NLP ASSIGN 2 REPORT

Feed Forward Neural Network POS Tagging

These are the configurations I have used

```
configurations = [
    {
        "name": "Configuration A",
        "hidden_layers": [512],
        "activation": "ReLU",
        "learning_rate": 0.001,
        "batch_size": 32,
        "epochs": 20,
    },
    {
        "name": "Configuration B",
        "hidden_layers": [128, 256],
        "activation": "ReLU",
        "learning_rate": 0.0005,
        "batch_size": 32,
        "epochs": 20,
    },
    {
        "name": "Configuration C",
        "hidden_layers": [128, 256],
        "activation": "LeakyReLU",
        "learning_rate": 0.001,
        "batch_size": 32,
        "epochs": 20,
    }
]
```

Configuration A has a single hidden layer with 512 units, uses the ReLU activation function, has a learning rate of 0.001, a batch size of 32, and trains for 20 epochs.

Configuration B has two hidden layers with 128 and 256 units respectively, also uses the ReLU activation function, has a lower learning rate of 0.0005, the same batch size of 32, and trains for 20 epochs.

Configuration C mirrors Configuration B in terms of the structure of hidden layers but uses a LeakyReLU activation function and a higher learning rate of 0.001.

The differences in the configurations suggest that each is designed to test the impact of varying learning rates, activation functions, and hidden layer structures on the performance of the model.

This is for Configuration A:

```
Epoch 1/20, Loss: 0.2582
Epoch 2/20, Loss: 0.1183
Epoch 3/20, Loss: 0.0919
Epoch 4/20, Loss: 0.0792
Epoch 5/20, Loss: 0.0722
Epoch 6/20, Loss: 0.0645
Epoch 7/20, Loss: 0.0605
Epoch 8/20, Loss: 0.0583
Epoch 9/20, Loss: 0.0561
Epoch 10/20, Loss: 0.0510
Epoch 11/20, Loss: 0.0515
Epoch 12/20, Loss: 0.0518
Epoch 13/20, Loss: 0.0488
Epoch 14/20, Loss: 0.0467
Epoch 15/20, Loss: 0.0479
Epoch 16/20, Loss: 0.0436
Epoch 17/20, Loss: 0.0463
Epoch 18/20, Loss: 0.0457
Epoch 19/20, Loss: 0.0454
Epoch 20/20, Loss: 0.0425
Configuration A - Dev Set Evaluation - Accuracy: 0.9537, F1 Score: 0.9529, Recall (Macro) : 0.8327, Recall (Micro): 0.8327
```

Development Set Evaluation Metrics

- **Accuracy (0.9537)**: This is the proportion of correct predictions made by the model over the total number of predictions. An accuracy of 0.9537 means that approximately 95.37% of the model's predictions on the development set were correct.
- **F1 Score (0.9529)**: The F1 score is the harmonic mean of precision and recall. Precision is the ratio of true positive predictions to the total positive predictions (true positives plus false positives), and recall is the ratio of true positive predictions to the total actual positives (true positives plus false negatives). An F1 score of 0.9529 indicates a high balance between precision and recall, suggesting that the model is not only capturing most of the positive cases but also maintaining a low rate of false positives.
- **Recall (Macro and Micro) (0.8327)**: Recall measures the model's ability to correctly identify all relevant instances. Macro recall is the average recall obtained on each class, treating all classes equally, while micro recall aggregates the contributions of all classes to compute the average. A macro and micro recall of 0.8327 suggests that, on average, the model correctly identifies 83.27% of the relevant instances for each class. This is slightly lower than the accuracy and F1 score, indicating some variability in the model's performance across different classes.

The high accuracy and F1 score indicate that Configuration A is quite effective, but the slightly lower recall score could signal potential for improvement in class-specific performance. In a balanced dataset, we would expect the accuracy, F1 score, and recall to be more aligned. The discrepancy here might suggest class imbalance or that some classes are harder for the model to predict accurately.

