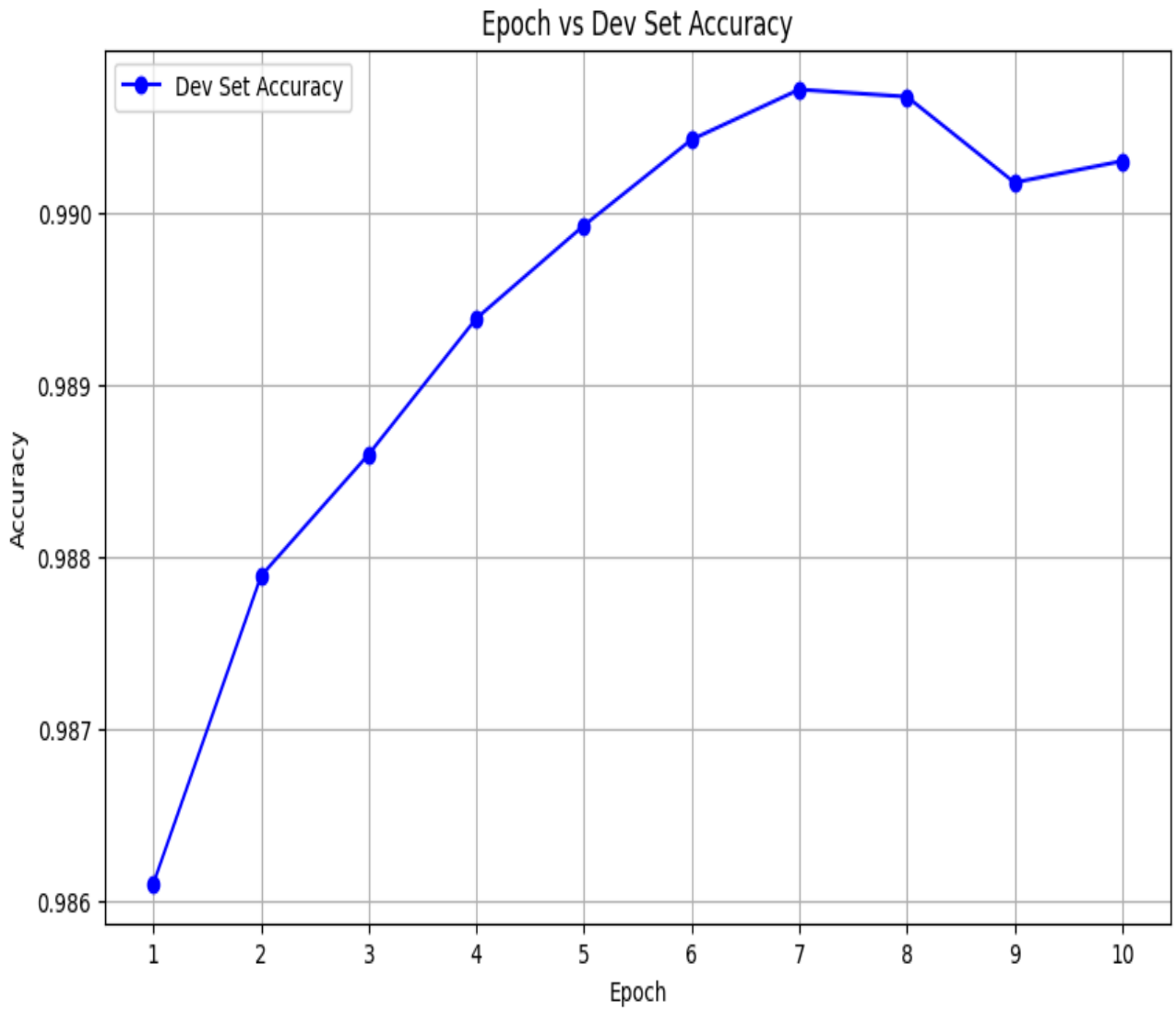
- Diagonal cells in the confusion matrix represent the number of correct predictions for each POS tag, indicating good performance when values are high relative to corresponding rows and columns.
- Tags like NOUN, PROPN, AUX, DET, and ADP show exceptionally high counts of true positives (correct predictions), reflected in darker cells on the diagonal, indicating good model performance for these tags.
- Off-diagonal cells represent confusion between tags, with lighter cells indicating infrequent misclassifications, a positive sign of model performance.
- Confusion between ADJ and NOUN is somewhat common due to adjectives sometimes being used as nouns (nominalization) depending on context.
- Similarly, confusion between VERB and AUX can be expected since auxiliary verbs can sometimes function as full verbs and vice versa.
- Micro-average recall and F1-score treat every instance (token prediction) equally, indicating overall good performance across individual instances.
- Macro-average recall and F1-score treat every class (POS tag) equally, suggesting balanced performance across different POS tags without significant bias toward more frequent tags.
- The model's accuracy improves significantly in the first few epochs, indicating rapid learning from training data and fitting to major patterns in the data.
- From epoch 4 onwards, there's a slight variability in accuracy, which could be attributed to the model's learning algorithm making incremental adjustments. These adjustments may either slightly improve or degrade performance on the development set.

