CSE556 NLP Assignment-2

Arnav Goel
2021519

Medha Hira
2021265

Siddharth Rajput
2021102

Amil Bhagat
2021309

1. TASK 1:

- No pre-processing was performed on the data. Custom splitting and tokenisation were performed. Please refer to the code for this.

Dataset after BIO Encoding:

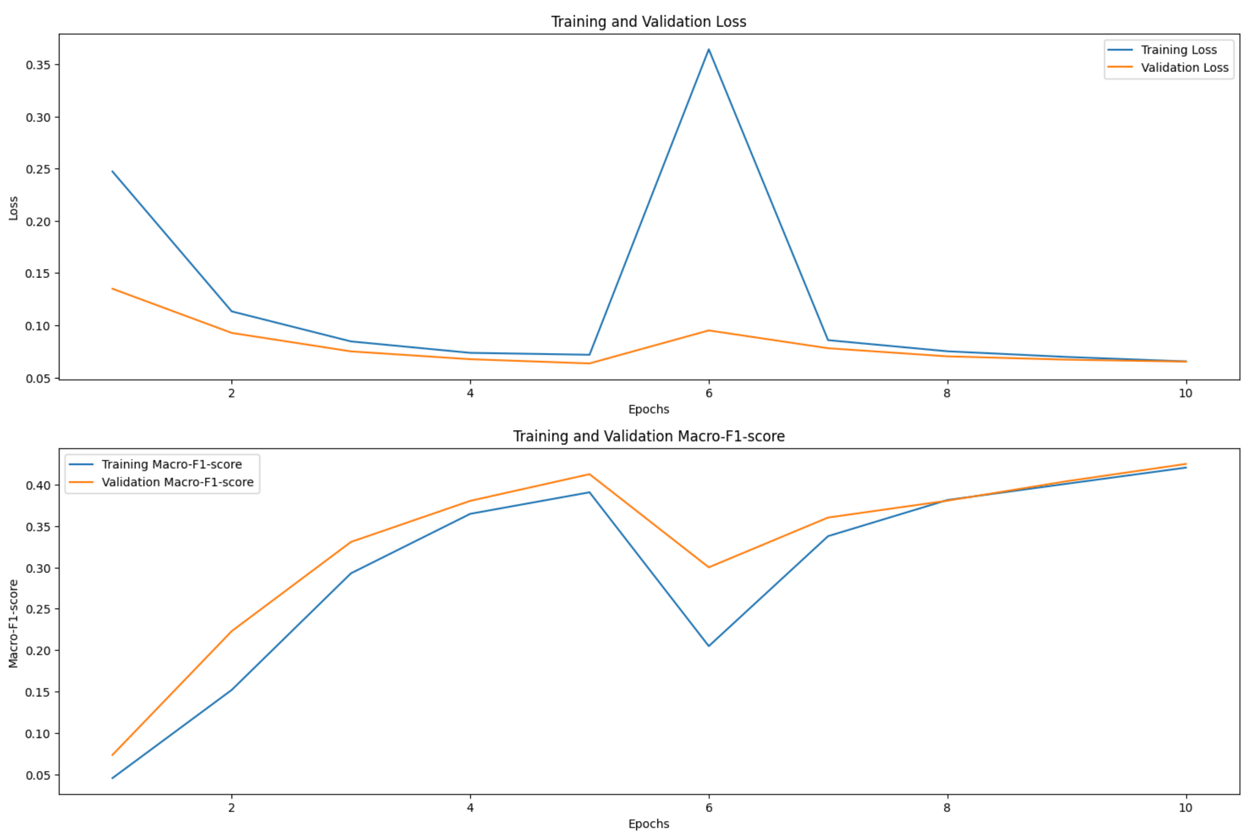
```
{
    "03f3901e95ed493b866bd7807f623bc0": {
        "text": "True, our Constitution has no 'due process' clause or the VIII Amendment;
but, in this branch of law, after R.C. Cooper v. Union of India, (1970) 1 SCC 248 and
Maneka Gandhi v. Union of India, (1978) 1 SCC 248, the consequence is the same.",
        "labels": [
            "O",
            "O",
            "B_STATUTE",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "B_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "O",
```

```
            "B_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "I_PRECEDENT",
            "O",
            "O",
            "O",
            "O",
            "O"
        ]
    }
```

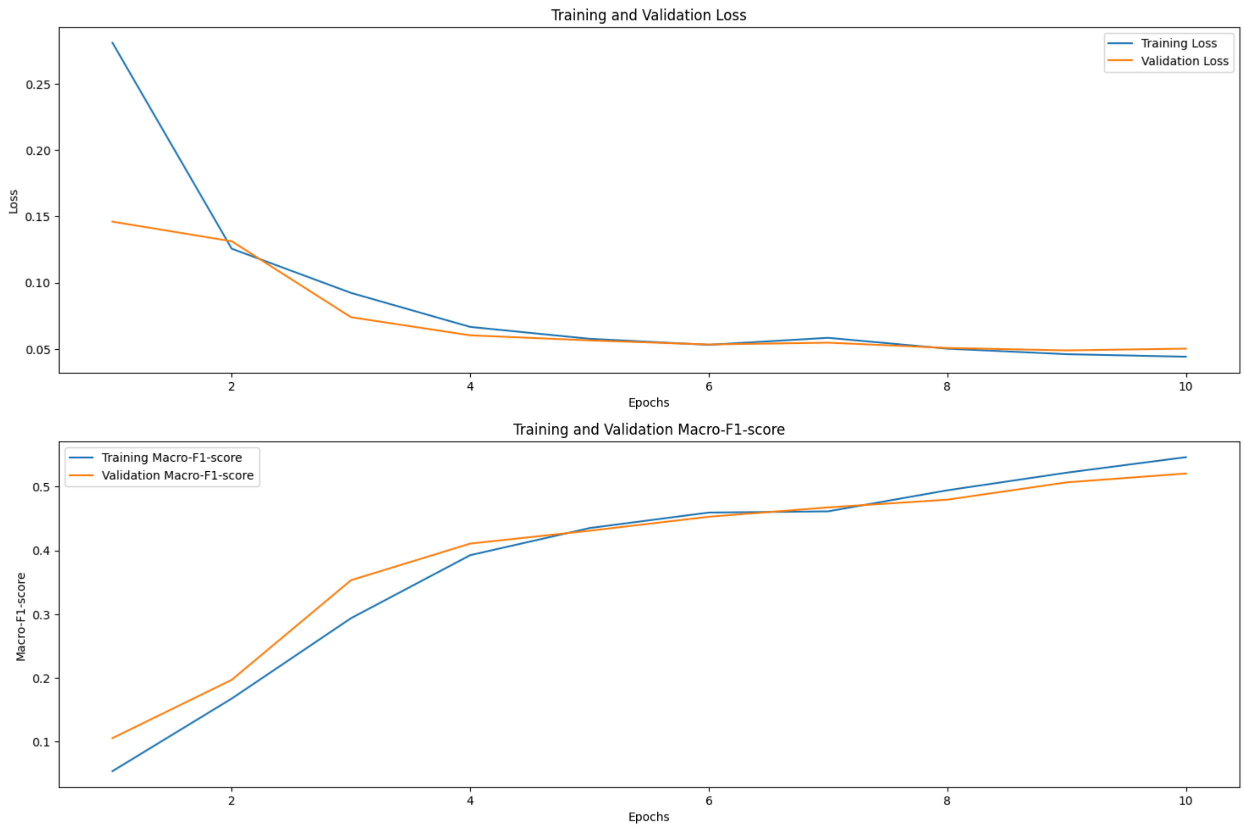
Graphs

MODEL_1(RNN)

