ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

National and Kapodistrian University of Athens

Department of Informatics and Telecommunications

Deep Learning for Natural Language Processing

Project III - Vaccine Sentiment Classifier using RNNs

Chalkias Spyridon

January 2022

Contents

1 Vaccine Sentiment Classifier

1.1 Task

The task is to perform sentiment analysis on a dataset consisting of tweets and be able to infer whether a new tweet is:

1. Neutral
2. Anti-vax
3. Pro-vax

by stacking *Bidirectional Recurrent Neural Networks.*

NOTE: Inside the project's folder, you will find a sub-folder named **Experiments**, where you can find **all the experiments performed for this project**.

1.2 Data Preprocessing

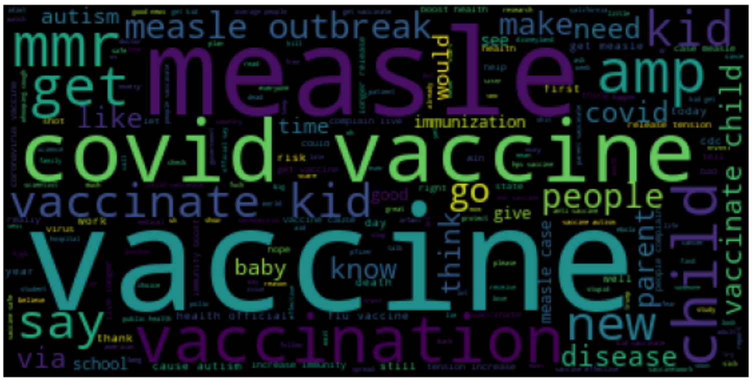
Real-world data tend to be incomplete, noisy, and inconsistent. This can lead to a poor quality of collected data and further to a low quality of models built on such data. In order to address these issues, Data Preprocessing provides operations which can organise the data into a proper form for better understanding in data mining process.
Some of the methods applied in order to prepare the data for analysis are:

- Data Cleaning
- Data Reduction
- Data Transformation

1.2.1 Data Cleaning

The cleaning process involved two (3) operations:

- Visualizing the tweets using **Wordclouds**, in order to expose the dataset's keywords. An indicative wordcloud is shown below:



- Checking whether there's need to drop data rows that contain *null* values. Luckily, the data had no such rows.
- Cleaning tweets from
  - URLs.
  - Noise contained in '< >'parenthesis.
  - The phrase 'RT' (meaning retweet).
  - '@' and the following twitter account name.
  - Any punctuation.
  - '\r\n' which is present in some strings.
  - Numbers, capitalization and white space.

  which are totally useless and add noise to the data.

1.2.2 Data Reduction

Data had a significant reduction process through *Tokenization and Lemmatization*, in order to group together the inflected forms of the words so they can be analysed in less items which are more independent and have greater semantic content. Also, all the stopwords were removed, so that data delivers more semantic content with as few words as possible. The stopwords container used was provided by *NLTK* library and the lemmatization process was supported by *spaCy* library.

1.2.3 Data Transformation

The data was transformed by using a **Word Embeddings** representation consisting of pre-trained vectors from GloVe's twitter dataset.

1.3 Tackling the imbalanced classes problem

An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. The distribution can vary from a slight bias to a severe imbalance where there is one example in the minority class for hundreds, thousands, or millions of examples in the majority class or classes.

In our case, the dataset suffers from the aforementioned problem, since the *Neutral* class consists of 7458 samples, the *Anti-Vax* class consists of 2073 samples and the *Pro-Vax* class consists of 6445 samples.

There are three (3) methods that have been tested in order to reduce the class imbalance:

- Upsampling by duplicating samples in minor class.
- Downsampling by cutting down samples from the major classes.
- Upsampling by generating synthetic samples from the minor class.

In the dataset provided, no method proved to be better than just using the raw data themselves.

In the given *.ipynb* notebook, there has not been applied any imbalance tackling method to the dataset.

1.4 Defining the Neural Network

A single class named **RNN** is defined and can support both *LSTM* and *GRU* Recurrent NN models. More specifically, it includes:

- **torch.nn.LSTM** or **torch.nn.GRU** models fed with the suitable hyperparameters.
- **Gradient Clipping** support.
- **Skip Connection** support.
- **Attention** support.

The aforementioned class is also built in order to support **Bidirectionality**.

1.5 Hyperparameter Tuning

In this project, several features were tested and fine-tuned in order maximize the network's generalization capabilities, which are discussed below.

1.5.1 Number of stacked RNNs

The optimal number of stacked RNNs seemed to be between three (3) and four (4) stacked RNNs, with the statistically best results produced by four (4).

1.5.2 Number of hidden features

The hidden features number seemed to be the best at around 50 features.

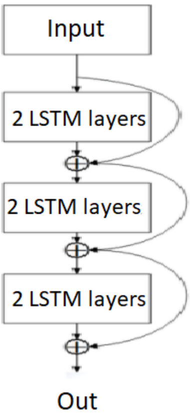
1.5.3 Type of cells (LSTM / GRU)

Both LSTM and GRU models produced quite similar results without pretty big differences. A few things to consider about their performance on the experiments are:

- LSTM models tended to overfit a bit more than GRU models.
- Both models performed even better with gradient clipping and skip connections.

1.5.4 Skip Connections

Both LSTM and GRU models were tested with skip connections. The current project's skip connection architecture is described below:



That way, by using a skip connection per two (2) stacked LSTM/GRU layers, we provide an alternative path for the gradient. The core idea is to backpropagate through the identity function, by just using a vector addition. Then, the gradient would simply be multiplied by one and its value will be maintained in the earlier layers. This is the main idea behind Residual Networks (ResNets): they stack these skip residual blocks together. That's why an identity function is used to preserve the gradient. Overall, skip connections were effective and slowed down the overfitting models.

1.5.5 Gradient Clipping

Gradient clipping ensures the gradient vector has norm at most n. In current project's implementation, $n = 3$. This helps gradient descent to have a reasonable behaviour even if the loss landscape of the model is irregular.

1.5.6 Dropout

Although dropout ratio's effectiveness may vary from one model to another, the overall best value tended to be close to 0.5, as it slowed down the models from overfitting, and produced pretty good results.

1.6 A few words about Word Embeddings

Word Embeddings model is a pretty efficient and manageable model. In order to train it, GloVe's pre-trained Twitter dataset was downloaded and used.

Some of the key advantages of this model are:

- Its small input size. The input size of the aforementioned model can be at max 200 units, since GloVe's pre-trained vectors can be maximum 200-dimensional. That way, the training sessions are way faster than every other model's.
- It enforces the word vectors to capture sub-linear relationships in the vector space.
- GloVe does not rely just on local statistics (local context information of words), but incorporates global statistics (word co-occurrence) to obtain word vectors.

The Word Embeddings model proved to performed pretty high, by scoring a weighted average f1-score of 71%.

1.7 Scores

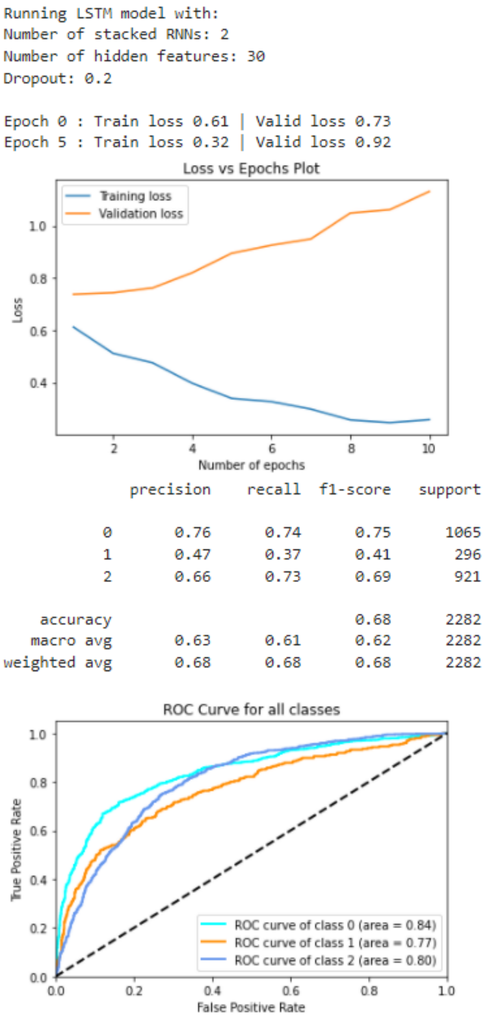
The following diagrams are showcasing some important training sessions that were critical for the choosing of the overall best performing models.

Note that the training sessions corresponding to the following diagrams do not appear in the provided notebook, because that would make it huge. The notebook only showcases the best performing version of every model.

1.7.1 LSTM cells without Gradient Clipping and Skip Connections

Below are some runs using:

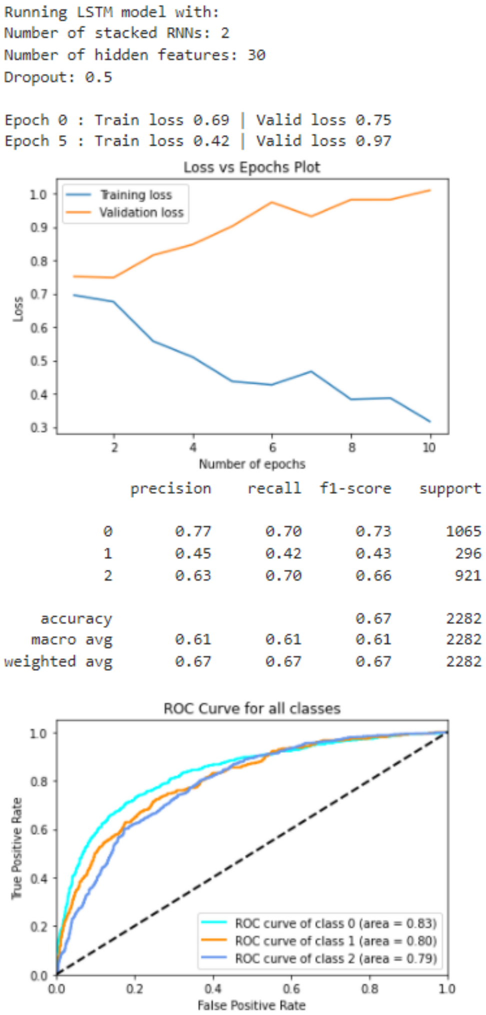
- Stacked LSTM layers.
- No Gradient Clipping.
- No Skip Connections.



```
Running LSTM model with:
Number of stacked RNNs: 2
Number of hidden features: 30
Dropout: 0.2

Epoch 0 : Train loss 0.61 | Valid loss 0.73
Epoch 5 : Train loss 0.32 | Valid loss 0.92
```

```
              precision    recall  f1-score   support

           0       0.76      0.74      0.75      1065
           1       0.47      0.37      0.41       296
           2       0.66      0.73      0.69       921

    accuracy                           0.68      2282
   macro avg       0.63      0.61      0.62      2282
weighted avg       0.68      0.68      0.68      2282
```