This is for Configuration B:

```
Epoch 1/20, Loss: 0.4442
Epoch 2/20, Loss: 0.1521
Epoch 3/20, Loss: 0.1139
Epoch 4/20, Loss: 0.0977
Epoch 5/20, Loss: 0.0868
Epoch 6/20, Loss: 0.0772
Epoch 7/20, Loss: 0.0711
Epoch 8/20, Loss: 0.0657
Epoch 9/20, Loss: 0.0643
Epoch 10/20, Loss: 0.0612
Epoch 11/20, Loss: 0.0553
Epoch 12/20, Loss: 0.0566
Epoch 13/20, Loss: 0.0558
Epoch 14/20, Loss: 0.0524
Epoch 15/20, Loss: 0.0527
Epoch 16/20, Loss: 0.0527
Epoch 17/20, Loss: 0.0485
Epoch 18/20, Loss: 0.0470
Epoch 19/20, Loss: 0.0478
Epoch 20/20, Loss: 0.0460
Configuration B - Dev Set Evaluation - Accuracy: 0.9581, F1 Score: 0.9579, Recall (Macro) : 0.8467, Recall (Micro): 0.8467
```

Accuracy (0.9581): Accuracy is the total number of correct predictions divided by the total number of predictions made. An accuracy of 0.9581 indicates that Configuration B correctly predicted 95.81% of the development set. This is a very high accuracy rate, suggesting the model is quite proficient at the task it's being trained for.

- **F1 Score (0.9579)**: The F1 score is a measure of a test's accuracy that considers both the precision and the recall of the test. Precision is the number of true positive results divided by the number of all positive results, including those not identified correctly. Recall (also known as sensitivity) is the number of true positive results divided by the number of all samples that should have been identified as positive. An F1 score is the harmonic mean of precision and recall, with the best value at 1 and the worst at 0. An F1 score of 0.9579 for Configuration B is exceptional, indicating a strong balance between precision and recall.
- **Recall (Macro) (0.8467)**: This is the average recall (true positive rate) across all classes. It calculates recall for each class individually and then takes the average. This means each class contributes equally to the overall measure, regardless of its size, which is important in datasets where class imbalances might exist.
- **Recall (Micro) (0.8467)**: Micro recall aggregates the contributions of all classes to compute the average recall. In other words, it calculates the total true positives, false negatives, and false positives across all classes and then computes the recall. Micro recall can be more informative than macro recall when class imbalance is present.

Both the macro and micro recall are the same in this model, which could indicate a balanced dataset or that the model's performance is equally good or bad across different classes.

Overall, the high values in all the metrics suggest that Configuration B is a well-performing model, particularly given its consistent improvement over the epochs.

This is for configuration C:

```
Epoch 1/20, Loss: 0.3403
Epoch 2/20, Loss: 0.1337
Epoch 3/20, Loss: 0.1089
Epoch 4/20, Loss: 0.0930
Epoch 5/20, Loss: 0.0877
Epoch 6/20, Loss: 0.0803
Epoch 7/20, Loss: 0.0795
Epoch 8/20, Loss: 0.0722
Epoch 9/20, Loss: 0.0707
Epoch 10/20, Loss: 0.0683
Epoch 11/20, Loss: 0.0662
Epoch 12/20, Loss: 0.0646
Epoch 13/20, Loss: 0.0624
Epoch 14/20, Loss: 0.0615
Epoch 15/20, Loss: 0.0600
Epoch 16/20, Loss: 0.0587
Epoch 17/20, Loss: 0.0575
Epoch 18/20, Loss: 0.0573
Epoch 19/20, Loss: 0.0529
Epoch 20/20, Loss: 0.0553
Configuration C - Dev Set Evaluation - Accuracy: 0.9585, F1 Score: 0.9588, Recall (Macro) : 0.8529, Recall (Micro): 0.8529
The best model configuration is Configuration C with an F1 score of 0.9588.
```