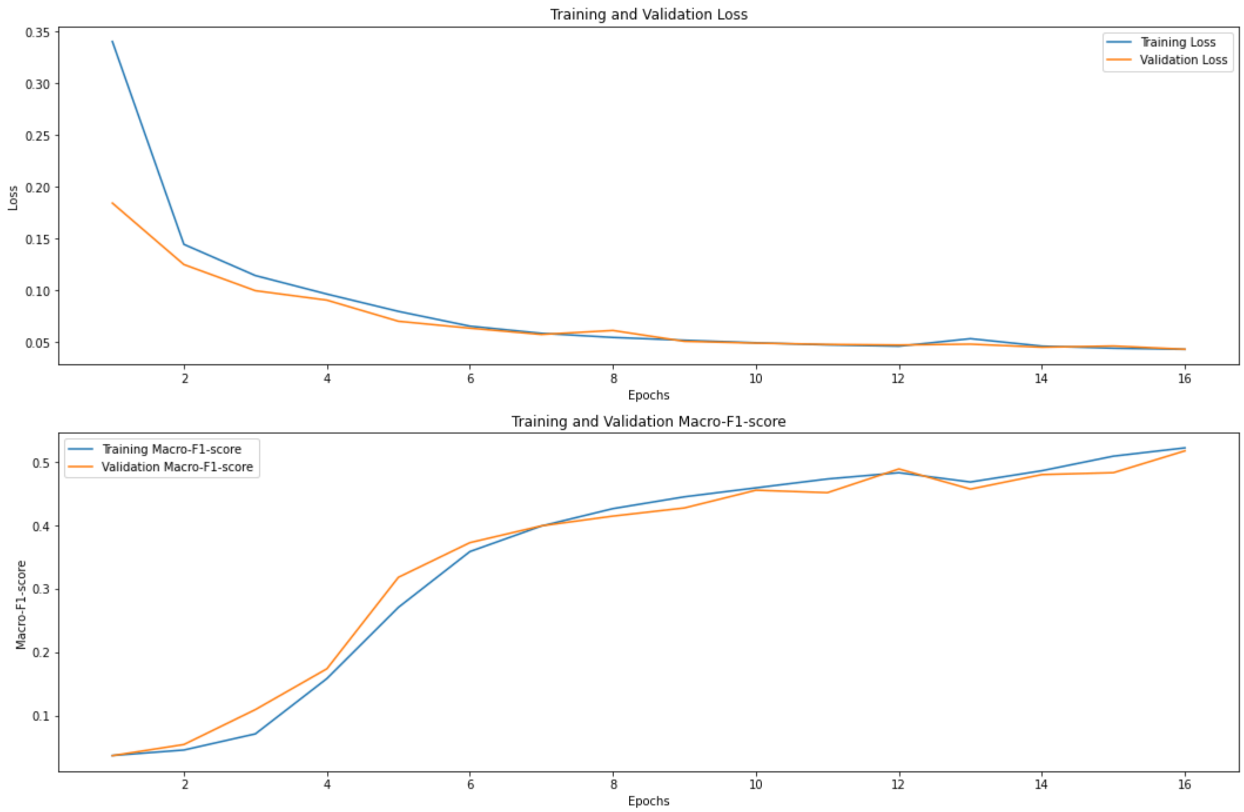
<u>Word2Vec</u>



GloVe



FastText



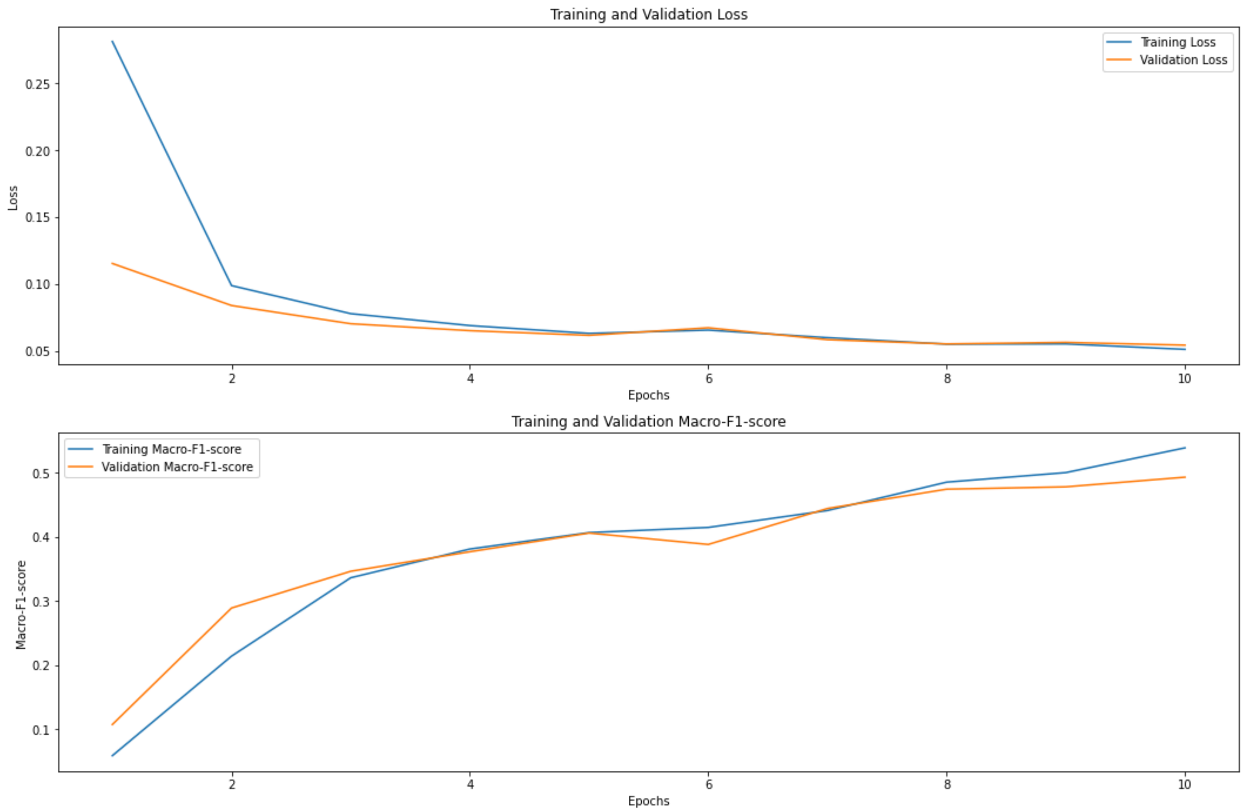
Analysis for RNN:-

- **Word2Vec Model:** Exhibits potential instability or overfitting, as seen by the significant spike in validation loss and corresponding dip in validation F1 score at epoch 6. This model might require further investigation or adjustment such as regularization or hyperparameter optimization.
- **GloVe Model:** Shows a steady decrease in both training and validation loss, suggesting a stable learning process. However, there's a consistent gap between training and validation loss, indicating a degree of overfitting. The F1 scores increase steadily without any erratic changes, which points to good generalization.
- **FastText Model:** This model presents a strong performance with a smooth decrease in loss and a close convergence between training and validation loss, implying better generalization. The F1 scores also increase steadily and surpass those of the GloVe model in later epochs, indicating that the model is effectively learning the task.

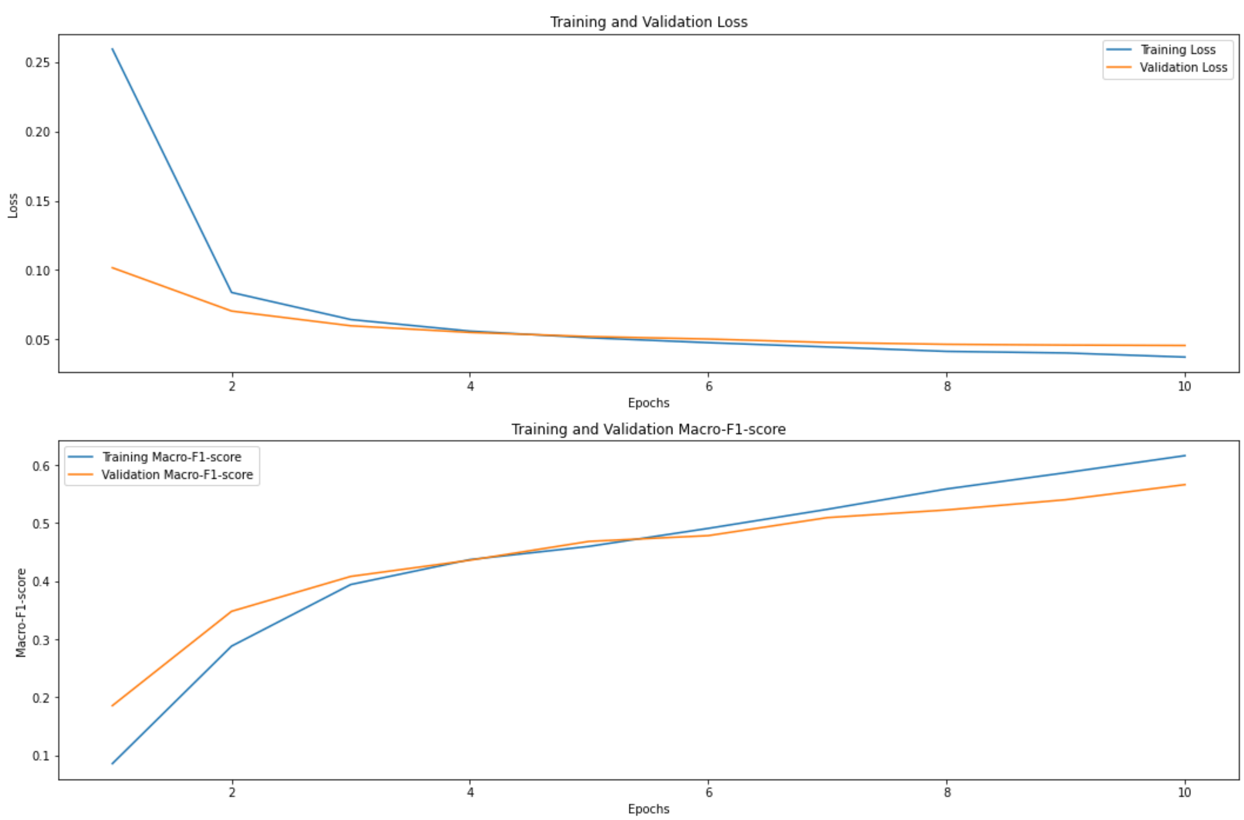
In essence, the FastText embedding seems to be the most suitable for this RNN architecture, given its steady learning and generalization. The GloVe model is stable but may be slightly overfitting, while the Word2Vec model shows signs of instability that would need to be addressed for better performance.

MODEL_2(LSTM)