ROC Curve for all classes
True Positive Rate
False Positive Rate
ROC curve of class 0 (area = 0.84)
ROC curve of class 1 (area = 0.77)
ROC curve of class 2 (area = 0.80)
0.0
0.2
0.4
0.6
0.8
1.0



```
Running LSTM model with:
Number of stacked RNNs: 2
Number of hidden features: 30
Dropout: 0.5

Epoch 0 : Train loss 0.69 | Valid loss 0.75
Epoch 5 : Train loss 0.42 | Valid loss 0.97
```

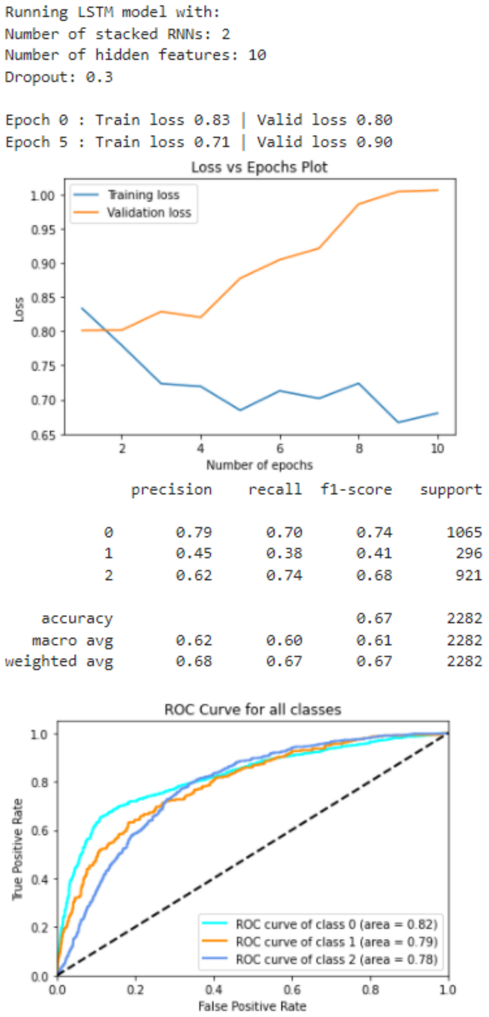
```
              precision    recall  f1-score   support

           0       0.77      0.70      0.73      1065
           1       0.45      0.42      0.43       296
           2       0.63      0.70      0.66       921

    accuracy                           0.67      2282
   macro avg       0.61      0.61      0.61      2282
weighted avg       0.67      0.67      0.67      2282
```

ROC Curve for all classes
True Positive Rate
False Positive Rate
ROC curve of class 0 (area = 0.83)
ROC curve of class 1 (area = 0.80)
ROC curve of class 2 (area = 0.79)
0.0
0.2
0.4
0.6
0.8
1.0

1.7.2 LSTM cells with Gradient Clipping and no Skip Connections

Below are some runs using:

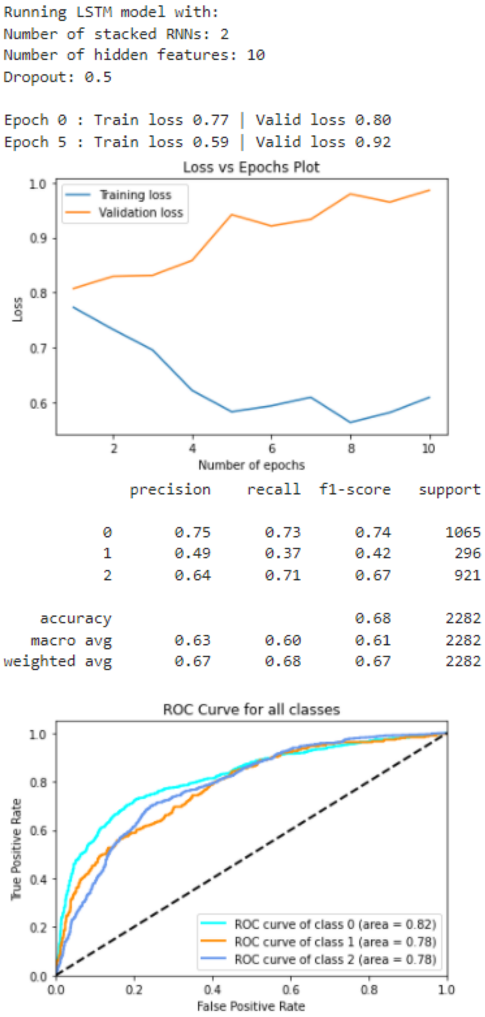
- Stacked LSTM layers.
- Gradient Clipping (max norm = 3).
- No Skip Connections.



```
Running LSTM model with:
Number of stacked RNNs: 2
Number of hidden features: 10
Dropout: 0.3

Epoch 0 : Train loss 0.83 | Valid loss 0.80
Epoch 5 : Train loss 0.71 | Valid loss 0.90
```

```
              precision    recall  f1-score   support

           0       0.79      0.70      0.74      1065
           1       0.45      0.38      0.41       296
           2       0.62      0.74      0.68       921

    accuracy                           0.67      2282
   macro avg       0.62      0.60      0.61      2282
weighted avg       0.68      0.67      0.67      2282
```

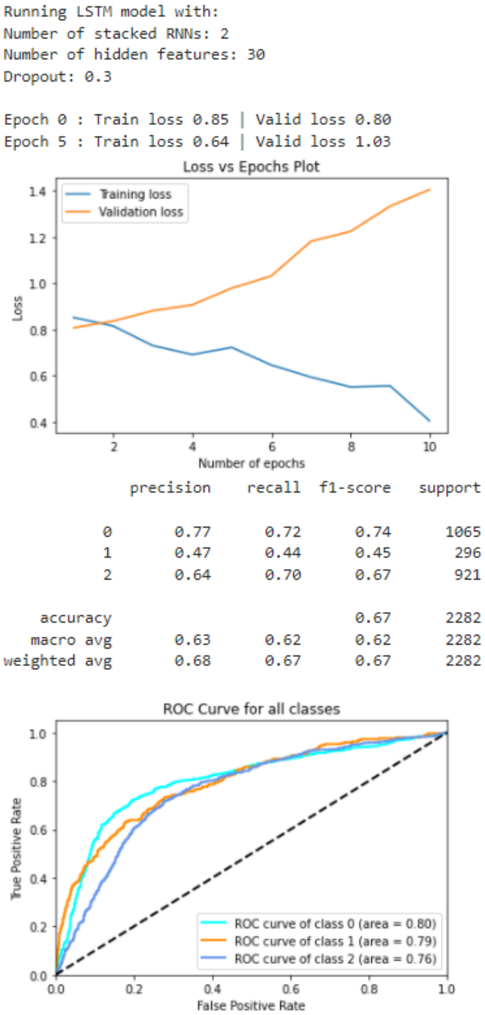
ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.82)
ROC curve of class 1 (area = 0.79)
ROC curve of class 2 (area = 0.78)



```
Running LSTM model with:
Number of stacked RNNs: 2
Number of hidden features: 10
Dropout: 0.5

Epoch 0 : Train loss 0.77 | Valid loss 0.80
Epoch 5 : Train loss 0.59 | Valid loss 0.92
```

```
              precision    recall  f1-score   support

           0       0.75      0.73      0.74      1065
           1       0.49      0.37      0.42       296
           2       0.64      0.71      0.67       921

    accuracy                           0.68      2282
   macro avg       0.63      0.60      0.61      2282
weighted avg       0.67      0.68      0.67      2282
```

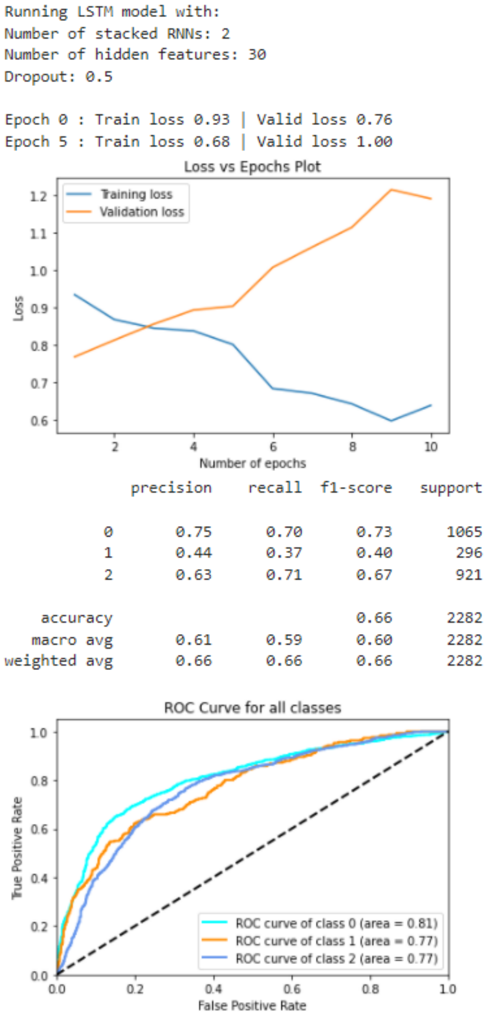
ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.82)
ROC curve of class 1 (area = 0.78)
ROC curve of class 2 (area = 0.78)



```
Running LSTM model with:
Number of stacked RNNs: 2
Number of hidden features: 30
Dropout: 0.3

Epoch 0 : Train loss 0.85 | Valid loss 0.80
Epoch 5 : Train loss 0.64 | Valid loss 1.03
```

```
              precision    recall  f1-score   support

           0       0.77      0.72      0.74      1065
           1       0.47      0.44      0.45       296
           2       0.64      0.70      0.67       921

    accuracy                           0.67      2282
   macro avg       0.63      0.62      0.62      2282
weighted avg       0.68      0.67      0.67      2282
```

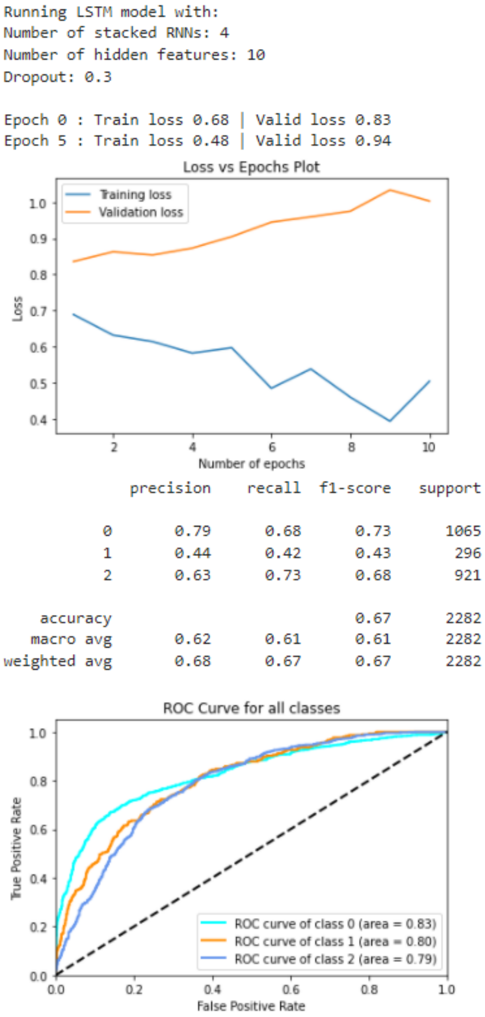
ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.80)
ROC curve of class 1 (area = 0.79)
ROC curve of class 2 (area = 0.76)



```
Running LSTM model with:
Number of stacked RNNs: 2
Number of hidden features: 30
Dropout: 0.5

Epoch 0 : Train loss 0.93 | Valid loss 0.76
Epoch 5 : Train loss 0.68 | Valid loss 1.00
```

```
              precision    recall  f1-score   support

           0       0.75      0.70      0.73      1065
           1       0.44      0.37      0.40       296
           2       0.63      0.71      0.67       921

    accuracy                           0.66      2282
   macro avg       0.61      0.59      0.60      2282
weighted avg       0.66      0.66      0.66      2282
```

ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.81)
ROC curve of class 1 (area = 0.77)
ROC curve of class 2 (area = 0.77)



```
Running LSTM model with:
Number of stacked RNNs: 4
Number of hidden features: 10
Dropout: 0.3

Epoch 0 : Train loss 0.68 | Valid loss 0.83
Epoch 5 : Train loss 0.48 | Valid loss 0.94
```

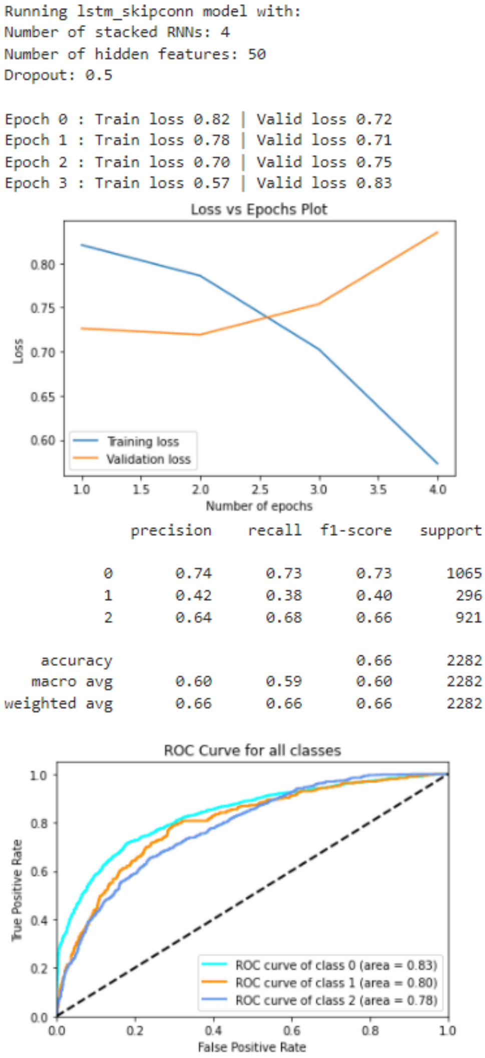
```
              precision    recall  f1-score   support

           0       0.79      0.68      0.73      1065
           1       0.44      0.42      0.43       296
           2       0.63      0.73      0.68       921

    accuracy                           0.67      2282
   macro avg       0.62      0.61      0.61      2282
weighted avg       0.68      0.67      0.67      2282
```

ROC Curve for all classes
True Positive Rate
False Positive Rate
ROC curve of class 0 (area = 0.83)
ROC curve of class 1 (area = 0.80)
ROC curve of class 2 (area = 0.79)
0.0
0.2
0.4
0.6
0.8
1.0

1.7.3 LSTM cells without Gradient Clipping but with Skip Connections

Below are some runs using:

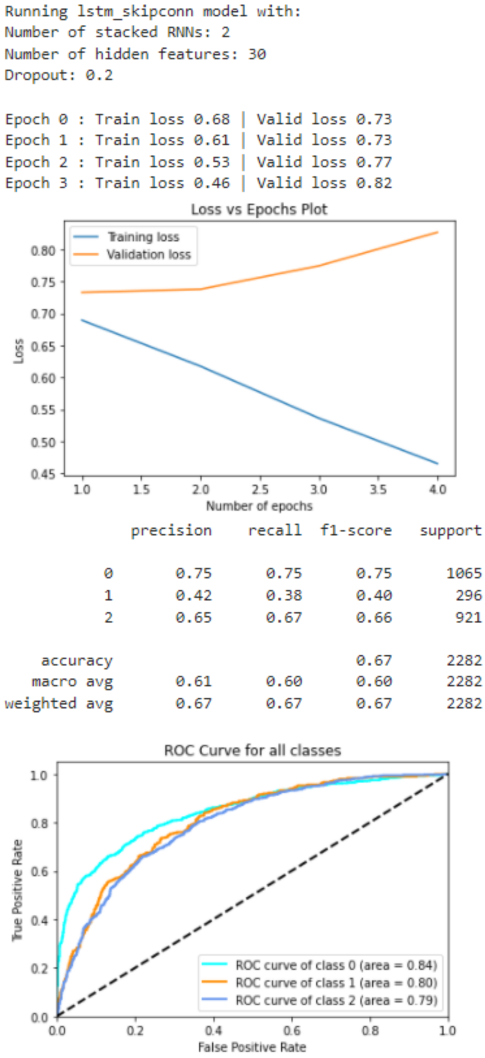
- Stacked LSTM layers.
- No Gradient Clipping.
- Skip Connections.



```
Running lstm_skipconn model with:
Number of stacked RNNs: 4
Number of hidden features: 50
Dropout: 0.5

Epoch 0 : Train loss 0.82 | Valid loss 0.72
Epoch 1 : Train loss 0.78 | Valid loss 0.71
Epoch 2 : Train loss 0.70 | Valid loss 0.75
Epoch 3 : Train loss 0.57 | Valid loss 0.83
```

```
              precision    recall  f1-score   support

           0       0.74      0.73      0.73      1065
           1       0.42      0.38      0.40       296
           2       0.64      0.68      0.66       921

    accuracy                           0.66      2282
   macro avg       0.60      0.59      0.60      2282
weighted avg       0.66      0.66      0.66      2282
```

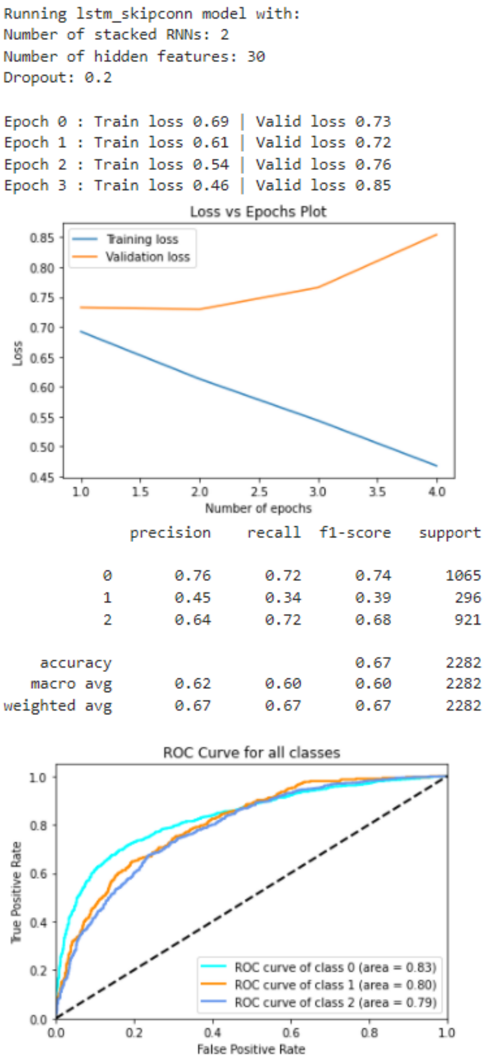
ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
ROC curve of class 0 (area = 0.83)
ROC curve of class 1 (area = 0.80)
ROC curve of class 2 (area = 0.78)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate



```
Running lstm_skipconn model with:
Number of stacked RNNs: 2
Number of hidden features: 30
Dropout: 0.2

Epoch 0 : Train loss 0.68 | Valid loss 0.73
Epoch 1 : Train loss 0.61 | Valid loss 0.73
Epoch 2 : Train loss 0.53 | Valid loss 0.77
Epoch 3 : Train loss 0.46 | Valid loss 0.82
```

```
              precision    recall  f1-score   support

           0       0.75      0.75      0.75      1065
           1       0.42      0.38      0.40       296
           2       0.65      0.67      0.66       921

    accuracy                           0.67      2282
   macro avg       0.61      0.60      0.60      2282
weighted avg       0.67      0.67      0.67      2282
```

ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
ROC curve of class 0 (area = 0.84)
ROC curve of class 1 (area = 0.80)
ROC curve of class 2 (area = 0.79)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate



```
Running lstm_skipconn model with:
Number of stacked RNNs: 2
Number of hidden features: 30
Dropout: 0.2

Epoch 0 : Train loss 0.69 | Valid loss 0.73
Epoch 1 : Train loss 0.61 | Valid loss 0.72
Epoch 2 : Train loss 0.54 | Valid loss 0.76
Epoch 3 : Train loss 0.46 | Valid loss 0.85
```

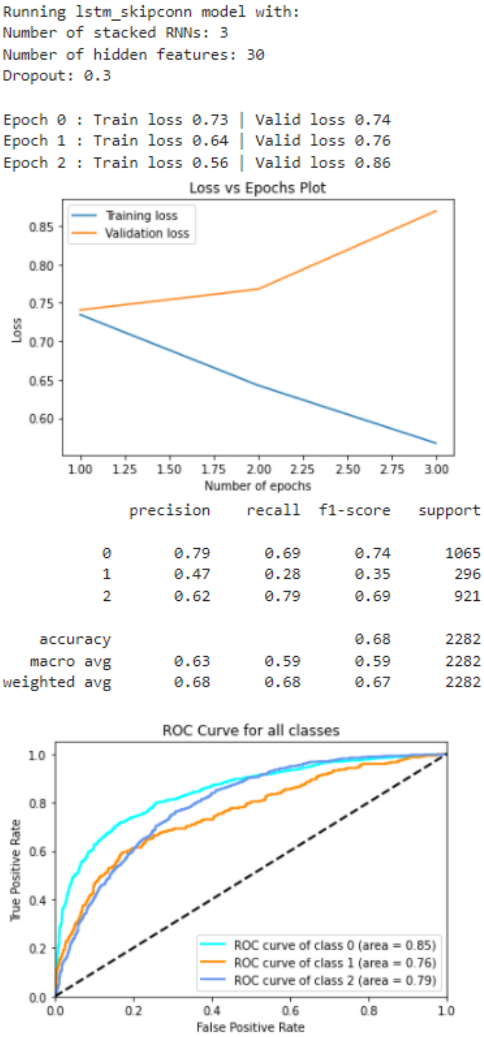
```
              precision    recall  f1-score   support

           0       0.76      0.72      0.74      1065
           1       0.45      0.34      0.39       296
           2       0.64      0.72      0.68       921

    accuracy                           0.67      2282
   macro avg       0.62      0.60      0.60      2282
weighted avg       0.67      0.67      0.67      2282
```

ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.83)
ROC curve of class 1 (area = 0.80)
ROC curve of class 2 (area = 0.79)

1.7.4 LSTM cells with both Gradient Clipping and Skip Connections

Below are some runs using:

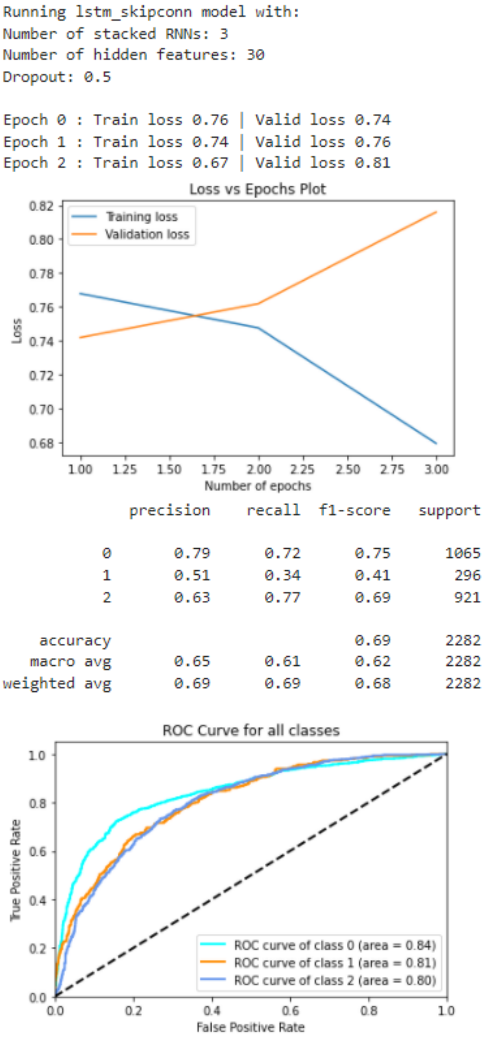
- Stacked LSTM layers.
- Gradient Clipping.
- Skip Connections.



```
Running lstm_skipconn model with:
Number of stacked RNNs: 3
Number of hidden features: 30
Dropout: 0.3

Epoch 0 : Train loss 0.73 | Valid loss 0.74
Epoch 1 : Train loss 0.64 | Valid loss 0.76
Epoch 2 : Train loss 0.56 | Valid loss 0.86
```

```
              precision    recall  f1-score   support

           0       0.79      0.69      0.74      1065
           1       0.47      0.28      0.35       296
           2       0.62      0.79      0.69       921

    accuracy                           0.68      2282
   macro avg       0.63      0.59      0.59      2282
weighted avg       0.68      0.68      0.67      2282
```

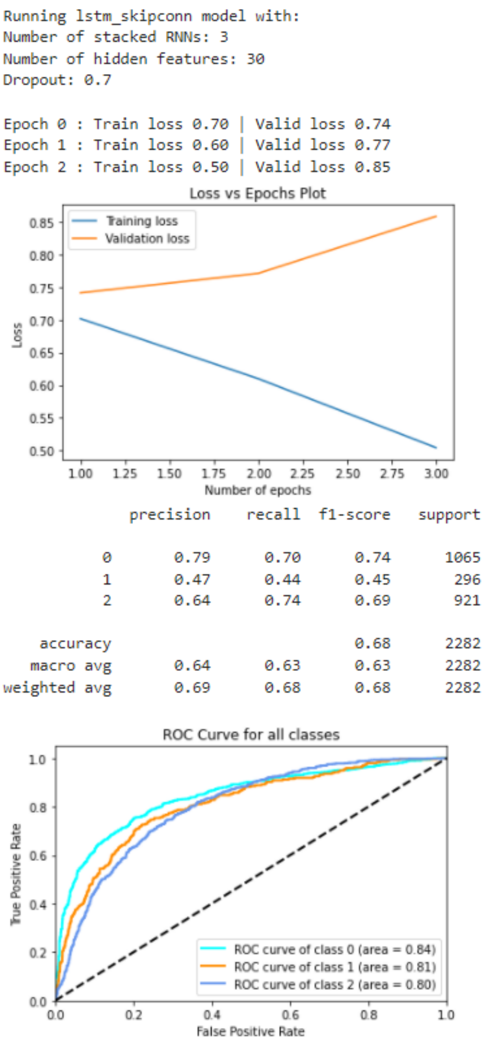
ROC Curve for all classes
ROC curve of class 0 (area = 0.85)
ROC curve of class 1 (area = 0.76)
ROC curve of class 2 (area = 0.79)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0



```
Running lstm_skipconn model with:
Number of stacked RNNs: 3
Number of hidden features: 30
Dropout: 0.5

Epoch 0 : Train loss 0.76 | Valid loss 0.74
Epoch 1 : Train loss 0.74 | Valid loss 0.76
Epoch 2 : Train loss 0.67 | Valid loss 0.81
```

```
              precision    recall  f1-score   support

           0       0.79      0.72      0.75      1065
           1       0.51      0.34      0.41       296
           2       0.63      0.77      0.69       921

    accuracy                           0.69      2282
   macro avg       0.65      0.61      0.62      2282
weighted avg       0.69      0.69      0.68      2282
```

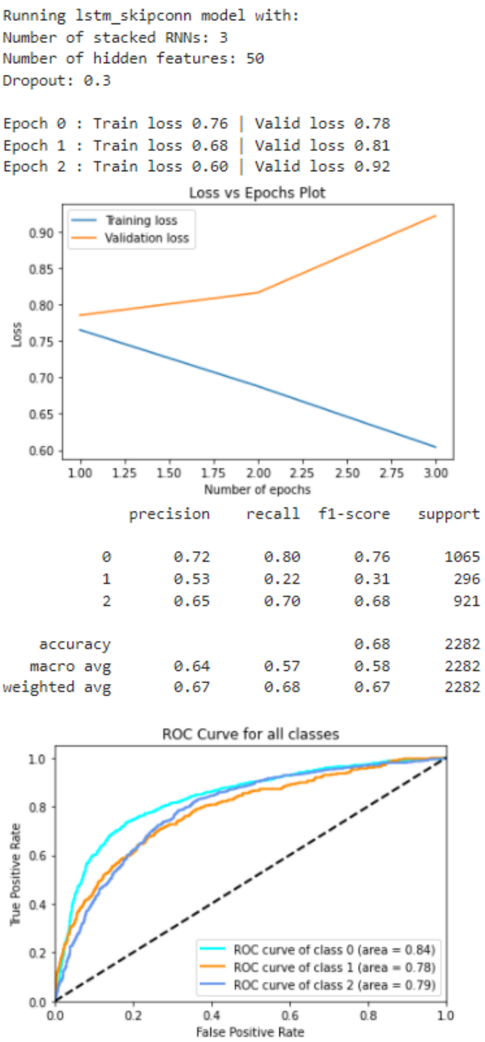
ROC Curve for all classes
ROC curve of class 0 (area = 0.84)
ROC curve of class 1 (area = 0.81)
ROC curve of class 2 (area = 0.80)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0



```
Running lstm_skipconn model with:
Number of stacked RNNs: 3
Number of hidden features: 30
Dropout: 0.7

Epoch 0 : Train loss 0.70 | Valid loss 0.74
Epoch 1 : Train loss 0.60 | Valid loss 0.77
Epoch 2 : Train loss 0.50 | Valid loss 0.85
```

```
              precision    recall  f1-score   support

           0       0.79      0.70      0.74      1065
           1       0.47      0.44      0.45       296
           2       0.64      0.74      0.69       921

    accuracy                           0.68      2282
   macro avg       0.64      0.63      0.63      2282
weighted avg       0.69      0.68      0.68      2282
```

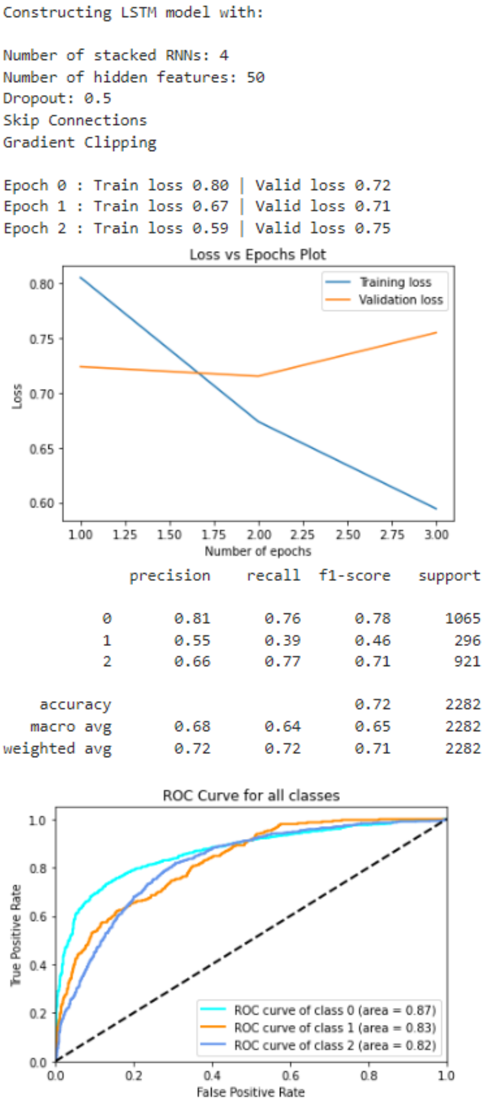
ROC Curve for all classes
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.84)
ROC curve of class 1 (area = 0.81)
ROC curve of class 2 (area = 0.80)



```
Running lstm_skipconn model with:
Number of stacked RNNs: 3
Number of hidden features: 50
Dropout: 0.3

Epoch 0 : Train loss 0.76 | Valid loss 0.78
Epoch 1 : Train loss 0.68 | Valid loss 0.81
Epoch 2 : Train loss 0.60 | Valid loss 0.92
```

```
              precision    recall  f1-score   support

           0       0.72      0.80      0.76      1065
           1       0.53      0.22      0.31       296
           2       0.65      0.70      0.68       921

    accuracy                           0.68      2282
   macro avg       0.64      0.57      0.58      2282
weighted avg       0.67      0.68      0.67      2282
```

ROC Curve for all classes
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.84)
ROC curve of class 1 (area = 0.78)
ROC curve of class 2 (area = 0.79)



```
Constructing LSTM model with:

Number of stacked RNNs: 4
Number of hidden features: 50
Dropout: 0.5
Skip Connections
Gradient Clipping

Epoch 0 : Train loss 0.80 | Valid loss 0.72
Epoch 1 : Train loss 0.67 | Valid loss 0.71
Epoch 2 : Train loss 0.59 | Valid loss 0.75
```