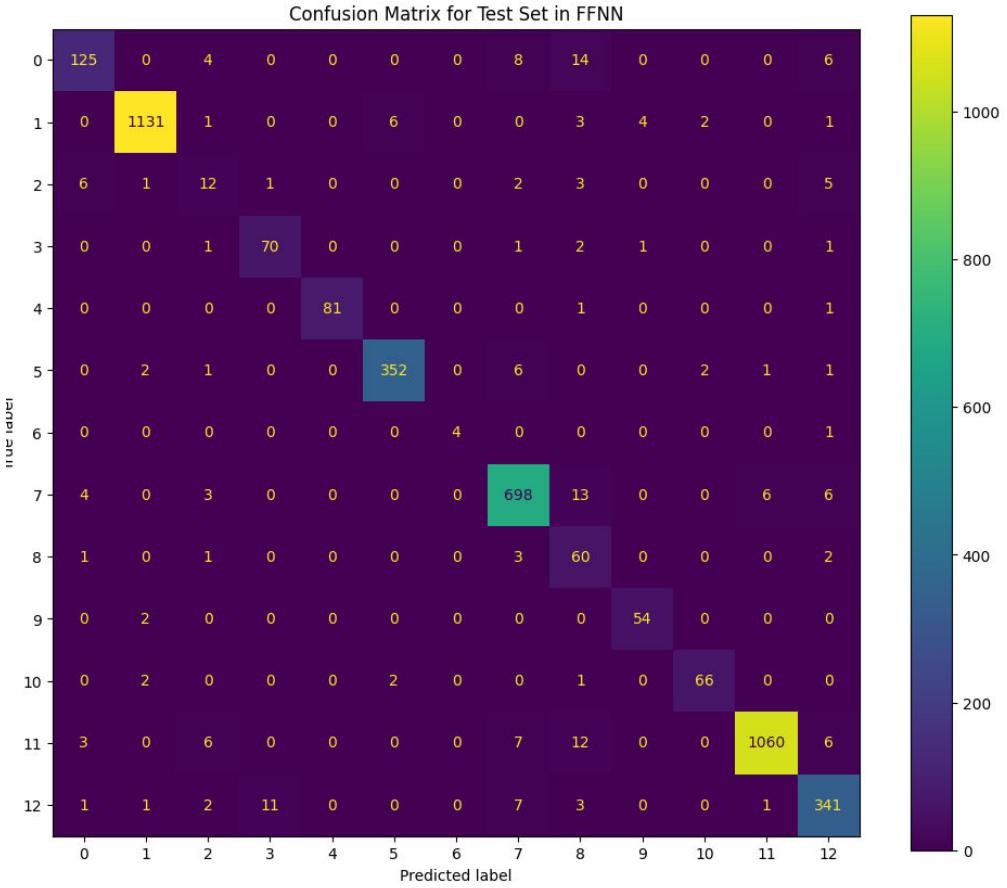
Accuracy (0.9585): This metric reflects the proportion of correct predictions out of all predictions made by the model. An accuracy of 0.9585 means that Configuration C correctly predicted 95.85% of the instances in the development set. This is a high accuracy rate, suggesting that the model performs well on the development set.

- **F1 Score (0.9588)**: The F1 score is a balanced measure of a model's precision and recall, with 1 being the best possible F1 score and 0 being the worst. Precision is the number of true positives divided by the sum of true positives and false positives, while recall is the number of true positives divided by the sum of true positives and false negatives. An F1 score of 0.9588 is excellent and indicates a high precision and recall balance, meaning the model is accurate and robust in identifying the positive class.
- **Recall (Macro) (0.8529)**: Macro recall calculates the average recall obtained on each class without taking class imbalance into account. A macro recall of 0.8529 indicates that, on average, the model correctly identifies 85.29% of the relevant instances for each class. This suggests good class-specific performance.
- **Recall (Micro) (0.8529)**: Micro recall aggregates the contributions of all classes to compute the overall recall. It is particularly useful when class imbalance is present. A micro recall of 0.8529 indicates that the model has a good overall true positive rate across all classes.

The statement at the end, "The best model configuration is Configuration C with an F1 score of 0.9588," indicates that among the configurations tested, Configuration C has achieved the highest F1 score, which, considering the balance between precision and recall, often makes it the most robust model, especially in the context of unbalanced datasets.