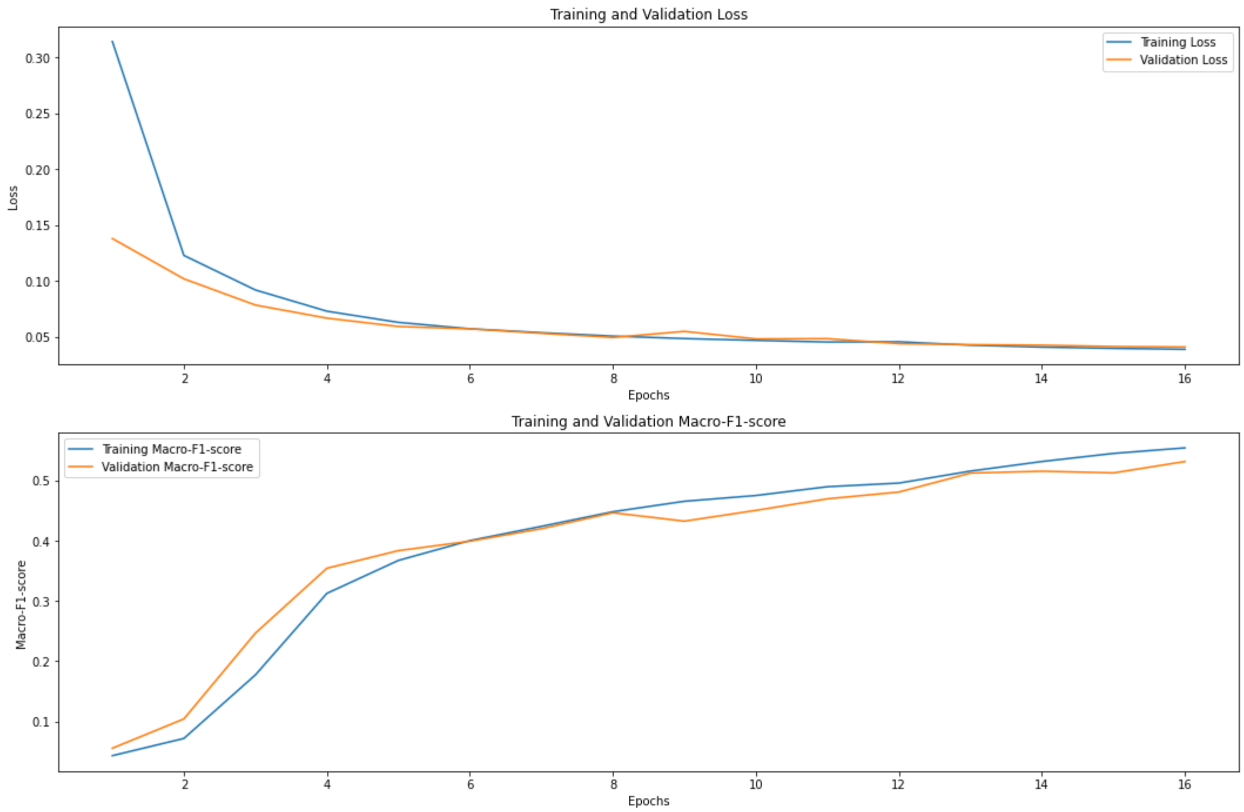
Word2Vec



GloVe



FastText



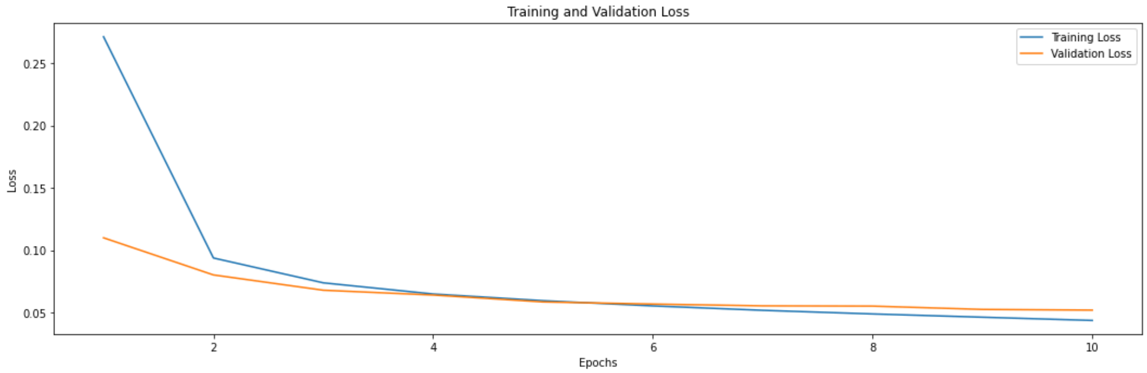
Analysis for LSTM:-

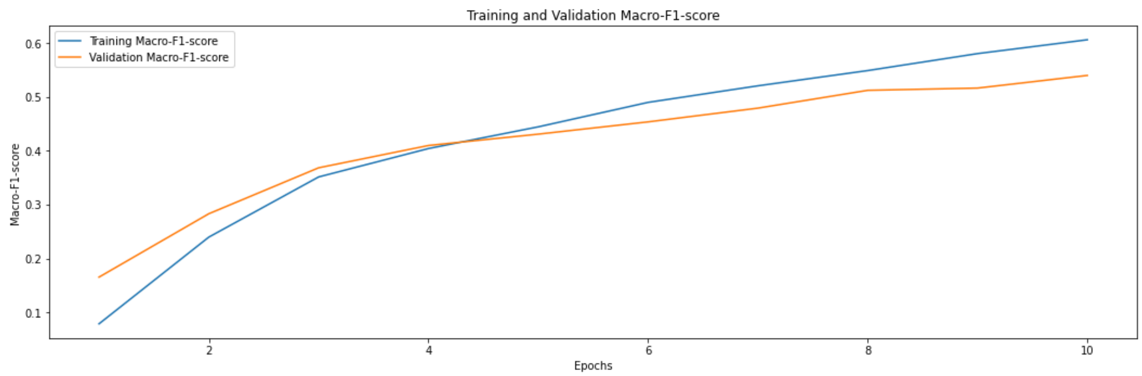
- **Word2Vec Embedding:** The LSTM model with Word2Vec shows a well-fitting trend, with the training and validation loss decreasing smoothly. The F1 scores improve consistently, suggesting that the model generalises well, though there might be a slight room for improvement compared to other embeddings.
- **GloVe Embedding:** This model demonstrates a stable and consistent learning pattern, with a very close training and validation loss, indicative of effective generalisation. The F1 scores rise steadily and maintain a close gap, suggesting that the GloVe embedding provides a slightly better representation of this LSTM model.
- **FastText Embedding:** The FastText embedding shows a good fit with a slight divergence between training and validation loss, hinting at a potential for minor overfitting. The F1 scores for both training and validation are on an upward trajectory, displaying good learning capability, although the F1 score trend is marginally lower than that of the GloVe embedding.

In summary, for the LSTM architecture, all three embeddings result in stable and promising models. Still, the GloVe embedding stands out slightly, offering the best balance between learning and generalisation capabilities as indicated by the convergence of the loss and F1 score plots.

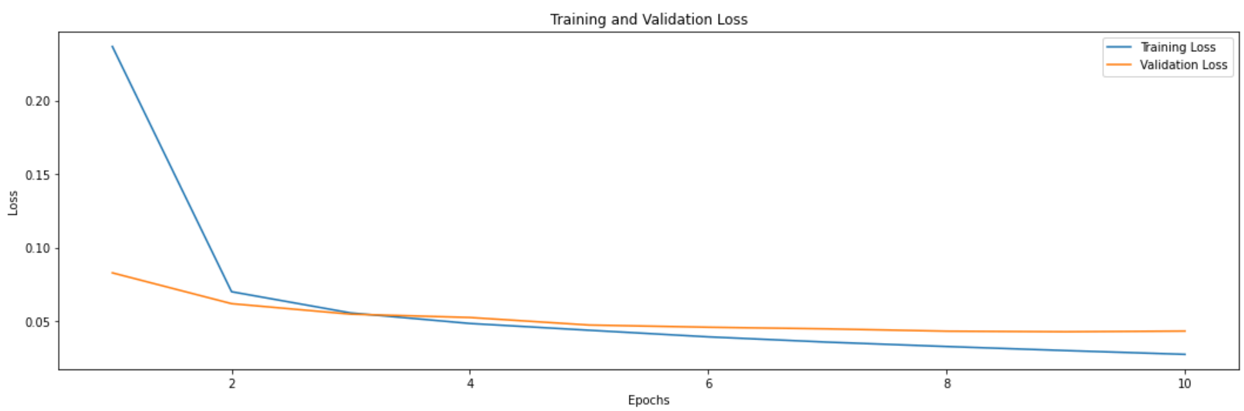
MODEL_3(GRU)

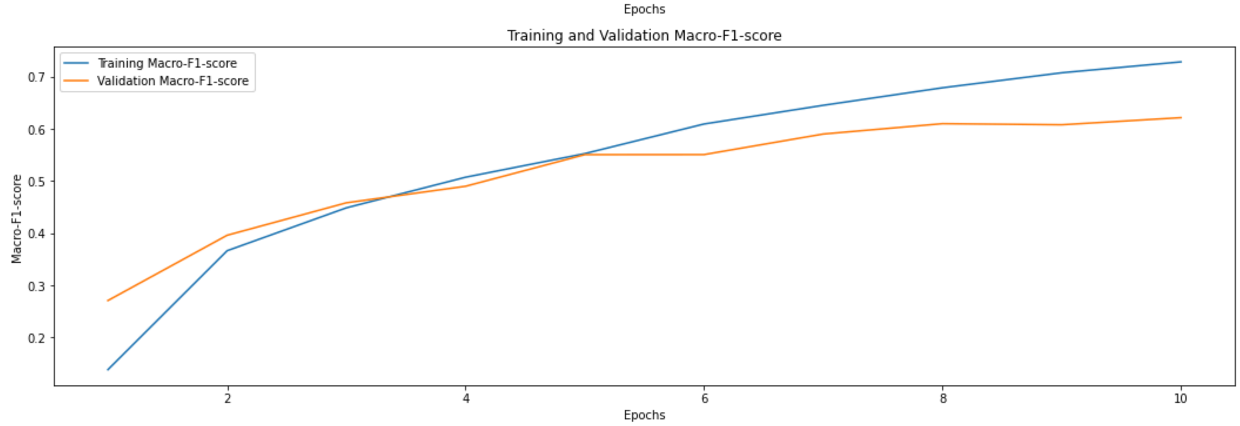
Word2Vec



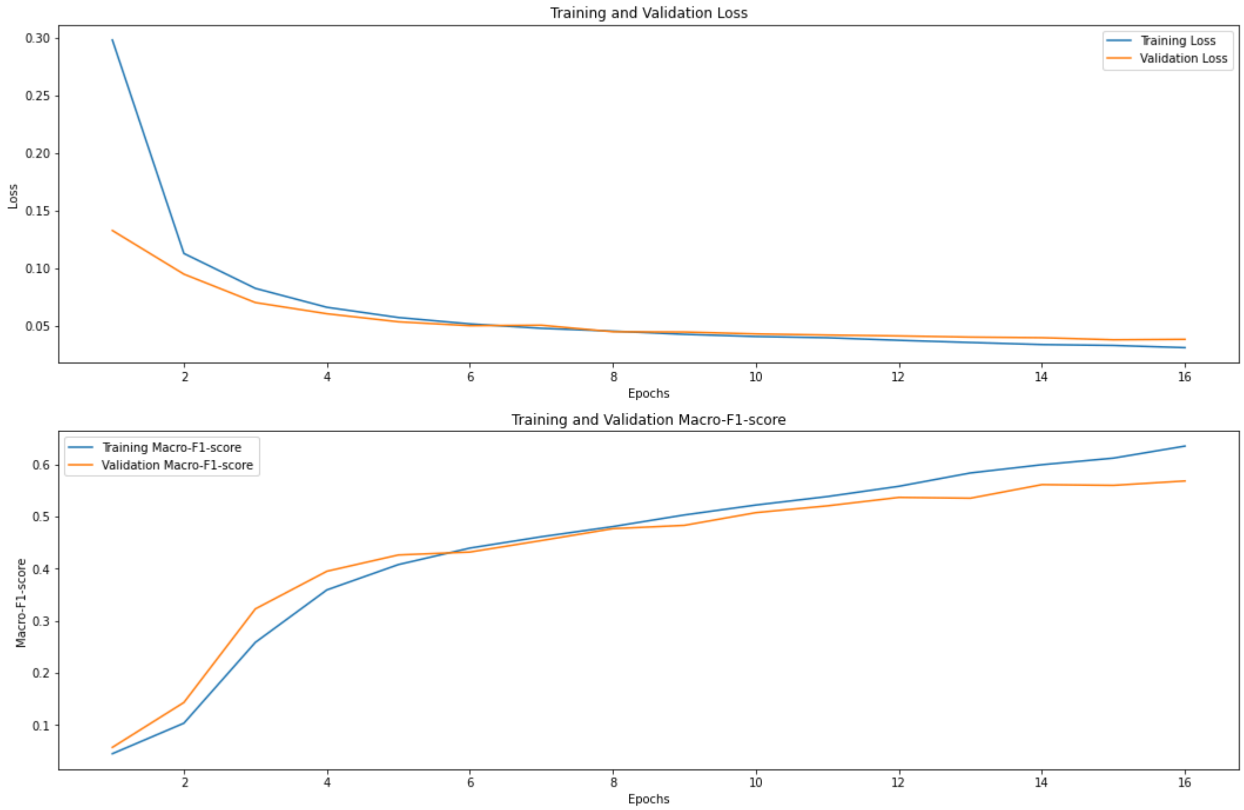


GloVe





FastText



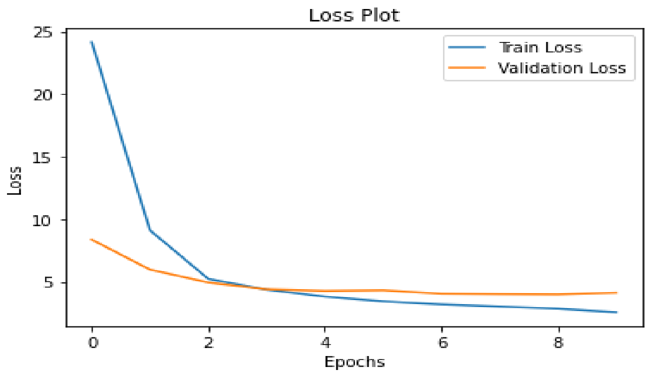
Analysis for GRU:-

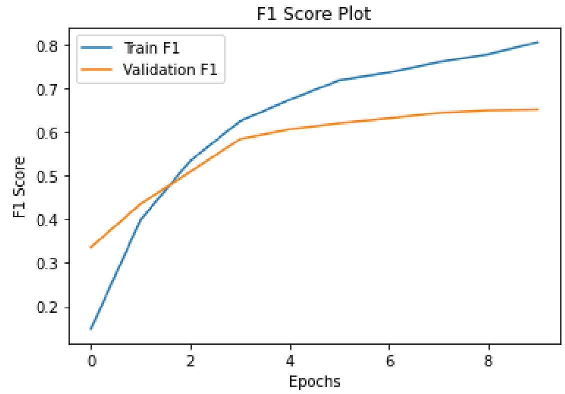
- **Word2Vec Embedding:** The GRU model with Word2Vec shows a strong and stable learning curve, with both the training and validation loss decreasing uniformly. The model's F1 scores for both training and validation rise consistently, suggesting the model generalises well to new data.
- **GloVe Embedding:** The GloVe embedding with the GRU model demonstrates a steady and consistent reduction in loss, indicating stable learning. The F1 scores are also on an upward trajectory with a tight alignment between training and validation, which is indicative of effective generalisation and robust learning.
- **FastText Embedding:** FastText with GRU displays a smooth and consistent decrease in training and validation loss, similar to the GloVe embedding. However, the F1 score, while improving steadily, shows a small gap between training and validation, hinting at slight overfitting compared to GloVe.