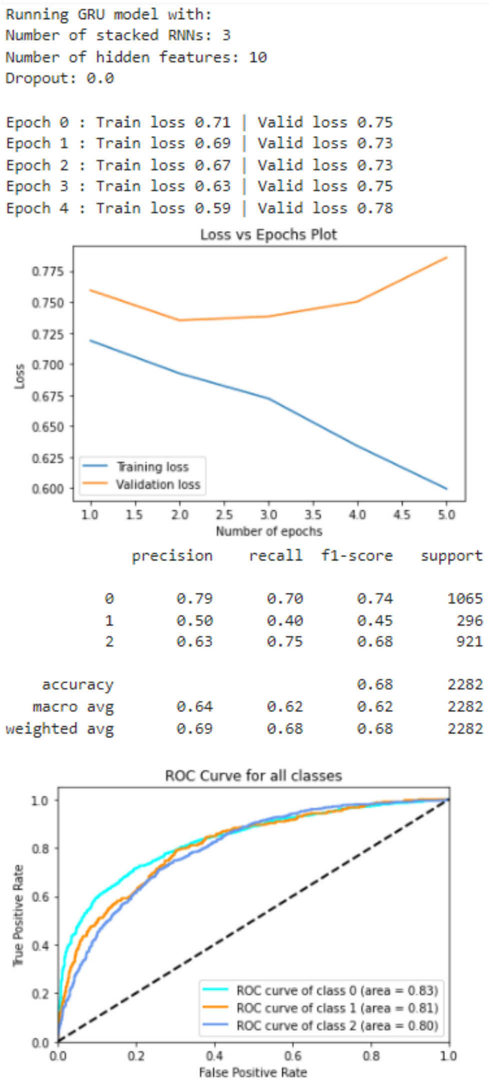
```
              precision    recall  f1-score   support

           0       0.81      0.76      0.78      1065
           1       0.55      0.39      0.46       296
           2       0.66      0.77      0.71       921

    accuracy                           0.72      2282
   macro avg       0.68      0.64      0.65      2282
weighted avg       0.72      0.72      0.71      2282
```

ROC Curve for all classes
ROC curve of class 0 (area = 0.87)
ROC curve of class 1 (area = 0.83)
ROC curve of class 2 (area = 0.82)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

1.7.5 GRU cells without Gradient Clipping and Skip Connections

Below are some runs using:

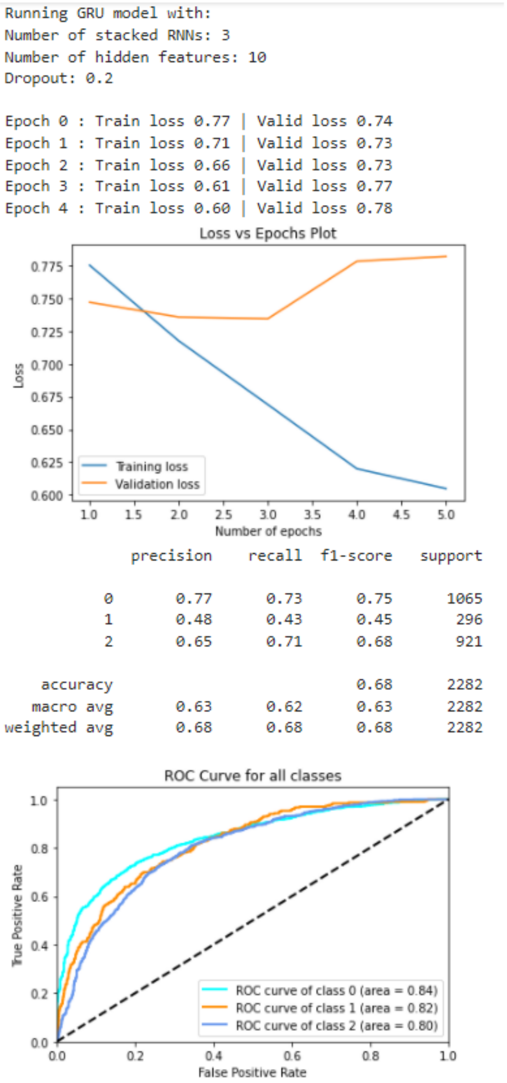
- Stacked GRU layers.
- No Gradient Clipping.
- No Skip Connections.



```
Running GRU model with:
Number of stacked RNNs: 3
Number of hidden features: 10
Dropout: 0.0

Epoch 0 : Train loss 0.71 | Valid loss 0.75
Epoch 1 : Train loss 0.69 | Valid loss 0.73
Epoch 2 : Train loss 0.67 | Valid loss 0.73
Epoch 3 : Train loss 0.63 | Valid loss 0.75
Epoch 4 : Train loss 0.59 | Valid loss 0.78
```

```
              precision    recall  f1-score   support

           0       0.79      0.70      0.74      1065
           1       0.50      0.40      0.45       296
           2       0.63      0.75      0.68       921

    accuracy                           0.68      2282
   macro avg       0.64      0.62      0.62      2282
weighted avg       0.69      0.68      0.68      2282
```

ROC Curve for all classes
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.83)
ROC curve of class 1 (area = 0.81)
ROC curve of class 2 (area = 0.80)



```
Running GRU model with:
Number of stacked RNNs: 3
Number of hidden features: 10
Dropout: 0.2

Epoch 0 : Train loss 0.77 | Valid loss 0.74
Epoch 1 : Train loss 0.71 | Valid loss 0.73
Epoch 2 : Train loss 0.66 | Valid loss 0.73
Epoch 3 : Train loss 0.61 | Valid loss 0.77
Epoch 4 : Train loss 0.60 | Valid loss 0.78
```

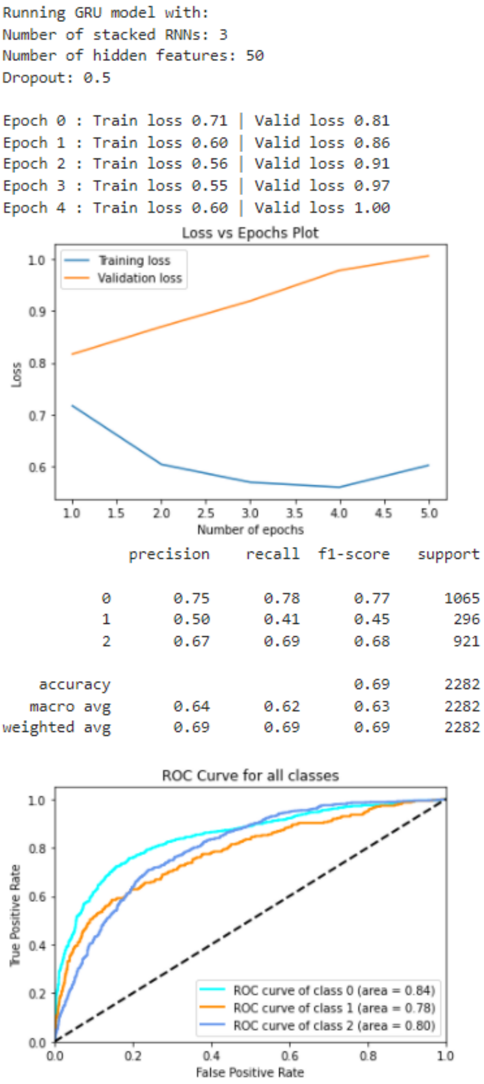
```
              precision    recall  f1-score   support

           0       0.77      0.73      0.75      1065
           1       0.48      0.43      0.45       296
           2       0.65      0.71      0.68       921

    accuracy                           0.68      2282
   macro avg       0.63      0.62      0.63      2282
weighted avg       0.68      0.68      0.68      2282
```

ROC Curve for all classes
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.84)
ROC curve of class 1 (area = 0.82)
ROC curve of class 2 (area = 0.80)

1.7.6 GRU cells with Gradient Clipping and no Skip Connections

Below are some runs using:

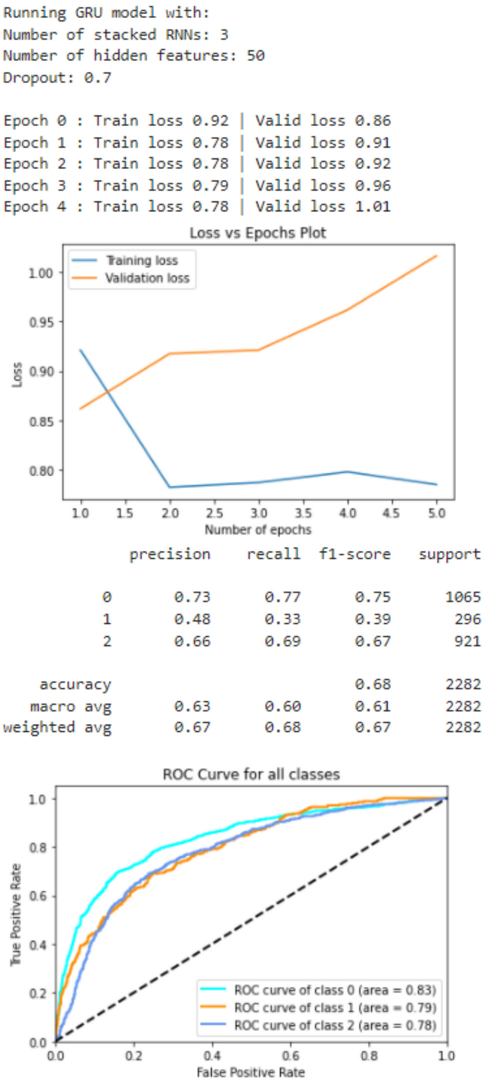
- Stacked GRU layers.
- Gradient Clipping (max norm = 3).
- No Skip Connections.



```
Running GRU model with:
Number of stacked RNNs: 3
Number of hidden features: 50
Dropout: 0.5

Epoch 0 : Train loss 0.71 | Valid loss 0.81
Epoch 1 : Train loss 0.60 | Valid loss 0.86
Epoch 2 : Train loss 0.56 | Valid loss 0.91
Epoch 3 : Train loss 0.55 | Valid loss 0.97
Epoch 4 : Train loss 0.60 | Valid loss 1.00
```

```
              precision    recall  f1-score   support

           0       0.75      0.78      0.77      1065
           1       0.50      0.41      0.45       296
           2       0.67      0.69      0.68       921

    accuracy                           0.69      2282
   macro avg       0.64      0.62      0.63      2282
weighted avg       0.69      0.69      0.69      2282
```

ROC Curve for all classes
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.84)
ROC curve of class 1 (area = 0.78)
ROC curve of class 2 (area = 0.80)



```
Running GRU model with:
Number of stacked RNNs: 3
Number of hidden features: 50
Dropout: 0.7

Epoch 0 : Train loss 0.92 | Valid loss 0.86
Epoch 1 : Train loss 0.78 | Valid loss 0.91
Epoch 2 : Train loss 0.78 | Valid loss 0.92
Epoch 3 : Train loss 0.79 | Valid loss 0.96
Epoch 4 : Train loss 0.78 | Valid loss 1.01
```