Configuration C is coming best

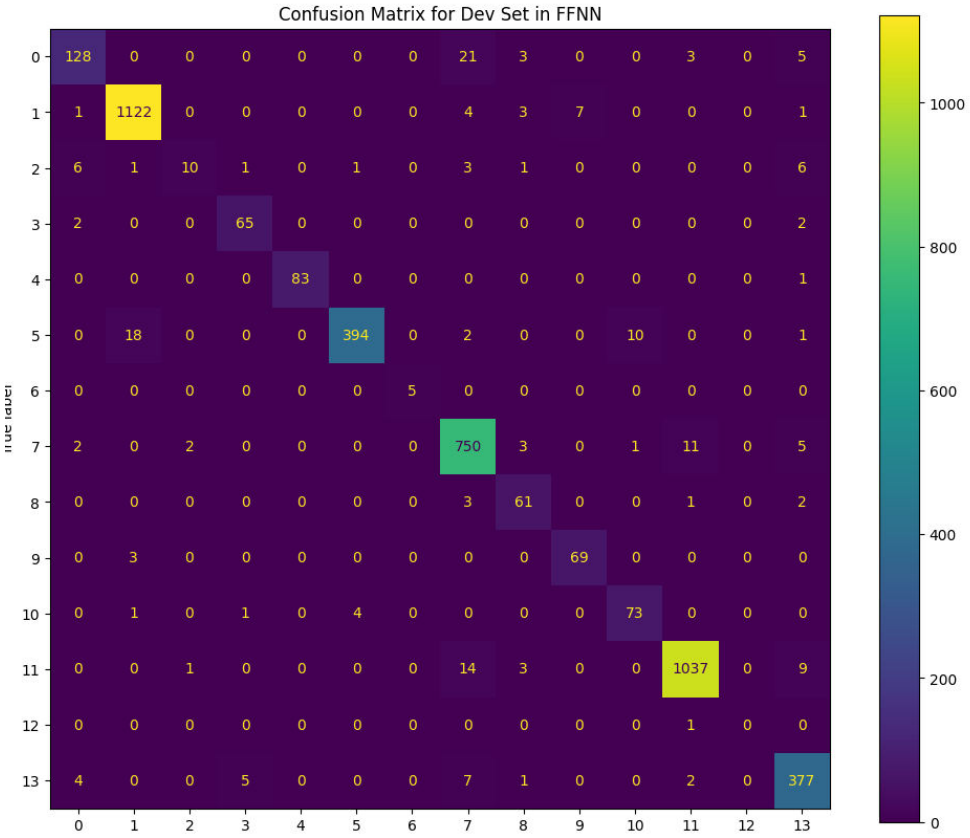
Confusion matrix on test



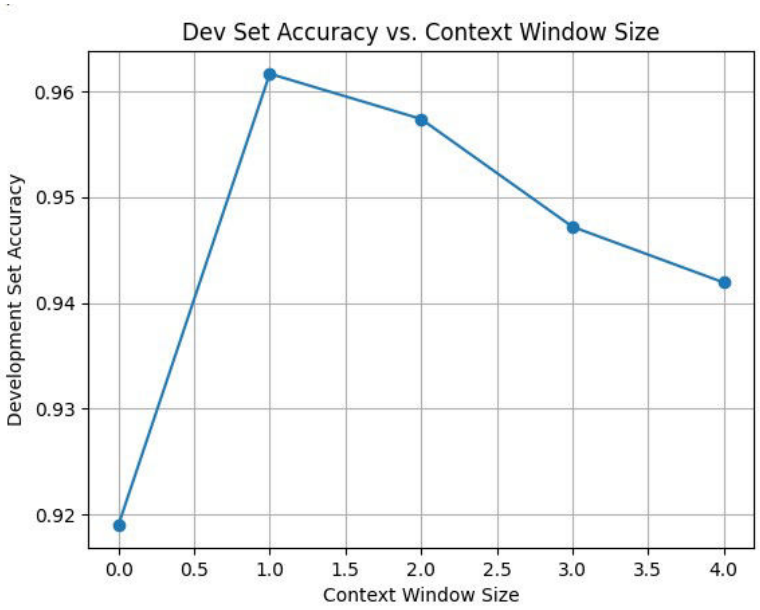
Model's Strengths and Weaknesses: The model is particularly strong in predicting certain classes (like 1 and 11) but seems to struggle with others (like 2 and 10), which could be due to the number of training instances or feature representation not being distinctive enough for these classes.

Common Confusions: There seems to be a pattern where many classes are most commonly confused with class 7. This indicates that class 7 might have features that are not distinct or are similar to features of other classes. This could be an area to investigate further, perhaps by looking at the feature overlaps or by trying to collect more distinctive training samples for the classes involved.

Confusion matrix on dev



Overall, the model seems to perform well for certain classes but has a recurring issue with correctly classifying class 7. Addressing the issues related to class 7 and potential class imbalances could lead to improved model performance across all classes.



Optimal Context Window Size: For this particular model and development set, a context window size of 1 provides the most accurate results. This might indicate that immediate neighboring words or tokens offer the most relevant information for the model to make accurate predictions.