Overall, all three embeddings work well with the GRU architecture, each showing that the model is capable of learning and generalising. However, the GloVe embedding stands out as providing the most stable performance across the evaluated metrics.

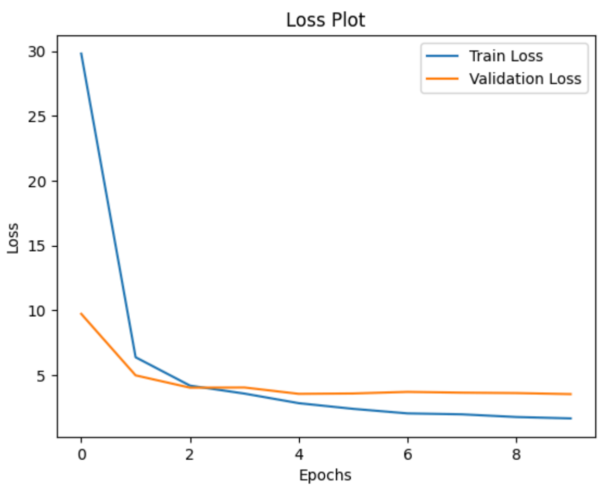
MODEL_4(BiLSTM-CRF)

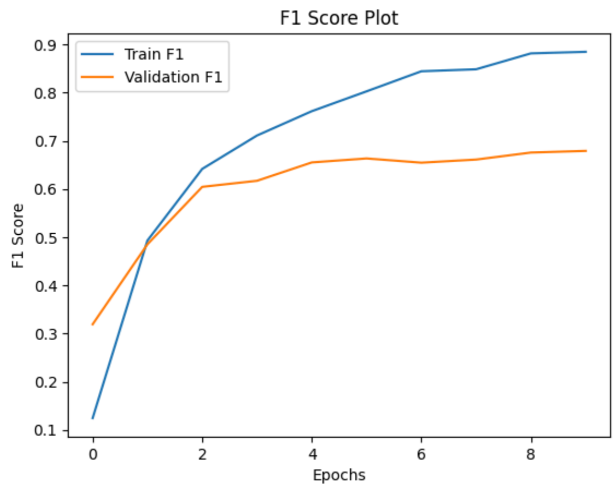
Word2Vec



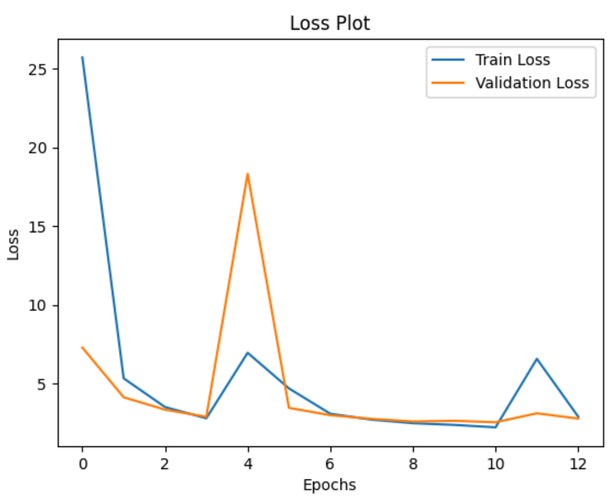


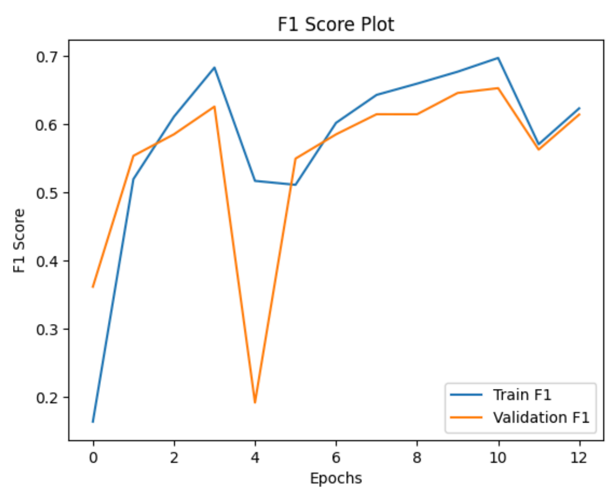
GloVe





FastText





Analysis for BiLSTM-CRF:-

Word2Vec Embedding: This model demonstrates a stable and consistent training process, as reflected in both loss and F1 score plots. The gradual and continuous improvement in the F1 score, along with closely aligned training and validation loss, suggests the model generalises well without overfitting.
GloVe Embedding: The loss plot indicates a possible minor overfitting, as the validation loss stabilises at a higher value compared to the training loss. However, the F1 score plot shows the highest performance, with validation scores improving steadily and closely following the training scores, indicating strong predictive capabilities.
FastText Embedding: There's considerable fluctuation in both the loss and F1 score plots, suggesting potential instability in the model training process. This could be due to various factors like model hyperparameters or the characteristics of the FastText embeddings themselves. Despite this, the model still manages to improve over time, but the variability suggests a need for careful tuning or regularisation.

Overall, GloVe captures the nuances of the data best, as indicated by the high F1 scores, but may benefit from regularisation to manage the overfitting potential. Word2Vec offers a good trade-off between performance and stability, while FastText requires more attention to ensure consistent training and validation results.

Table for TASK1:-

Model_No	Embedding_used	Accuracy	Macro_F1
Model_1(RNN)	Word2Vec	0.981	0.438
Model_1(RNN)	GloVe	0.985	0.529
Model_1(RNN)	FastText	0.986	0.525
Model_2(LSTM)	Word2Vec	0.984	0.507
Model_2(LSTM)	GloVe	0.986	0.569
Model_2(LSTM)	FastText	0.986	0.534
Model_3(GRU)	Word2Vec	0.984	0.556
Model_3(GRU)	GloVe	0.986	0.597
Model_3(GRU)	FastText	0.987	0.551
Model_4(BiLSTM-CRF)	Word2Vec	0.988	0.644
Model_4(BiLSTM-CRF)	GloVe	0.991	0.663
Model_4(BiLSTM-CRF)	FastText	0.990	0.624

2. TASK 2

- No pre-processing was performed on the data. Custom splitting and tokenisation were performed. Please refer to the code for this.

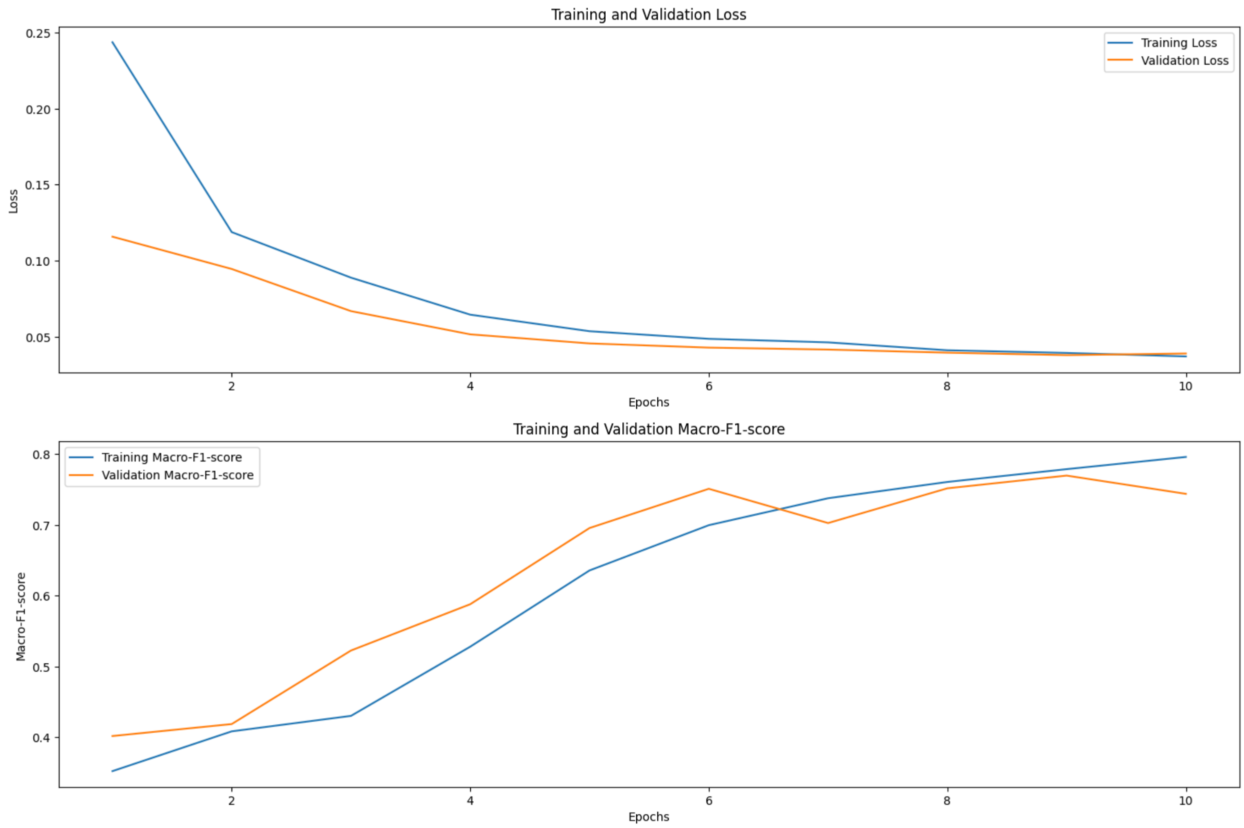
Dataset


```
    "9": {
        "text": "I am pleased with the fast log on , speedy WiFi connection and the long battery life ( > 6 hrs ) .",
        "labels": [
            "O",
            "O",
            "O",
            "O",
            "O",
            "O",
            "B",
            "I",
            "O",
            "O",
            "B",
            "I",
            "O",
            "O",
            "O",
            "B",
            "I",
            "O",
            "O",
            "O",
            "O",
            "O",
            "O"
        ]
    }
```