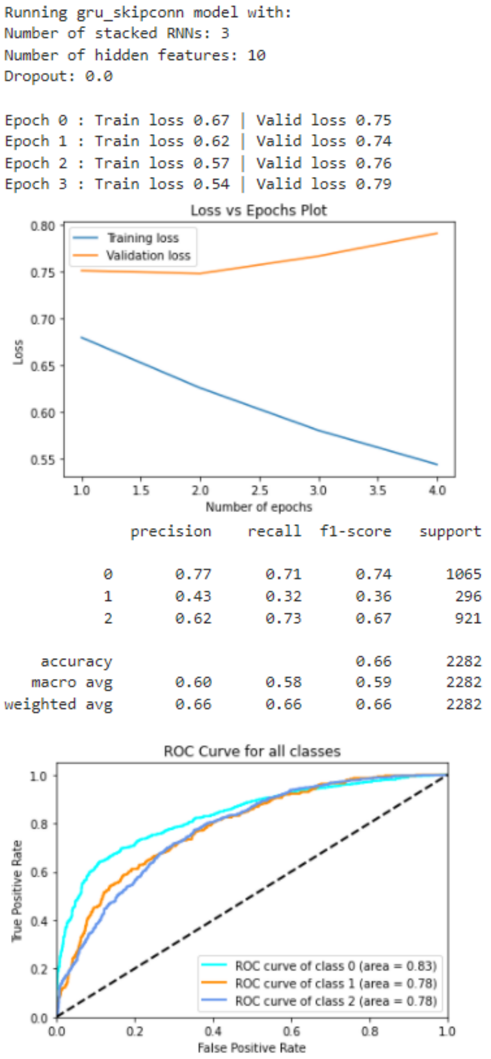
```
              precision    recall  f1-score   support

           0       0.73      0.77      0.75      1065
           1       0.48      0.33      0.39       296
           2       0.66      0.69      0.67       921

    accuracy                           0.68      2282
   macro avg       0.63      0.60      0.61      2282
weighted avg       0.67      0.68      0.67      2282
```

ROC Curve for all classes
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.83)
ROC curve of class 1 (area = 0.79)
ROC curve of class 2 (area = 0.78)

1.7.7 GRU cells without Gradient Clipping but with Skip Connections

Below are some runs using:

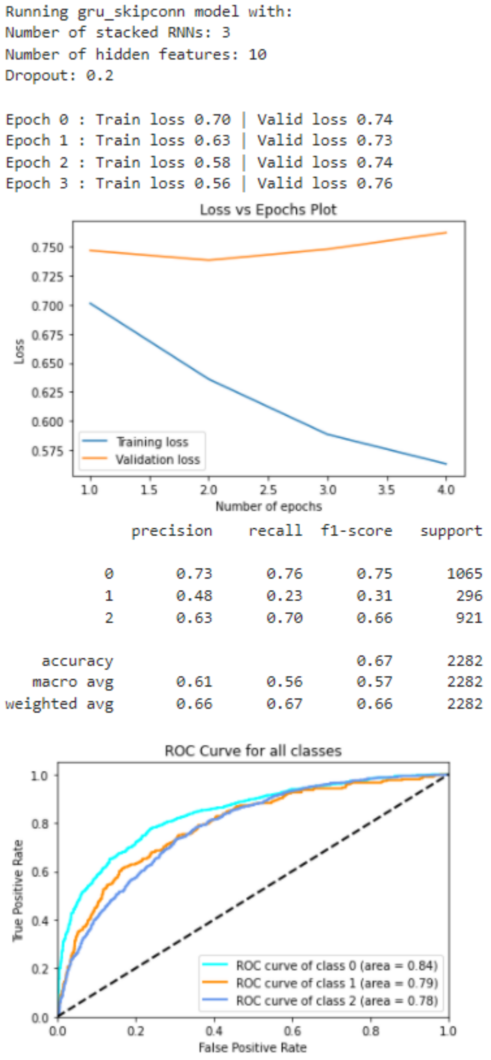
- Stacked GRU layers.
- No Gradient Clipping.
- Skip Connections.



```
Running gru_skipconn model with:
Number of stacked RNNs: 3
Number of hidden features: 10
Dropout: 0.0

Epoch 0 : Train loss 0.67 | Valid loss 0.75
Epoch 1 : Train loss 0.62 | Valid loss 0.74
Epoch 2 : Train loss 0.57 | Valid loss 0.76
Epoch 3 : Train loss 0.54 | Valid loss 0.79
```

```
              precision    recall  f1-score   support

           0       0.77      0.71      0.74      1065
           1       0.43      0.32      0.36       296
           2       0.62      0.73      0.67       921

    accuracy                           0.66      2282
   macro avg       0.60      0.58      0.59      2282
weighted avg       0.66      0.66      0.66      2282
```

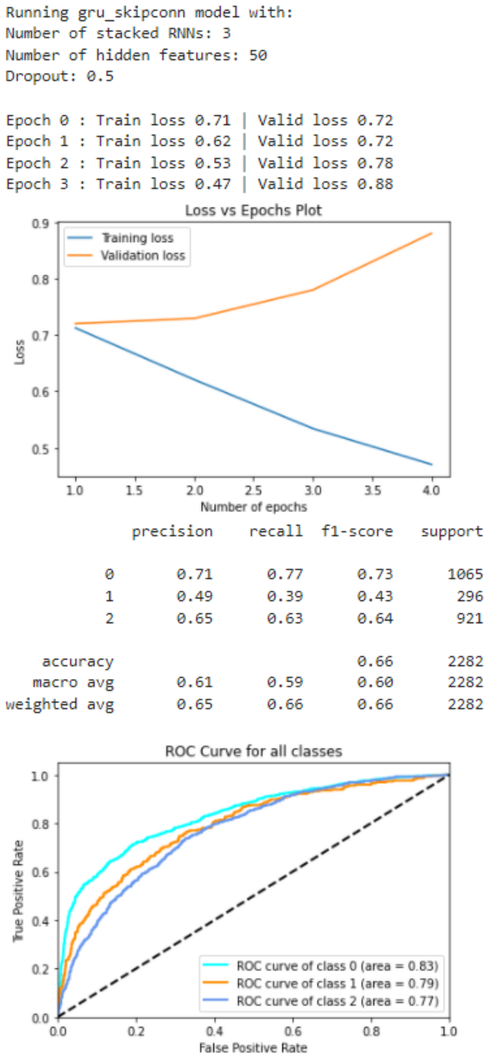
ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.83)
ROC curve of class 1 (area = 0.78)
ROC curve of class 2 (area = 0.78)



```
Running gru_skipconn model with:
Number of stacked RNNs: 3
Number of hidden features: 10
Dropout: 0.2

Epoch 0 : Train loss 0.70 | Valid loss 0.74
Epoch 1 : Train loss 0.63 | Valid loss 0.73
Epoch 2 : Train loss 0.58 | Valid loss 0.74
Epoch 3 : Train loss 0.56 | Valid loss 0.76
```

```
              precision    recall  f1-score   support

           0       0.73      0.76      0.75      1065
           1       0.48      0.23      0.31       296
           2       0.63      0.70      0.66       921

    accuracy                           0.67      2282
   macro avg       0.61      0.56      0.57      2282
weighted avg       0.66      0.67      0.66      2282
```

ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.84)
ROC curve of class 1 (area = 0.79)
ROC curve of class 2 (area = 0.78)



```
Running gru_skipconn model with:
Number of stacked RNNs: 3
Number of hidden features: 50
Dropout: 0.5

Epoch 0 : Train loss 0.71 | Valid loss 0.72
Epoch 1 : Train loss 0.62 | Valid loss 0.72
Epoch 2 : Train loss 0.53 | Valid loss 0.78
Epoch 3 : Train loss 0.47 | Valid loss 0.88
```

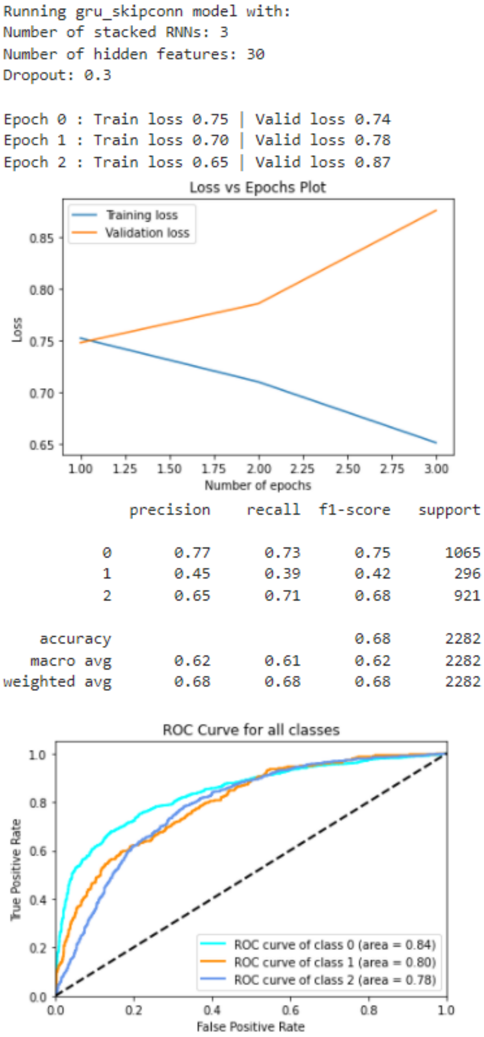
```
              precision    recall  f1-score   support

           0       0.71      0.77      0.73      1065
           1       0.49      0.39      0.43       296
           2       0.65      0.63      0.64       921

    accuracy                           0.66      2282
   macro avg       0.61      0.59      0.60      2282
weighted avg       0.65      0.66      0.66      2282
```

ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.83)
ROC curve of class 1 (area = 0.79)
ROC curve of class 2 (area = 0.77)

1.7.8 GRU cells with both Gradient Clipping and Skip Connections

Below are some runs using:

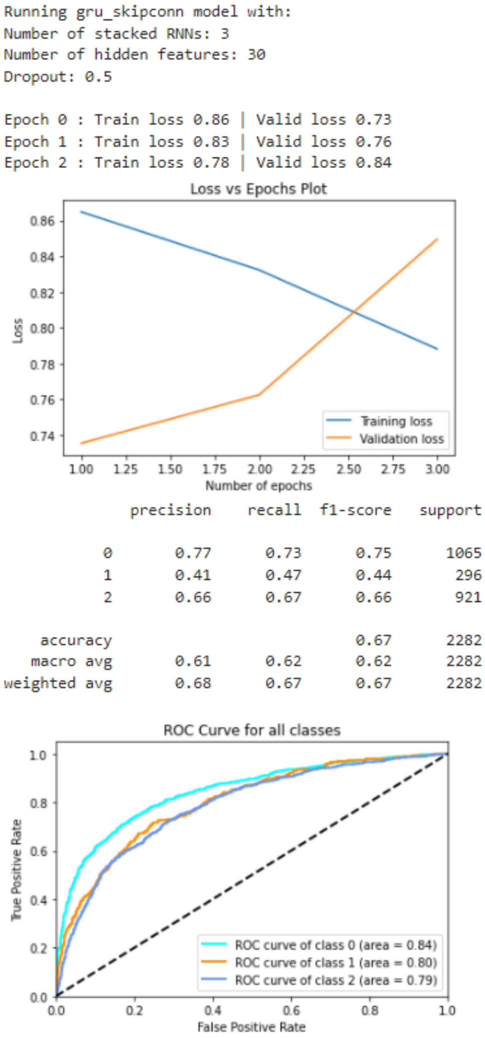
- Stacked GRU layers.
- Gradient Clipping.
- Skip Connections.



```
Running gru_skipconn model with:
Number of stacked RNNs: 3
Number of hidden features: 30
Dropout: 0.3

Epoch 0 : Train loss 0.75 | Valid loss 0.74
Epoch 1 : Train loss 0.70 | Valid loss 0.78
Epoch 2 : Train loss 0.65 | Valid loss 0.87
```

```
              precision    recall  f1-score   support

           0       0.77      0.73      0.75      1065
           1       0.45      0.39      0.42       296
           2       0.65      0.71      0.68       921

    accuracy                           0.68      2282
   macro avg       0.62      0.61      0.62      2282
weighted avg       0.68      0.68      0.68      2282
```