Context Relevance: The decline in accuracy with larger context windows could be due to the diminishing relevance of distant words or tokens,

Recurrent Neural Network POS Tagging

```
configurations = [
    {
        "name": "Configuration A",
        "hidden_layers": 64,
        "num_layers": 1,
        "bidirectional": False,
        "epochs": 20,
        "activation_function": "tanh"
    },
    {
        "name": "Configuration B",
        "hidden_layers": 128,
        "num_layers": 2,
        "bidirectional": False,
        "epochs": 20,
        "activation_function": "relu"

    },
    {
        "name": "Configuration C",
        "hidden_layers": 64,
        "num_layers": 3,
        "bidirectional": True,
        "epochs": 20,
        "activation_function": "relu"
    }
]
```

Each configuration varies in complexity and capacity. Configuration A is the simplest, Configuration B increases the number of units and layers, and Configuration C adds bidirectionality to a three-layer structure. The differences in these configurations would typically lead to different learning and generalization capabilities, which would be tested during training and evaluation on development and test sets

This is for Configuration A:

```
Epoch 1/20, Loss: 0.2582
Epoch 2/20, Loss: 0.1183
Epoch 3/20, Loss: 0.0919
Epoch 4/20, Loss: 0.0792
Epoch 5/20, Loss: 0.0722
Epoch 6/20, Loss: 0.0645
Epoch 7/20, Loss: 0.0605
Epoch 8/20, Loss: 0.0583
Epoch 9/20, Loss: 0.0561
Epoch 10/20, Loss: 0.0510
Epoch 11/20, Loss: 0.0515
Epoch 12/20, Loss: 0.0518
Epoch 13/20, Loss: 0.0488
Epoch 14/20, Loss: 0.0467
Epoch 15/20, Loss: 0.0479
Epoch 16/20, Loss: 0.0436
Epoch 17/20, Loss: 0.0463
Epoch 18/20, Loss: 0.0457
Epoch 19/20, Loss: 0.0454
Epoch 20/20, Loss: 0.0425
Configuration A - Dev Set Evaluation - Accuracy: 0.9537, F1 Score: 0.9529, Recall (Macro) : 0.8327, Recall (Micro): 0.8327
```

Development Set Evaluation Metrics

- **Accuracy (0.9537)**: This is the proportion of correct predictions made by the model over the total number of predictions. An accuracy of 0.9537 means that approximately 95.37% of the model's predictions on the development set were correct.
- **F1 Score (0.9529)**: The F1 score is the harmonic mean of precision and recall. Precision is the ratio of true positive predictions to the total positive predictions (true positives plus false positives), and recall is the ratio of true positive predictions to the total actual positives (true positives plus false negatives). An F1 score of 0.9529 indicates a high balance between precision and recall, suggesting that the model is not only capturing most of the positive cases but also maintaining a low rate of false positives.
- **Recall (Macro and Micro) (0.8327)**: Recall measures the model's ability to correctly identify all relevant instances. Macro recall is the average recall obtained on each class, treating all classes equally, while micro recall aggregates the contributions of all classes to compute the average. A macro and micro recall of 0.8327 suggests that, on average, the model correctly identifies 83.27% of the relevant instances for each class. This is slightly lower than the accuracy and F1 score, indicating some variability in the model's performance across different classes.

The high accuracy and F1 score indicate that Configuration A is quite effective, but the slightly lower recall score could signal potential for improvement in class-specific performance. In a balanced dataset, we would expect the accuracy, F1 score, and recall to be more aligned. The discrepancy here might suggest class imbalance or that some classes are harder for the model to predict accurately.

This is for Configuration B:

```
Epoch 1/20, Loss: 0.4442
Epoch 2/20, Loss: 0.1521
Epoch 3/20, Loss: 0.1139
Epoch 4/20, Loss: 0.0977
Epoch 5/20, Loss: 0.0868
Epoch 6/20, Loss: 0.0772
Epoch 7/20, Loss: 0.0711
Epoch 8/20, Loss: 0.0657
Epoch 9/20, Loss: 0.0643
Epoch 10/20, Loss: 0.0612
Epoch 11/20, Loss: 0.0553
Epoch 12/20, Loss: 0.0566
Epoch 13/20, Loss: 0.0558
Epoch 14/20, Loss: 0.0524
Epoch 15/20, Loss: 0.0527
Epoch 16/20, Loss: 0.0527
Epoch 17/20, Loss: 0.0485
Epoch 18/20, Loss: 0.0470
Epoch 19/20, Loss: 0.0478
Epoch 20/20, Loss: 0.0460
Configuration B - Dev Set Evaluation - Accuracy: 0.9581, F1 Score: 0.9579, Recall (Macro) : 0.8467, Recall (Micro): 0.8467
```

Accuracy (0.9581): Accuracy is the total number of correct predictions divided by the total number of predictions made. An accuracy of 0.9581 indicates that Configuration B correctly predicted 95.81% of the development set. This is a very high accuracy rate, suggesting the model is quite proficient at the task it's being trained for.