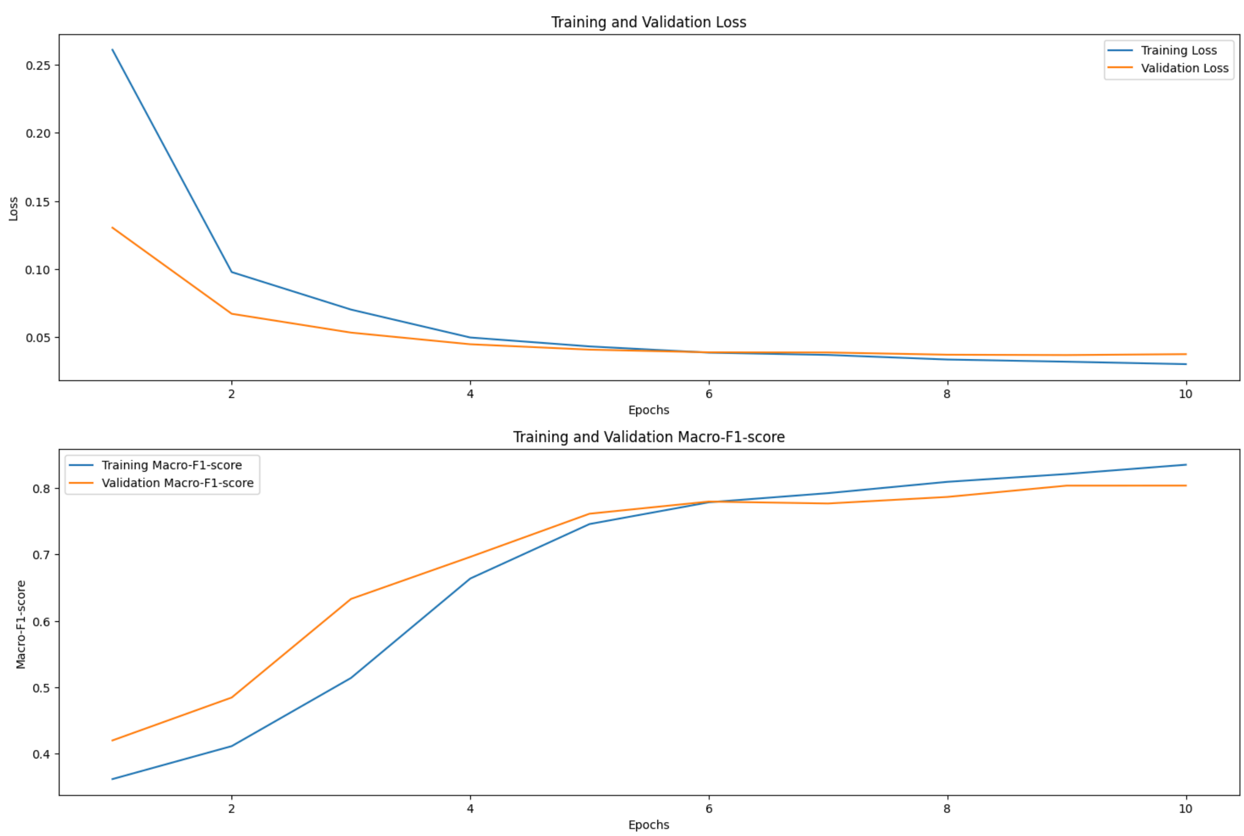

Graphs

MODEL_1(RNN)

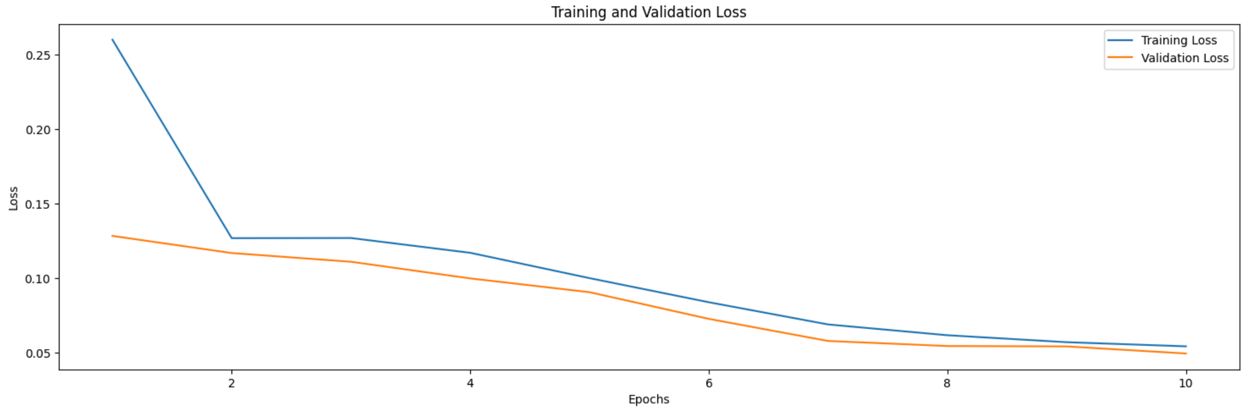
<u>Word2Vec</u>

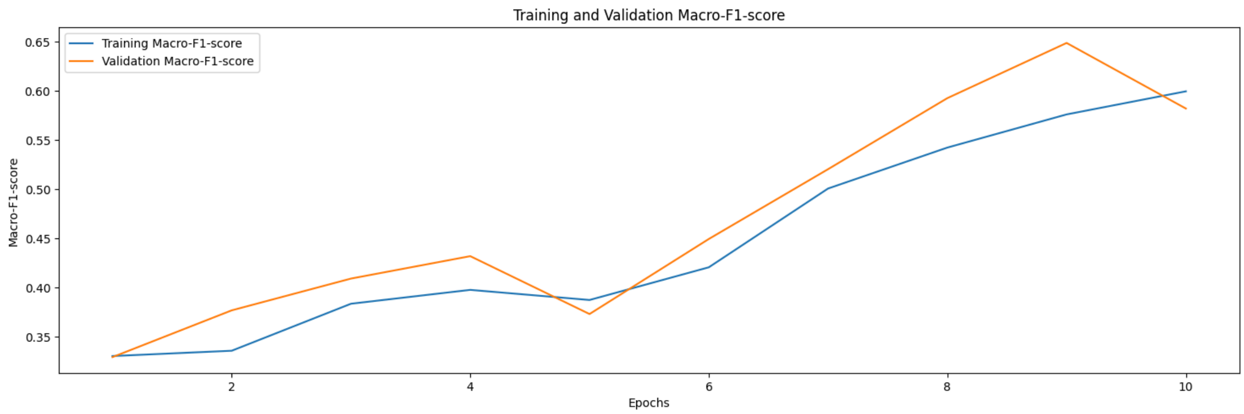


GloVe



FastText





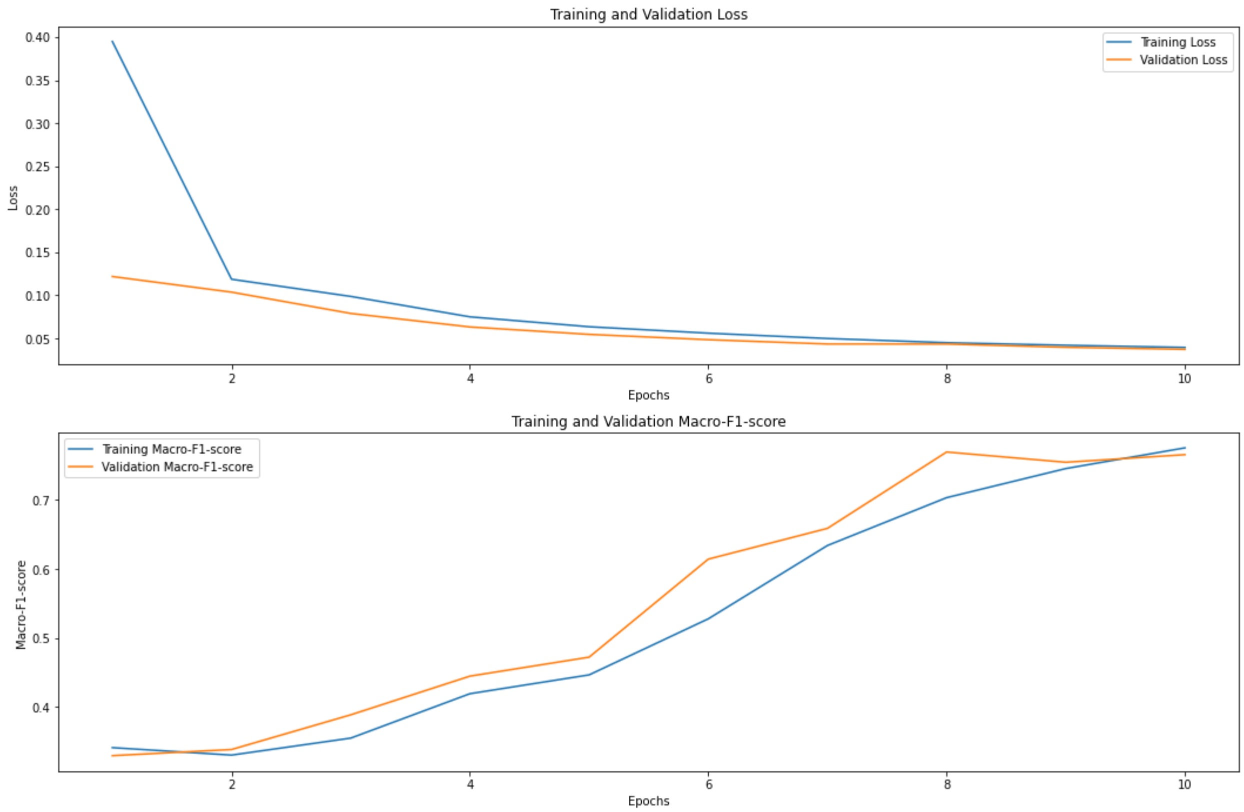
<u>Analysis for RNN:-</u>

- **Word2Vec Embedding:** The RNN model with Word2Vec shows a solid performance with a consistent decline in loss and a steady rise in the Macro-F1 score. This indicates that the model is learning effectively and generalising well to the validation data without significant overfitting.
- **GloVe Embedding:** The GloVe embedding reveals a strong learning capability with a pronounced increase in the Macro-F1 score and a stable decrease in loss. Although the validation loss is slightly higher than the training loss, the F1 scores increase consistently, which suggests effective learning with a potential for minor overfitting.
- **FastText Embedding:** FastText shows a stable loss decrease but with a notable uptick in validation loss towards the later epochs, suggesting the onset of overfitting. The Macro-F1 scores for training and validation increase, although the validation score shows some volatility, indicating that the model might not be as robust to variations in the data as the others.