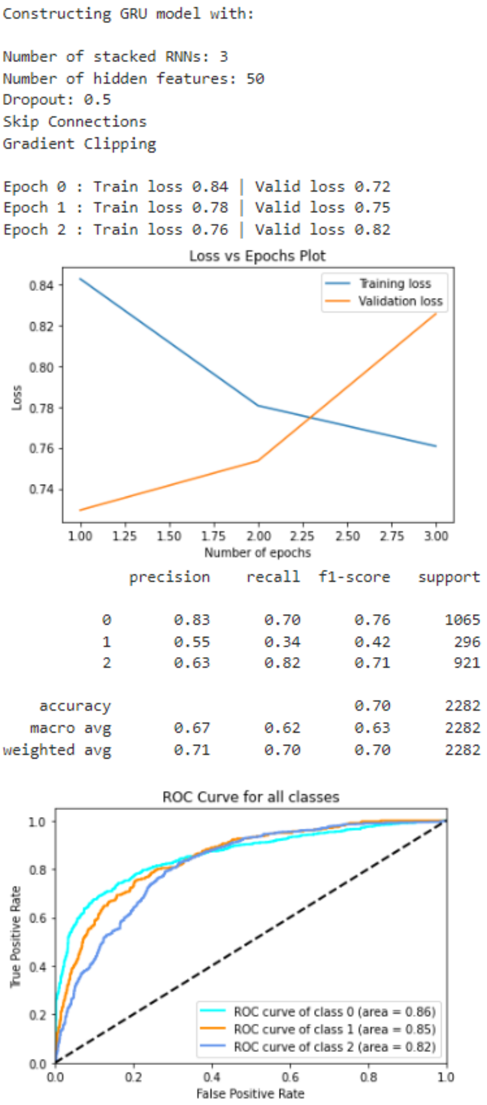
ROC Curve for all classes
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
ROC curve of class 0 (area = 0.84)
ROC curve of class 1 (area = 0.80)
ROC curve of class 2 (area = 0.78)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate



```
Running gru_skipconn model with:
Number of stacked RNNs: 3
Number of hidden features: 30
Dropout: 0.5

Epoch 0 : Train loss 0.86 | Valid loss 0.73
Epoch 1 : Train loss 0.83 | Valid loss 0.76
Epoch 2 : Train loss 0.78 | Valid loss 0.84
```

```
              precision    recall  f1-score   support

           0       0.77      0.73      0.75      1065
           1       0.41      0.47      0.44       296
           2       0.66      0.67      0.66       921

    accuracy                           0.67      2282
   macro avg       0.61      0.62      0.62      2282
weighted avg       0.68      0.67      0.67      2282
```

ROC Curve for all classes
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
ROC curve of class 0 (area = 0.84)
ROC curve of class 1 (area = 0.80)
ROC curve of class 2 (area = 0.79)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate



```
Constructing GRU model with:

Number of stacked RNNs: 3
Number of hidden features: 50
Dropout: 0.5
Skip Connections
Gradient Clipping

Epoch 0 : Train loss 0.84 | Valid loss 0.72
Epoch 1 : Train loss 0.78 | Valid loss 0.75
Epoch 2 : Train loss 0.76 | Valid loss 0.82
```

```
              precision    recall  f1-score   support

           0       0.83      0.70      0.76      1065
           1       0.55      0.34      0.42       296
           2       0.63      0.82      0.71       921

    accuracy                           0.70      2282
   macro avg       0.67      0.62      0.63      2282
weighted avg       0.71      0.70      0.70      2282
```

ROC Curve for all classes
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.86)
ROC curve of class 1 (area = 0.85)
ROC curve of class 2 (area = 0.82)

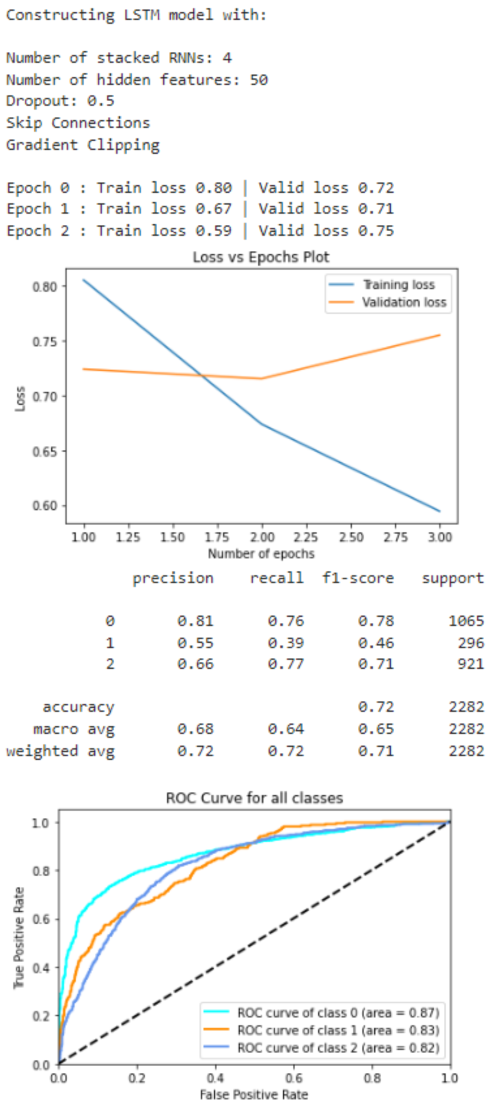
1.8 Best Performing Models

Based on all the testing, tuning and experimenting, the overall **best performing configurations** for each cell type are:

1.8.1 LSTM

- 4 stacked LSTM layers.
- 50 hidden features in each layer.
- Skip Connections.

- Gradient Clipping (max norm = 3).
- 0.5 dropout ratio.



```
Constructing LSTM model with:

Number of stacked RNNs: 4
Number of hidden features: 50
Dropout: 0.5
Skip Connections
Gradient Clipping

Epoch 0 : Train loss 0.80 | Valid loss 0.72
Epoch 1 : Train loss 0.67 | Valid loss 0.71
Epoch 2 : Train loss 0.59 | Valid loss 0.75
```

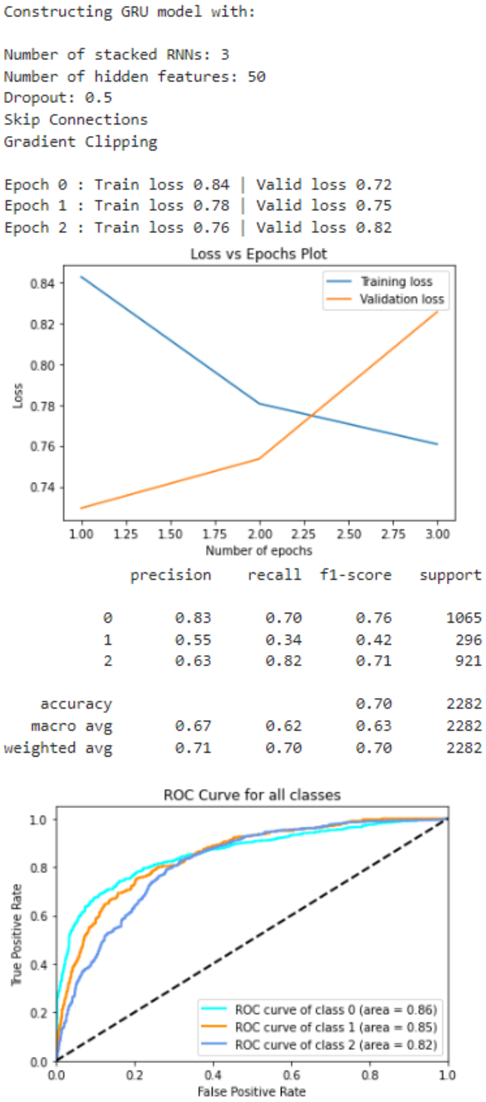
```
              precision    recall  f1-score   support

           0       0.81      0.76      0.78      1065
           1       0.55      0.39      0.46       296
           2       0.66      0.77      0.71       921

    accuracy                           0.72      2282
   macro avg       0.68      0.64      0.65      2282
weighted avg       0.72      0.72      0.71      2282
```

ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.87)
ROC curve of class 1 (area = 0.83)
ROC curve of class 2 (area = 0.82)

1.8.2 GRU

- 3 stacked GRU layers.
- 50 hidden features in each layer.
- Skip Connections.
- Gradient Clipping (max norm = 3).

- 0.5 dropout ratio.



```
Constructing GRU model with:

Number of stacked RNNs: 3
Number of hidden features: 50
Dropout: 0.5
Skip Connections
Gradient Clipping

Epoch 0 : Train loss 0.84 | Valid loss 0.72
Epoch 1 : Train loss 0.78 | Valid loss 0.75
Epoch 2 : Train loss 0.76 | Valid loss 0.82
```

```
              precision    recall  f1-score   support

           0       0.83      0.70      0.76      1065
           1       0.55      0.34      0.42       296
           2       0.63      0.82      0.71       921

    accuracy                           0.70      2282
   macro avg       0.67      0.62      0.63      2282
weighted avg       0.71      0.70      0.70      2282
```

ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.86)
ROC curve of class 1 (area = 0.85)
ROC curve of class 2 (area = 0.82)

We can see that LSTM model scored a weighted average score (precision, recall and f1-score) of 72%, 72% and 71% respectively, while the GRU one scored a little lower on 71%, 70% and 70%. Although the LSTM model had an overall better fit in the ROC curve, both models produced pretty truthful results for all 3 classes. However, both models tend to overfit, so the ideal number of epochs in order for them to avoid excessive overfitting, is around 3 epochs of training.

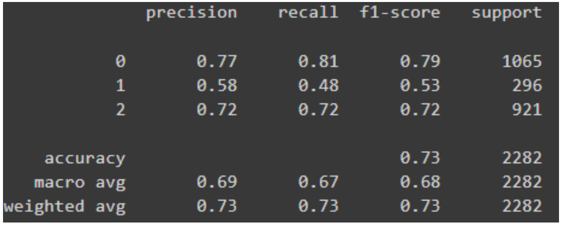
1.9 Comparison with Feed Forward Neural Networks

Feed-forward Networks allow signals to travel one way only, from input to output. There are no feedback (loops); i.e., the output of any layer does not affect that same layer. FFNNs tend to be straightforward networks that associate inputs with outputs. They are extensively used in pattern recognition. This type of organisation is also referred to as bottom-up or top-down.