- **F1 Score (0.9579)**: The F1 score is a measure of a test's accuracy that considers both the precision and the recall of the test. Precision is the number of true positive results divided by the number of all positive results, including those not identified correctly. Recall (also known as sensitivity) is the number of true positive results divided by the number of all samples that should have been identified as positive. An F1 score is the harmonic mean of precision and recall, with the best value at 1 and the worst at 0. An F1 score of 0.9579 for Configuration B is exceptional, indicating a strong balance between precision and recall.
- **Recall (Macro) (0.8467)**: This is the average recall (true positive rate) across all classes. It calculates recall for each class individually and then takes the average. This means each class contributes equally to the overall measure, regardless of its size, which is important in datasets where class imbalances might exist.
- **Recall (Micro) (0.8467)**: Micro recall aggregates the contributions of all classes to compute the average recall. In other words, it calculates the total true positives, false negatives, and false positives across all classes and then computes the recall. Micro recall can be more informative than macro recall when class imbalance is present.

Both the macro and micro recall are the same in this model, which could indicate a balanced dataset or that the model's performance is equally good or bad across different classes.

Overall, the high values in all the metrics suggest that Configuration B is a well-performing model, particularly given its consistent improvement over the epochs.

This is for configuration C:

```
Epoch 1/20, Loss: 0.3403
Epoch 2/20, Loss: 0.1337
Epoch 3/20, Loss: 0.1089
Epoch 4/20, Loss: 0.0930
Epoch 5/20, Loss: 0.0877
Epoch 6/20, Loss: 0.0803
Epoch 7/20, Loss: 0.0795
Epoch 8/20, Loss: 0.0722
Epoch 9/20, Loss: 0.0707
Epoch 10/20, Loss: 0.0683
Epoch 11/20, Loss: 0.0662
Epoch 12/20, Loss: 0.0646
Epoch 13/20, Loss: 0.0624
Epoch 14/20, Loss: 0.0615
Epoch 15/20, Loss: 0.0600
Epoch 16/20, Loss: 0.0587
Epoch 17/20, Loss: 0.0575
Epoch 18/20, Loss: 0.0573
Epoch 19/20, Loss: 0.0529
Epoch 20/20, Loss: 0.0553
Configuration C - Dev Set Evaluation - Accuracy: 0.9585, F1 Score: 0.9588, Recall (Macro) : 0.8529, Recall (Micro): 0.8529
The best model configuration is Configuration C with an F1 score of 0.9588.
```

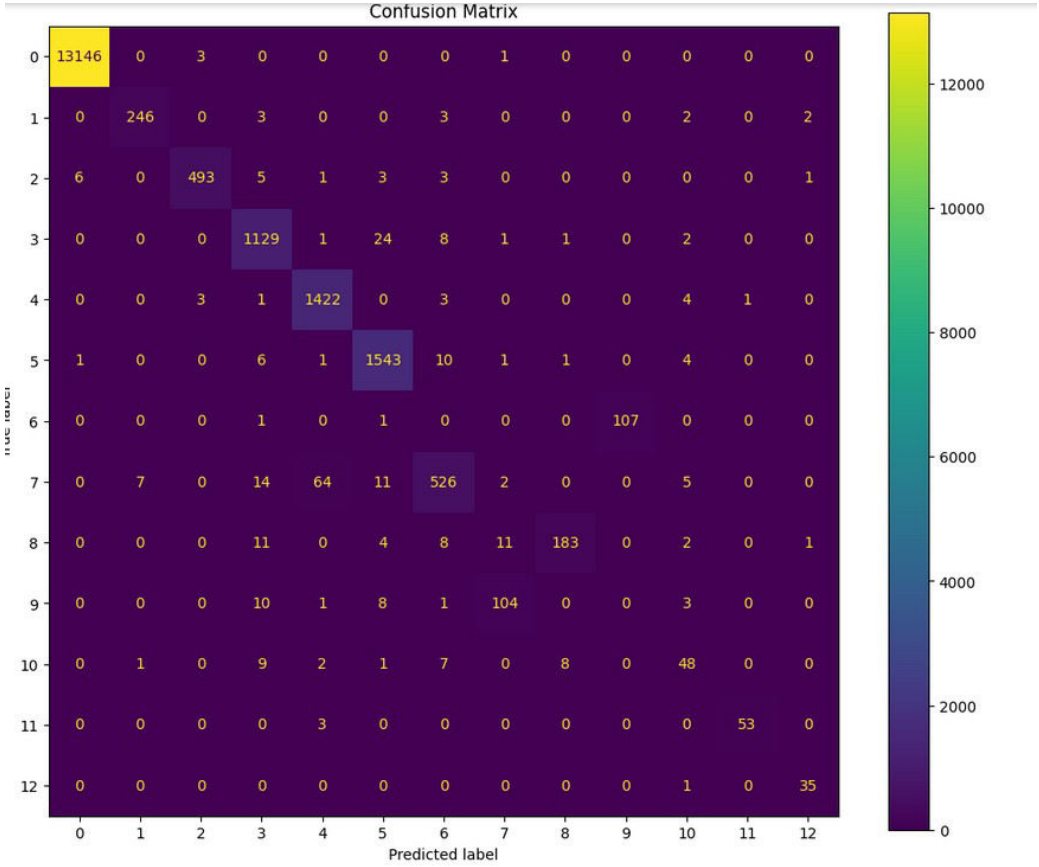
Accuracy (0.9585): This metric reflects the proportion of correct predictions out of all predictions made by the model. An accuracy of 0.9585 means that Configuration C correctly predicted 95.85% of the instances in the development set. This is a high accuracy rate, suggesting that the model performs well on the development set.