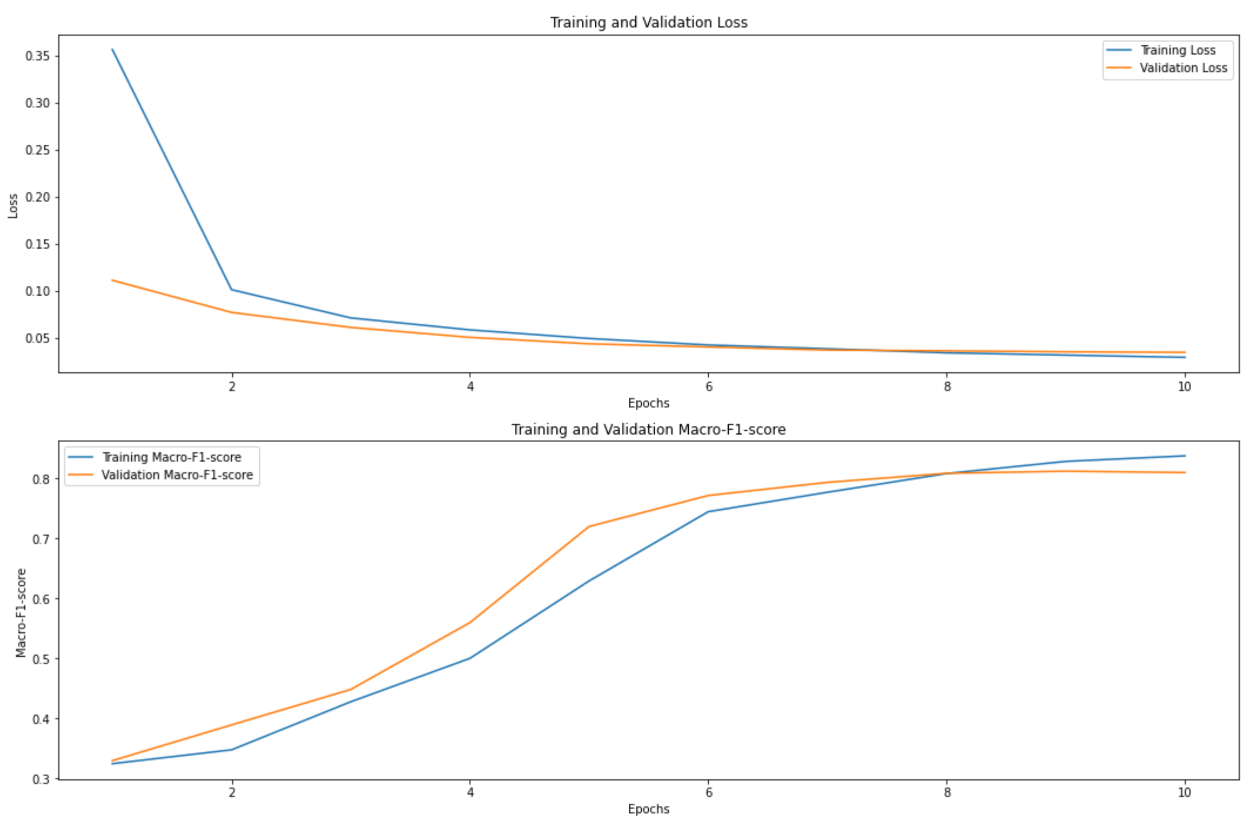
In this comparison, GloVe takes the lead in terms of performance, with Word2Vec providing the most balanced approach in terms of learning and generalisation. FastText, while promising, might need model tuning or data preprocessing adjustments to mitigate the risk of overfitting.

MODEL_2(LSTM)

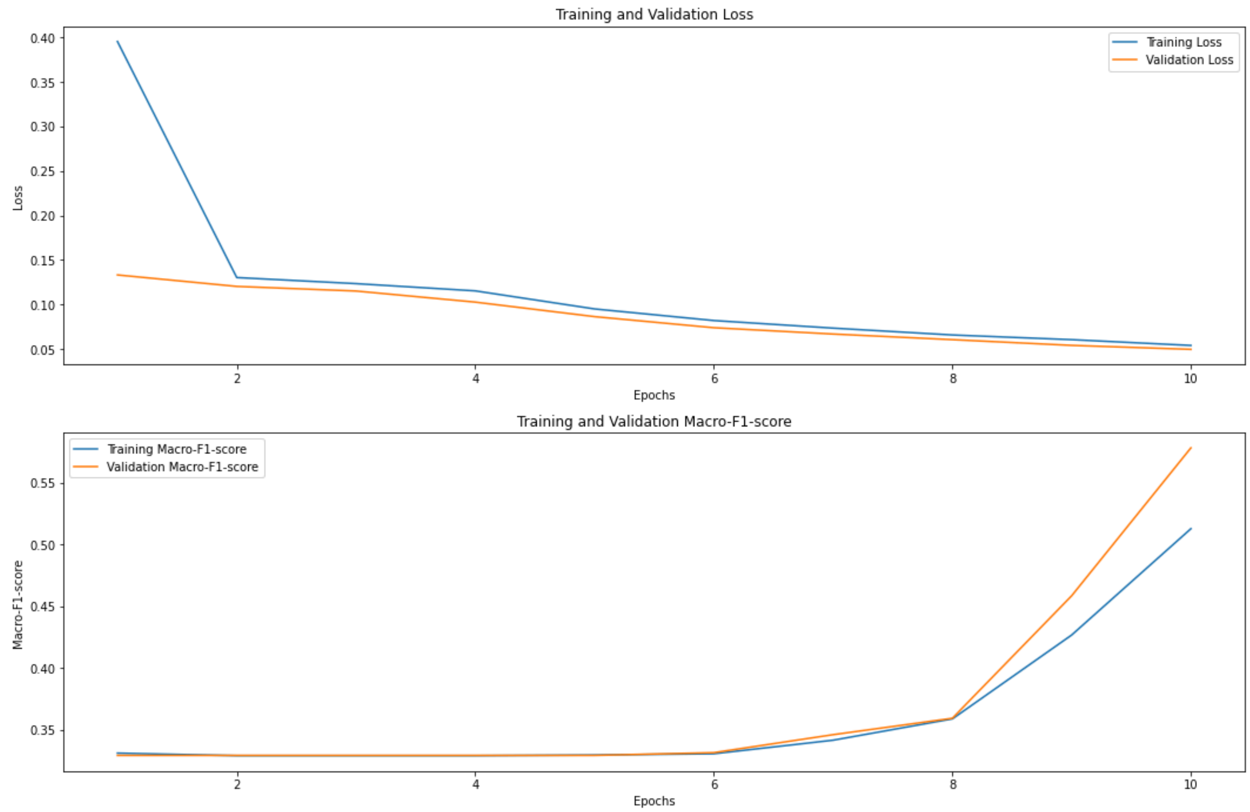
<u>*Word2Vec*</u>



GloVe



FastText

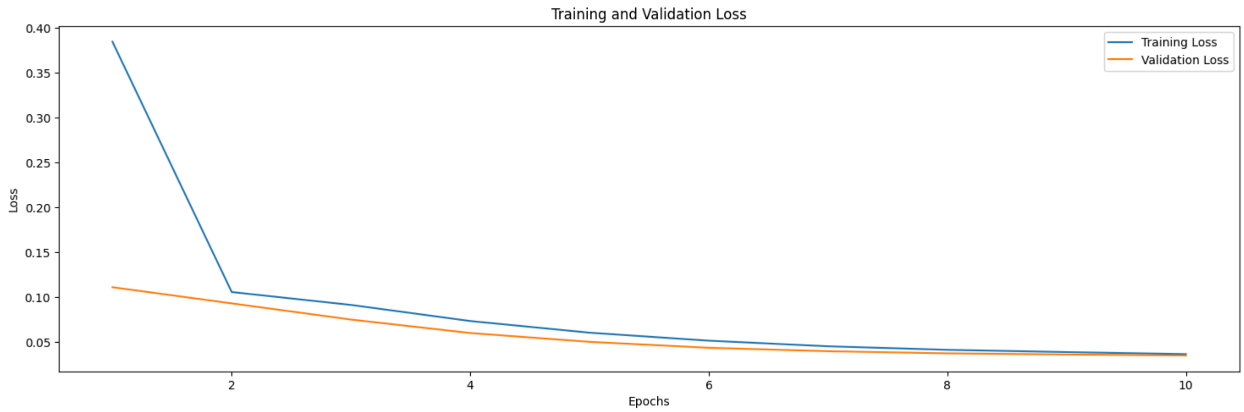


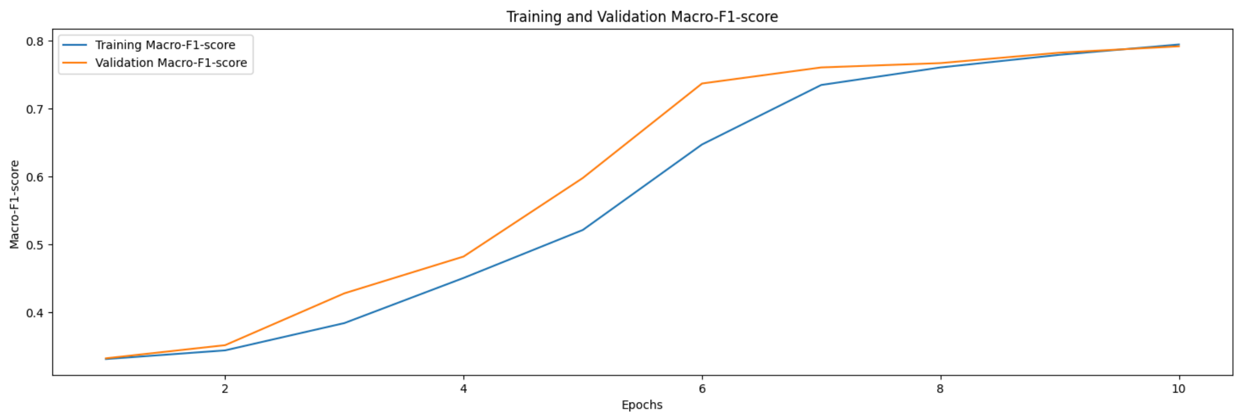
Analysis for LSTM:-

Word2Vec Embedding: The LSTM model trained with Word2Vec demonstrates a consistent reduction in both training and validation loss, suggesting the model fits well to the dataset without overfitting. The Macro-F1 scores for both training and validation display a steady rise, indicating the model is learning and generalising effectively.
GloVe Embedding: The LSTM model using GloVe embeddings shows a very similar pattern in loss reduction to Word2Vec, with the validation loss slightly higher, which is common as the model tries to generalise from training to validation data. The Macro-F1 score is increasing at a good pace, with validation scores closely following the training scores, which indicates robust learning.
FastText Embedding: This model presents a stable decline in training loss, while the validation loss remains flat after an initial drop, which may suggest the model isn't learning much after a certain point or may need further tuning. The Macro-F1 scores show significant improvement, especially in the validation score, which surpasses the training score later in the training.
In summary, the LSTM models perform well with all embeddings. The choice between Word2Vec and GloVe may come down to the specific nuances of the dataset and the validation F1 score.