- Variability in accuracy can occur as the model starts to overfit to the training data or navigates areas of the parameter space where optimization surfaces are complex.

Model 2: Length of hidden dimension is 64, the activation function used is tanh,MNumber of hidden layer is 1 ,bidirectionality is used.

```
Accuracy: 0.9896
Macro Recall: 0.8829
Macro F1: 0.8830
Micro Recall: 0.9896
Micro F1: 0.9896
```



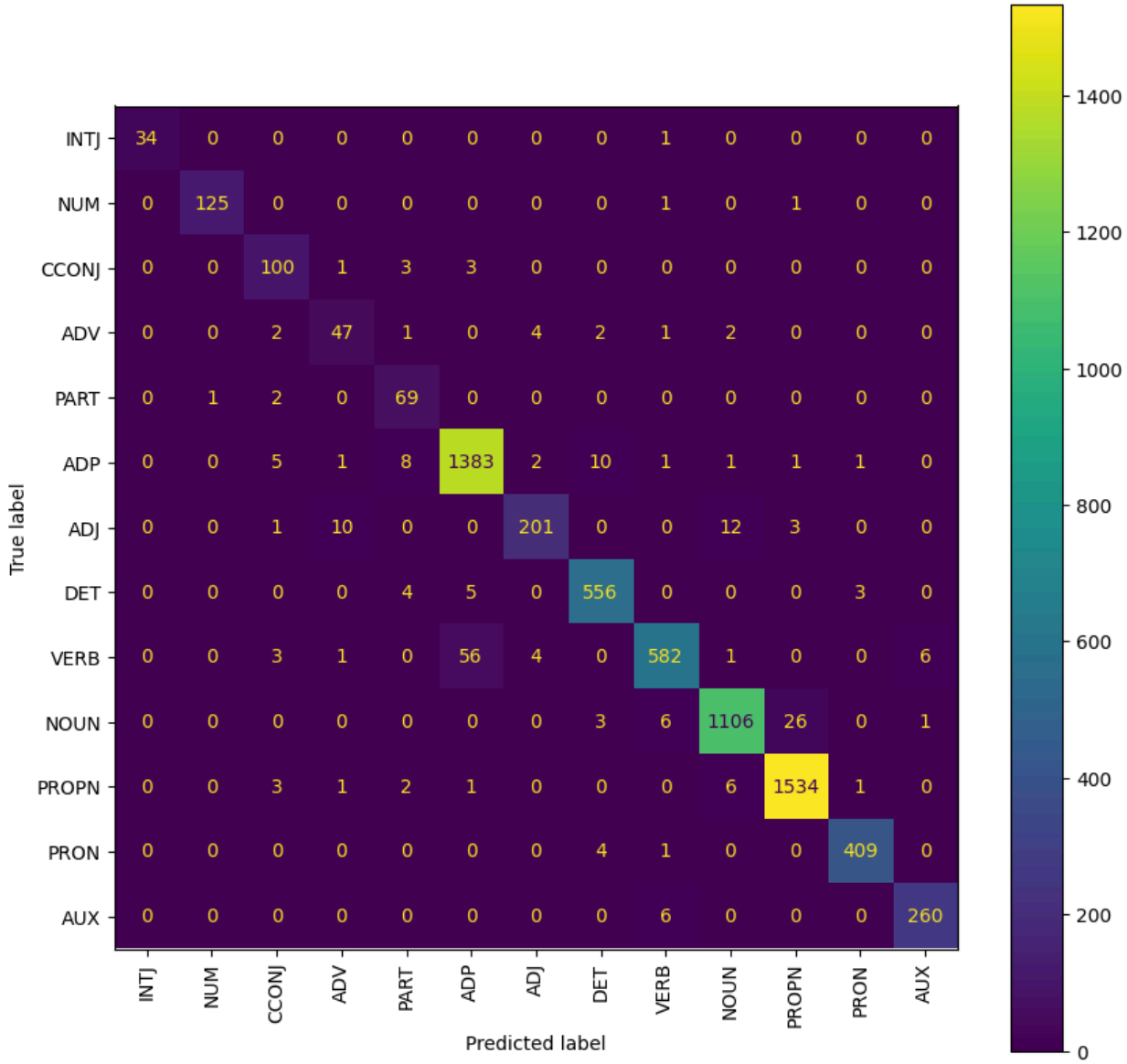


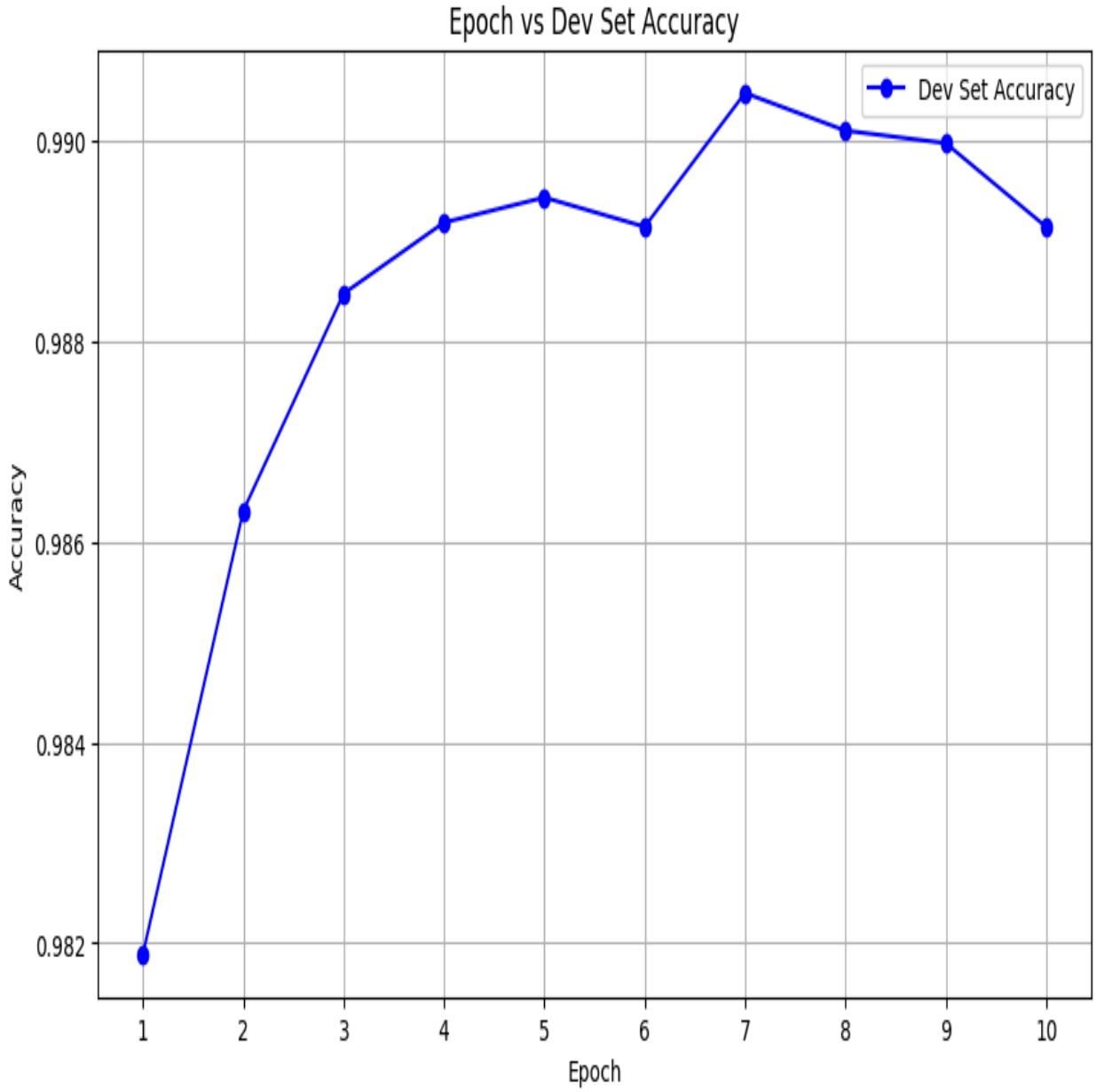
Analysis:

- The confusion matrix's diagonal elements are predominantly high, indicating that the model makes correct predictions most of the time for each POS tag.
- There are very few off-diagonal elements with non-zero values, indicating minimal misclassifications compared to the first model, signifying an improvement.
- Certain POS tags like NOUN, PROPN, VERB, AUX, and DET are predicted with high accuracy. However, correct predictions for NOUN and PROPN have slightly decreased, while VERB, AUX, and DET have seen an increase.
- Misclassifications between ADJ and NOUN, and between VERB and AUX appear to be reduced in this model, indicating better learning distinction between these POS tags.
- The graph illustrates a steady increase in accuracy, plateauing around epoch 5. Unlike the first model, this model continues to improve slightly until epoch 7 before showing some fluctuation.
- Minor fluctuations in accuracy suggest the model is not drastically changing its parameters after a certain point, indicating potential convergence.

Model 3: There are 3 LSTM layers.the size of the hidden dimension is 256,the activation function used is elu,bidirectionality is also used here.

```
Accuracy: 0.9899
Macro Recall: 0.8838
Macro F1: 0.8800
Micro Recall: 0.9899
Micro F1: 0.9899
```





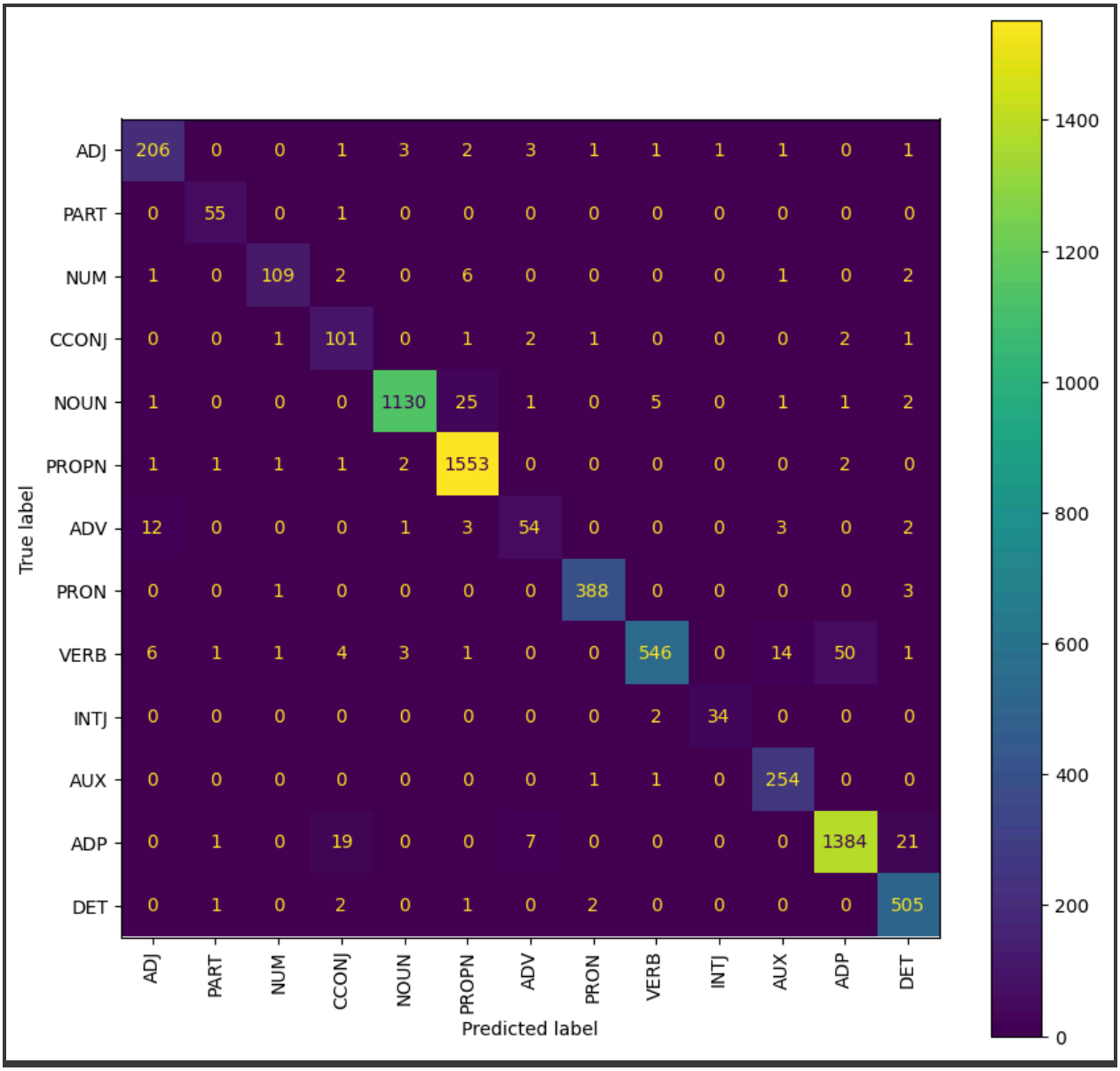
Analysis:

- The confusion matrix for Model 3 exhibits high values along the diagonal, indicating correct predictions for the majority of POS tags.
- Particularly high numbers on the diagonal for common tags like NOUN, PROPN, VERB, AUX, and ADJ highlight strong predictive performance for these tags.
- There's a notable misclassification between ADJ (adjective) and NOUN, likely due to the flexibility of English grammar where adjectives can sometimes function as nouns.

- The model also demonstrates confusion between PART (particle) and other tags, albeit at low frequencies, suggesting potential areas for improvement with additional context or training data.
- However, confusion between VERB and NOUN is relatively low, indicating the model's good understanding of the verb-noun distinction, crucial in POS tagging.
- Accuracy on the development set steadily increases over epochs, with a significant jump between epochs 1 and 4, suggesting effective learning from the training data.
- Post epoch 4, accuracy levels out with minor fluctuations, indicating near optimization of parameters for the current dataset, with little to no improvement from additional training.
- Model 3 achieves the highest accuracy among the discussed models, indicating positive impacts from changes made to its architecture or training process.
- The model demonstrates relatively stable and high accuracy across epochs, suggesting a robust understanding of underlying data patterns and good generalization to unseen data.

Best Model : The Best model is model 3 so here is the best model ‘s test set evaluation metrics

`Accuracy: 0.9864`
`Macro Recall: 0.9371`
`Macro F1: 0.9356`
`Micro Recall: 0.9864`
`Micro F1: 0.9864`



Analysis:

The high values along the diagonal indicate strong predictive capability for each POS tag, with most tags being predicted correctly most of the time.

The model performs extremely well with NOUN, PROPN, VERB, and AUX, as indicated by the very high correct predictions, seen from the dark squares on the diagonal.

Confusion exists between:

- ADJ and NOUN, a common area of confusion due to contextual dependencies.
- VERB and AUX, which may occur when verbs function both as helping verbs and as main verbs in sentences.
- ADJ and ADV, possibly due to adjectives being used adverbially or comparative forms of adjectives being mistaken for adverbs.
- PART, with misclassifications as ADP and CONJ, likely due to the multifunctional nature of particles in English grammar.

Accuracy on the test set is very high, suggesting the model generalizes well to unseen data and is robust across various sentence structures and vocabularies.

A high micro-average recall indicates the model is recovering almost all instances of each POS tag from the test set.

A high macro-average recall suggests consistent performance across different POS tags, without bias towards more frequent tags.

High F1 scores (both micro and macro) indicate a good balance between precision and recall, signifying accuracy and comprehensiveness in POS tagging.