Recurrent Neural Networks can have signals traveling in both directions by introducing loops in the network. Computations derived from earlier input are fed back into the network, which gives them a kind of memory. Feedback networks are dynamic; their 'state' is changing continuously until they reach an equilibrium point. They remain at the equilibrium point until the input changes and a new equilibrium needs to be found.

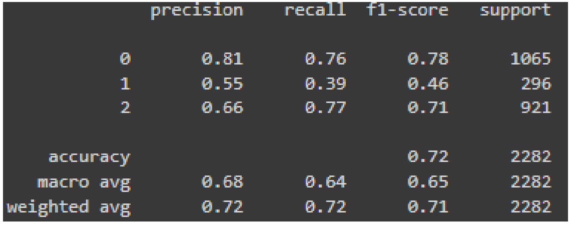
Compared to the best model of the previous exercise, current RNNs didn't score as high as FFNN model implemented with Bag of Words (BoW) method, but their results are pretty comparable to each other:

Feed Forward Neural Network



```
              precision    recall  f1-score   support

           0       0.77      0.81      0.79      1065
           1       0.58      0.48      0.53       296
           2       0.72      0.72      0.72       921

    accuracy                           0.73      2282
   macro avg       0.69      0.67      0.68      2282
weighted avg       0.73      0.73      0.73      2282
```

Recurrent Neural Network

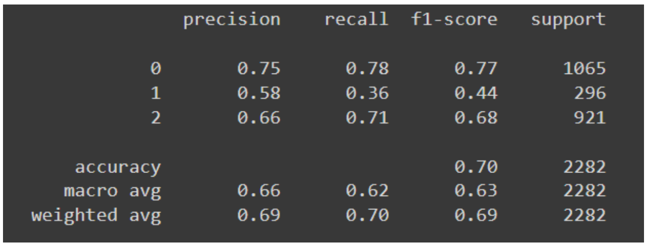


```
              precision    recall  f1-score   support

           0       0.81      0.76      0.78      1065
           1       0.55      0.39      0.46       296
           2       0.66      0.77      0.71       921

    accuracy                           0.72      2282
   macro avg       0.68      0.64      0.65      2282
weighted avg       0.72      0.72      0.71      2282
```

1.10 Comparison with Softmax Regression

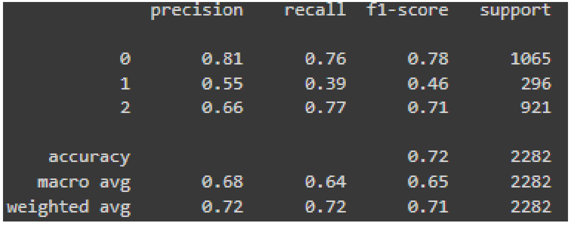
Neural networks are somewhat related to logistic regression. Basically, we can think of logistic regression as a one layer neural network. In fact, it is very common to use logistic sigmoid functions as activation functions in the hidden layer of a neural network.

Softmax Classifier



```
              precision    recall  f1-score   support

           0       0.75      0.78      0.77      1065
           1       0.58      0.36      0.44       296
           2       0.66      0.71      0.68       921

    accuracy                           0.70      2282
   macro avg       0.66      0.62      0.63      2282
weighted avg       0.69      0.70      0.69      2282
```

Recurrent Neural Network



```
              precision    recall  f1-score   support

           0       0.81      0.76      0.78      1065
           1       0.55      0.39      0.46       296
           2       0.66      0.77      0.71       921

    accuracy                           0.72      2282
   macro avg       0.68      0.64      0.65      2282
weighted avg       0.72      0.72      0.71      2282
```

In this case, we can see that the new RNN model performs better than the Softmax Classifier at all weighted avg. metrics.

2 Attention

In psychology, attention is the cognitive process of selectively concentrating on one or a few things while ignoring others.

A neural network is considered to be an effort to mimic human brain actions in a simplified manner. Attention Mechanism is also an attempt to implement the same action of selectively concentrating on a few relevant things, while ignoring others in deep neural networks. That way, the model learns to pay selective attention to some inputs and relate them to items in the output sequences.

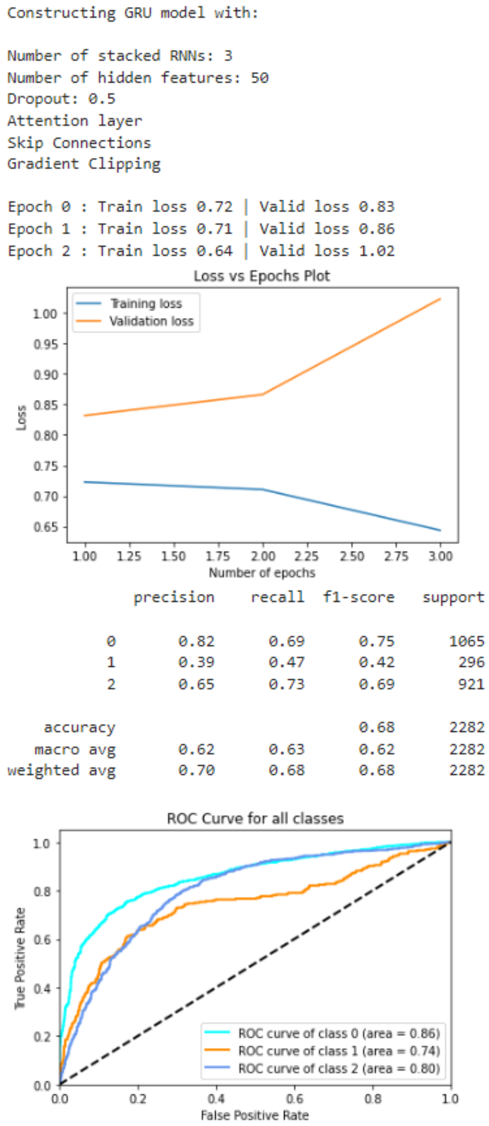
Unfortunately, attention does not always improve the models' overall scores. Actually, for the current implementation, attention mechanism not only did not increase the models' scores, but it made them slightly worse than they were before:

2.1 Attention to the best GRU model

Observing the diagram below, it is obvious that the attention mechanism reduced the model's weighted avg. precision, recall and f1-score. It is also worth mentioning, that accuracy went from 70% to 68%.

When looking to the loss-vs-epochs diagram, we can see that the model is prone to overfitting, so it would be safer to train it for a smaller amount of epochs.

The model's fit in the ROC diagram has a relatively worse fit compared to the previous one.



```
Constructing GRU model with:

Number of stacked RNNs: 3
Number of hidden features: 50
Dropout: 0.5
Attention layer
Skip Connections
Gradient Clipping

Epoch 0 : Train loss 0.72 | Valid loss 0.83
Epoch 1 : Train loss 0.71 | Valid loss 0.86
Epoch 2 : Train loss 0.64 | Valid loss 1.02
```

```
              precision    recall  f1-score   support

           0       0.82      0.69      0.75      1065
           1       0.39      0.47      0.42       296
           2       0.65      0.73      0.69       921

    accuracy                           0.68      2282
   macro avg       0.62      0.63      0.62      2282
weighted avg       0.70      0.68      0.68      2282
```

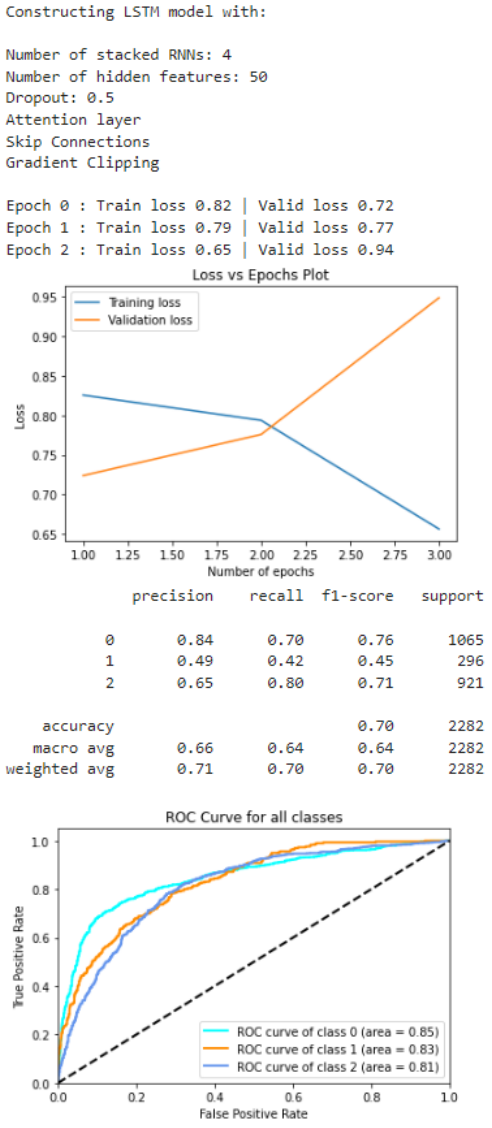
ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.86)
ROC curve of class 1 (area = 0.74)
ROC curve of class 2 (area = 0.80)

2.2 Attention to the best LSTM model

When applying attention to the best performing LSTM model, we can also see a decrease in performance by looking at weighted avg. precision, recall and f1-score. All the aforementioned metrics went from 72%, 72%, 71% to 71%, 70%, 70%. Similarly, accuracy was increased from 72% to 70%.

When looking to the loss-vs-epochs diagram, we can see that the model is prone to overfitting, but not as much as the GRU model.

The model's fit in the ROC diagram has a relatively worse fit compared to the previous one.



```
Constructing LSTM model with:

Number of stacked RNNs: 4
Number of hidden features: 50
Dropout: 0.5
Attention layer
Skip Connections
Gradient Clipping

Epoch 0 : Train loss 0.82 | Valid loss 0.72
Epoch 1 : Train loss 0.79 | Valid loss 0.77
Epoch 2 : Train loss 0.65 | Valid loss 0.94
```

```
              precision    recall  f1-score   support

           0       0.84      0.70      0.76      1065
           1       0.49      0.42      0.45       296
           2       0.65      0.80      0.71       921

    accuracy                           0.70      2282
   macro avg       0.66      0.64      0.64      2282
weighted avg       0.71      0.70      0.70      2282
```

ROC Curve for all classes
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve of class 0 (area = 0.85)
ROC curve of class 1 (area = 0.83)
ROC curve of class 2 (area = 0.81)

3 Final Choice

Finally, with all that said, the **overall best performing model** is a Word Embeddings LSTM model, consisting of:

- GloVe's pretrained 200-dimensional twitter vectors.

- 4 stacked LSTM layers.
- 50 hidden features in each layer.
- Skip Connections.
- Gradient Clipping (max norm = 3).
- No Attention mechanism.
- 0.5 dropout ratio
- Adam optimizer (learning rate = 0.001).
- Cross Entropy loss function.
- 3 epochs of training.

4 Sources

- Deal With Imbalanced Data
- Pytorch Manual
- General Info