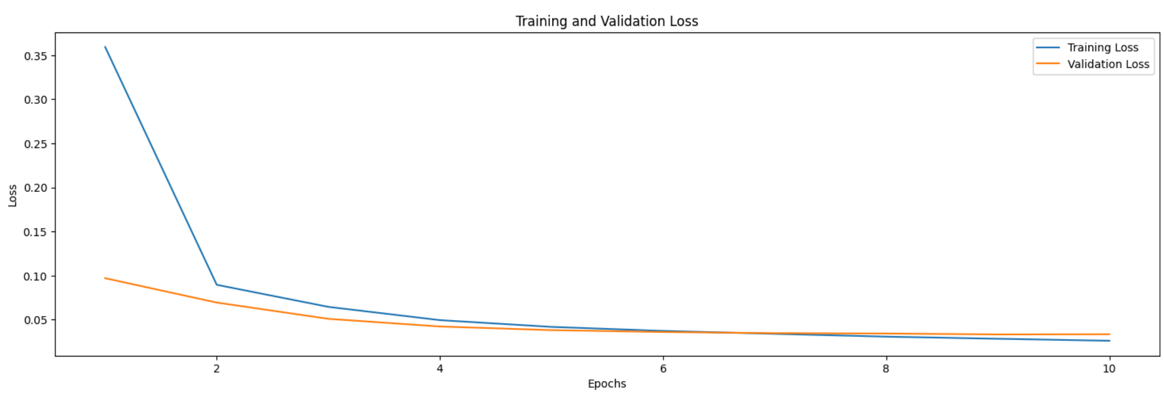
MODEL_3(GRU)

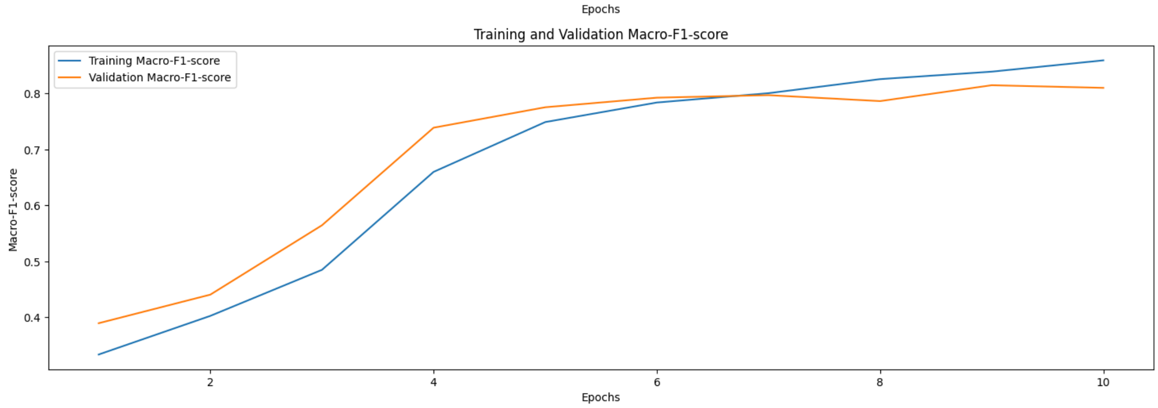
Word2Vec



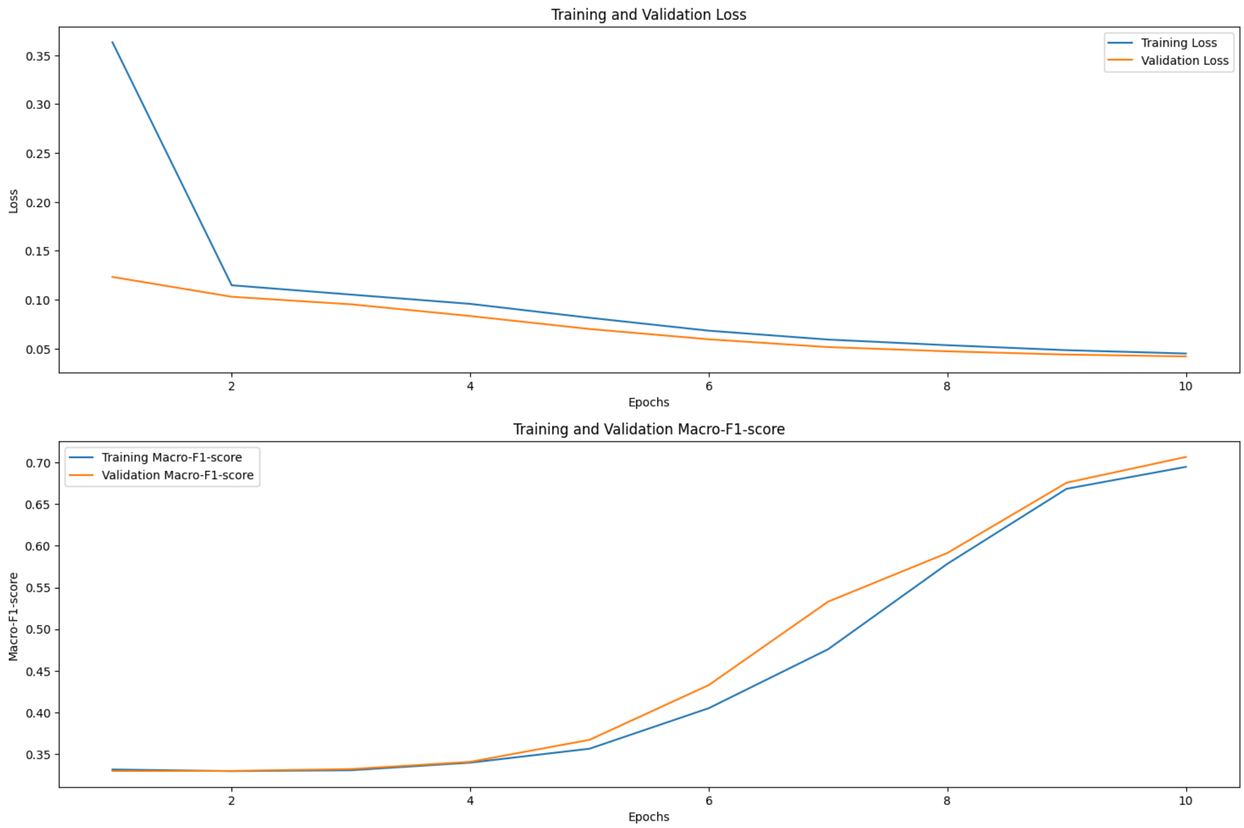


GloVe





FastText

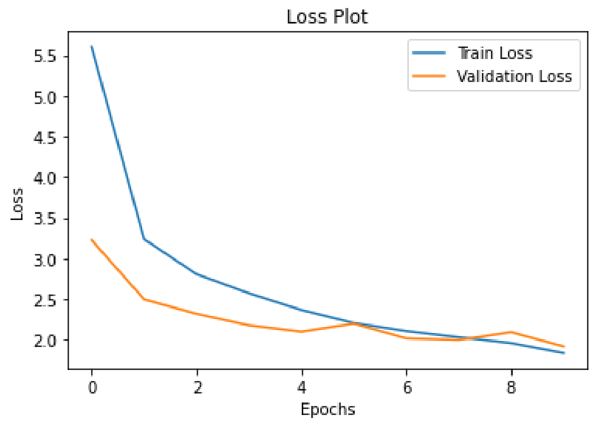


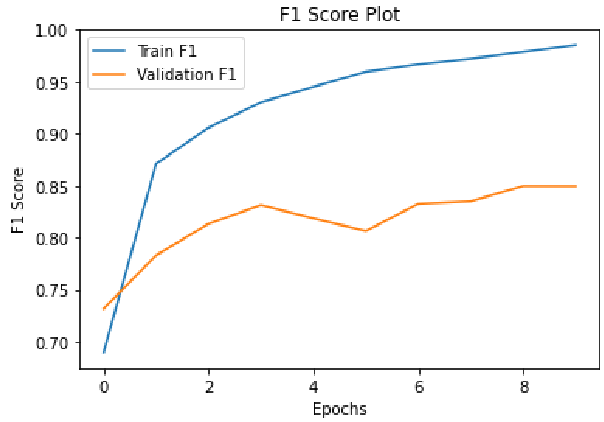
Analysis for GRU:-

Word2Vec Embedding: The GRU model with Word2Vec shows a steady decline in both training and validation loss, indicating a good fit to the data. The Macro-F1 scores for training and validation rise over epochs, with the validation F1 closely tracking the training F1, suggesting effective learning and generalisation capabilities.
GloVe Embedding: Similarly, the GRU model using GloVe embeddings shows a stable decrease in training loss, while the validation loss tends to plateau, indicating a good but slightly less perfect fit than Word2Vec. The Macro-F1 score increases steadily, and the gap between training and validation F1 scores is narrow, which is indicative of good model performance.
FastText Embedding: This model shows a more pronounced decrease in training loss compared to the other two, and the validation loss remains quite flat after the initial epochs, suggesting some degree of overfitting. The Macro-F1 scores increase notably for both training and validation, with the validation score overtaking the training score, which may suggest better generalisation on the specific features captured by the FastText embeddings or variability in the validation set.

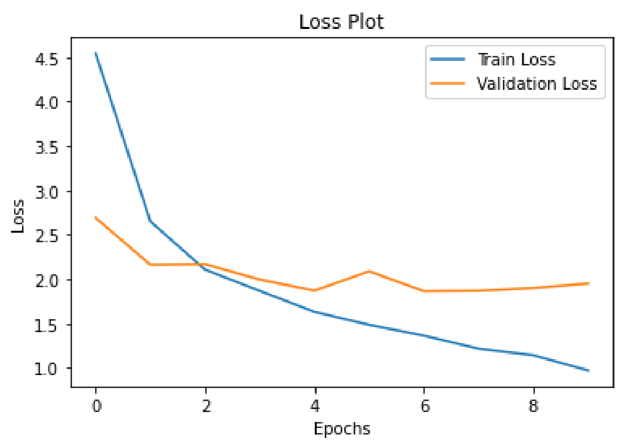
MODEL_4(BiLSTM-CRF)

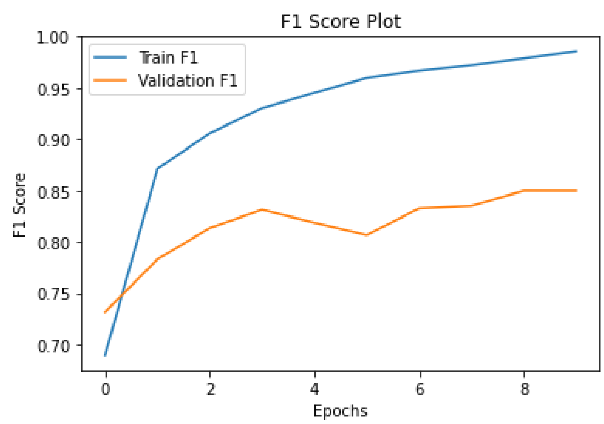
Word2Vec



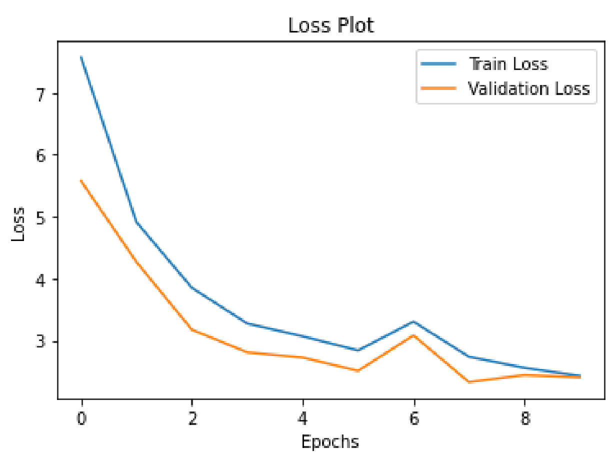


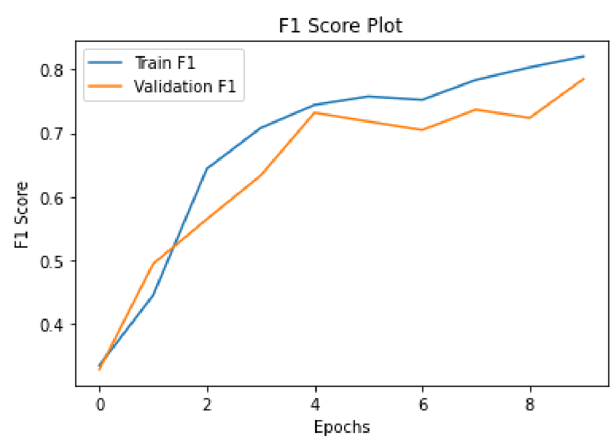
GloVe





FastText





Analysis for BiLSTM-CRF:-

Word2Vec Embedding: The BiLSTM-CRF model with Word2Vec displays a steady decline in training and validation loss, which flattens as epochs increase, indicative of a stable learning process. The Macro-F1 scores show a consistent upward trend, and the validation scores closely follow the training scores, suggesting good generalisation without significant overfitting.