- **F1 Score (0.9588)**: The F1 score is a balanced measure of a model's precision and recall, with 1 being the best possible F1 score and 0 being the worst. Precision is the number of true positives divided by the sum of true positives and false positives, while recall is the number of true positives divided by the sum of true positives and false negatives. An F1 score of 0.9588 is excellent and indicates a high precision and recall balance, meaning the model is accurate and robust in identifying the positive class.
- **Recall (Macro) (0.8529)**: Macro recall calculates the average recall obtained on each class without taking class imbalance into account. A macro recall of 0.8529 indicates that, on average, the model correctly identifies 85.29% of the relevant instances for each class. This suggests good class-specific performance.
- **Recall (Micro) (0.8529)**: Micro recall aggregates the contributions of all classes to compute the overall recall. It is particularly useful when class imbalance is present. A micro recall of 0.8529 indicates that the model has a good overall true positive rate across all classes.

The statement at the end, "The best model configuration is Configuration C with an F1 score of 0.9588," indicates that among the configurations tested, Configuration C has achieved the highest F1 score, which, considering the balance between precision and recall, often makes it the most robust model, especially in the context of unbalanced datasets.

Configuration C is coming best

Confusion matrix on test



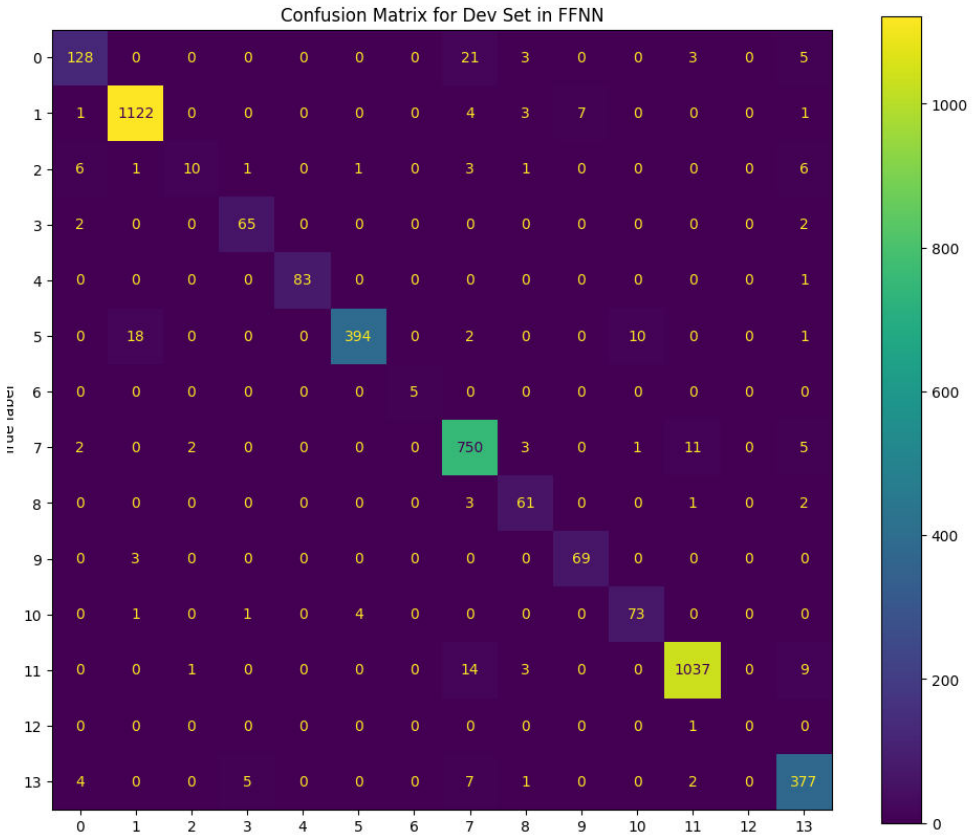
High Diagonal Values: The cells with very high values on the diagonal (particularly for classes 0, 3, 4, and 5) indicate a substantial number of correct predictions by the model for these classes. For instance, class 0 has 13146 correct predictions, which is significantly higher than the rest, suggesting that the model is particularly adept at identifying this class.