GloVe Embedding: This model shows a similar trend in loss reduction to Word2Vec but with a slight divergence between training and validation loss, potentially indicating minor overfitting. The Macro-F1 scores increase significantly, with the validation score closely trailing the training score, hinting at effective learning with good generalisation capabilities.

FastText Embedding: The loss plot for FastText shows greater fluctuation, suggesting potential instability or model sensitivity to the dataset's nuances. The Macro-F1 scores demonstrate an interesting pattern, with validation scores surpassing training scores at certain points, which could point to particular strengths in the model's generalisation or peculiarities in the validation data.

In summary, the GloVe embedding emerges with a marginally superior performance on the F1 score, though with a slightly higher likelihood of overfitting. Word2Vec shows the most stable and consistent performance, making it a reliable choice for tasks prioritising consistent model behaviour.

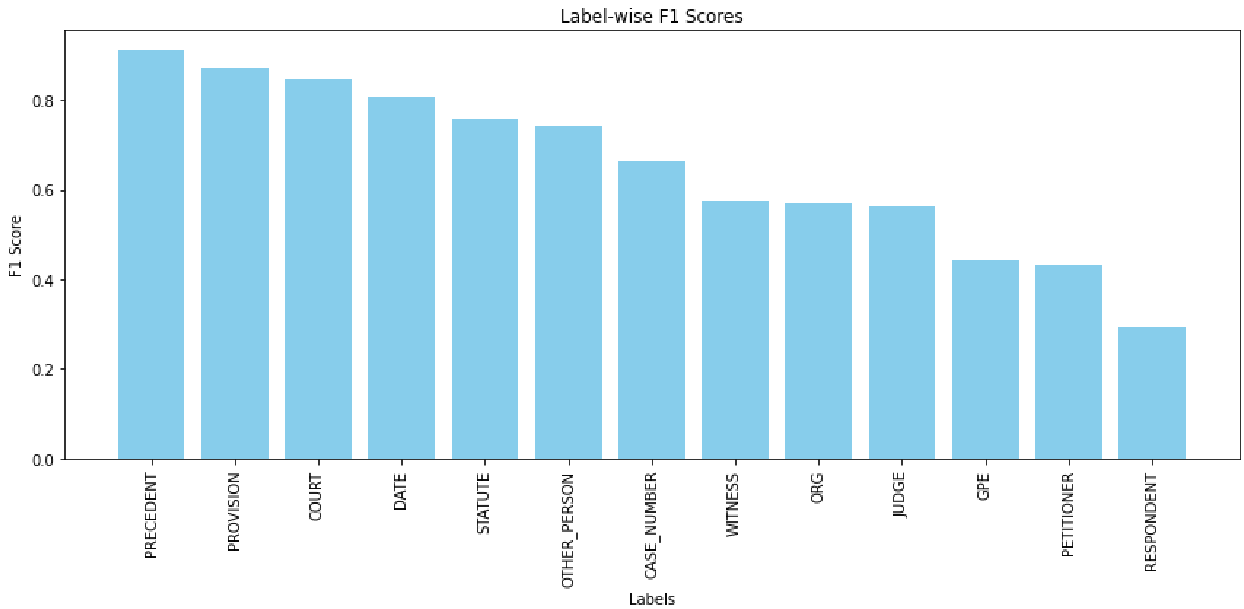
<u>Table for TASK2:-</u>

Model_No	Embedding_used	Accuracy	Macro_F1
Model_1(RNN)	Word2Vec	0.985	0.739
Model_1(RNN)	GloVe	0.986	0.792
Model_1(RNN)	FastText	0.977	0.563
Model_2(LSTM)	Word2Vec	0.985	0.734
Model_2(LSTM)	GloVe	0.986	0.781
Model_2(LSTM)	FastText	0.980	0.597
Model_3(GRU)	Word2Vec	0.986	0.789
Model_3(GRU)	GloVe	0.987	0.809
Model_3(GRU)	FastText	0.983	0.718
Model_4(BiLSTM-CRF)	Word2Vec	0.9871	0.7984
Model_4(BiLSTM-CRF)	GloVe	0.9890	0.8319
Model_4(BiLSTM-CRF)	FastText	0.9855	0.7937

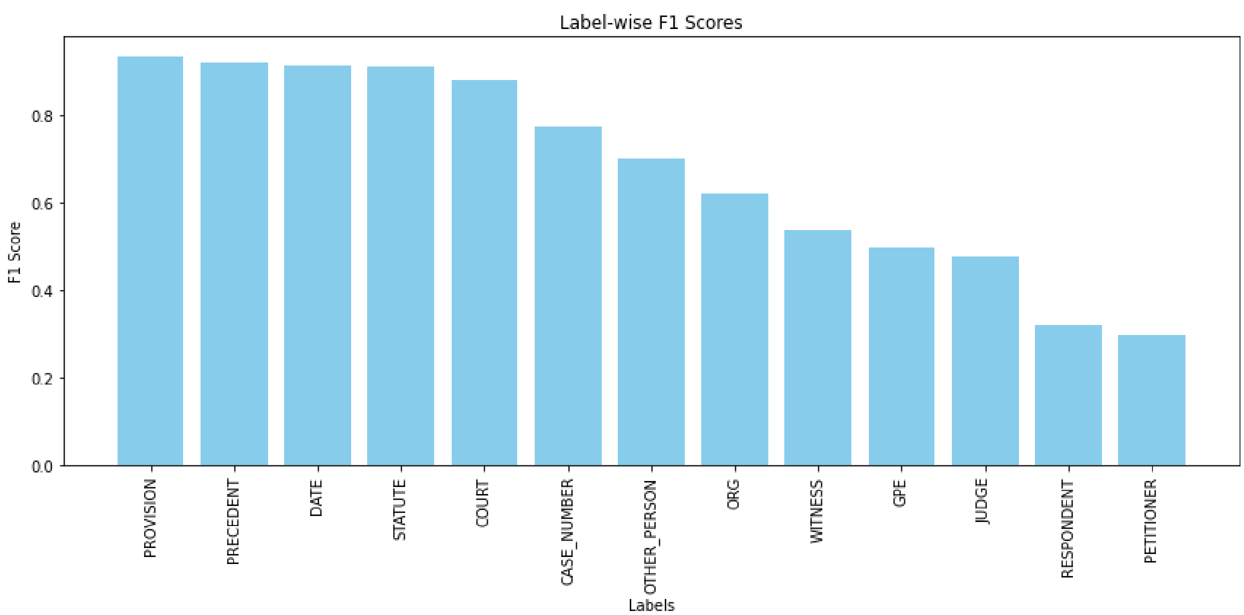
Label-Wise F1 score for Dataset 1:

Henceforth, the best-performing model is BiLSTM-CRF on all three embeddings. The plots are given as follows:

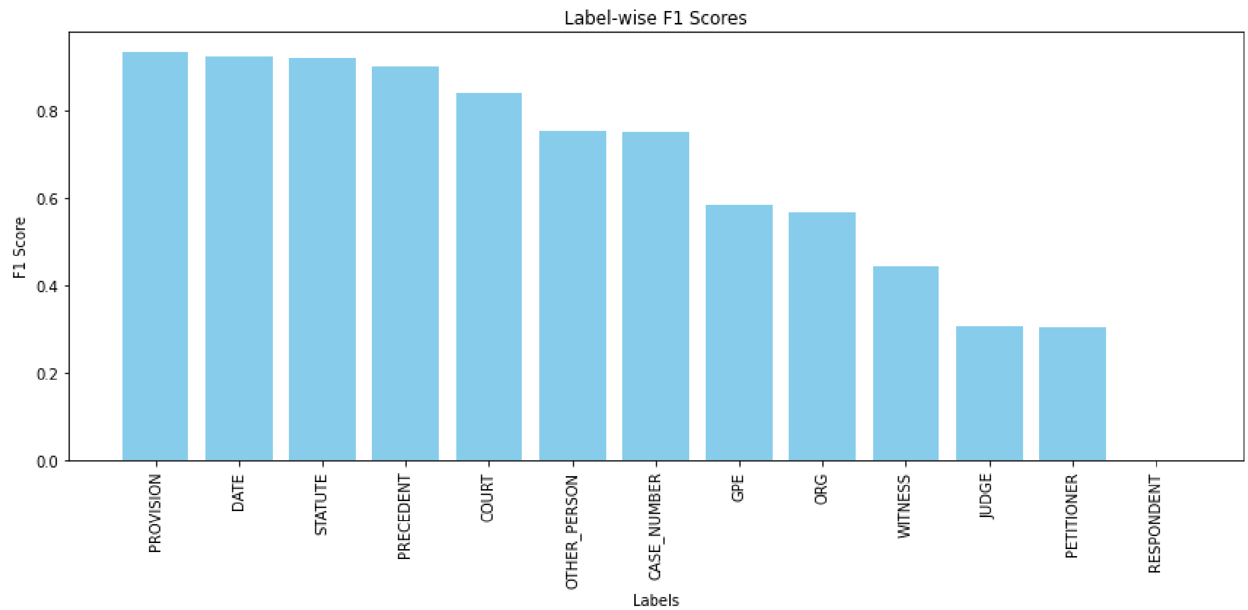
1) Word2Vec:



2) GLoVe:



3) Fasttext:



Credit Statement:

While each member contributed equally to all tasks, we are mentioning a division of tasks just for the sake of the question here.

- **Medha Hira (Roll No: 2021265)** - Fasttext Encoding, GRU Implementation, Report
- **Arnav Goel (Roll No: 2021519)** - Glove Encoding, Bi-LSTM CRF Implementation, Report
- **Siddharth Rajput (Roll No: 2021102)** - BIO-Encoding for Task1 and Task2, LSTM Implementation, Report
- **Amil Bhagat (Roll No: 2021309)** - Word2Vec Encoding, Code for RNN, Report