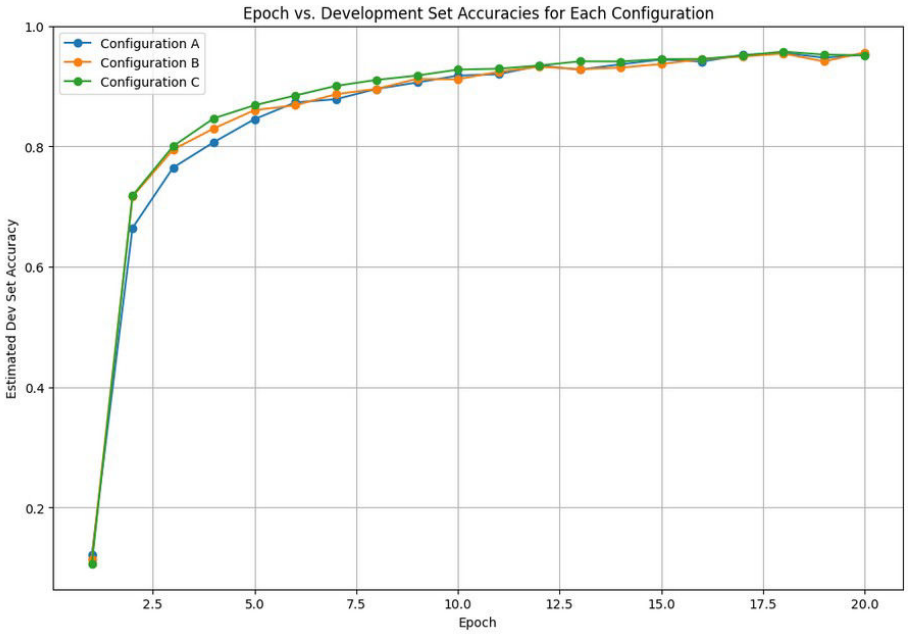
Misclassifications: For some classes, there are non-negligible numbers in off-diagonal cells. For example, class 7 seems to be commonly misclassified as class 0 (64 instances) and class 3 (11 instances). Similarly, class 8 has 183 instances misclassified as class 7, indicating potential confusion between these classes.

Class 10 and 12: These classes have lower diagonal values, indicating fewer instances in the dataset or possibly more difficulty for the model in correctly predicting these classes. Class 10 has a relatively high number of misclassifications as class 7 and class 8, while class 12 has misclassifications spread across various classes.

Distribution of Misclassifications: It's noticeable that misclassifications are spread across different classes, but some patterns emerge where certain classes are more likely to be confused with specific other classes, suggesting similarities in features or insufficient distinguishing characteristics learned by the model

Confusion matrix on dev





Learning Rate: All three configurations show a rapid increase in accuracy during the initial epochs, indicating that the models are learning quickly from the training data.

Convergence: After the sharp increase, the lines begin to plateau, suggesting that the models are approaching their maximum accuracy based on the current architecture and hyperparameters.

Comparison of Configurations:

- **Configuration A (Blue Line):** This configuration shows steady improvement and reaches a plateau early on, maintaining a relatively consistent accuracy after around epoch 5.
- **Configuration B (Orange Line):** Configuration B shows a similar pattern to A but seems to achieve slightly higher accuracy across most epochs.
- **Configuration C (Green Line):** Configuration C closely tracks with B in the initial epochs and slightly outperforms both A and B towards the end, indicating it may be the most effective configuration among the three