Detection of Sexist Statements Against Women

By: Merna Abdelbadie - 900203731
Zein Firas Noureddin - 900212591

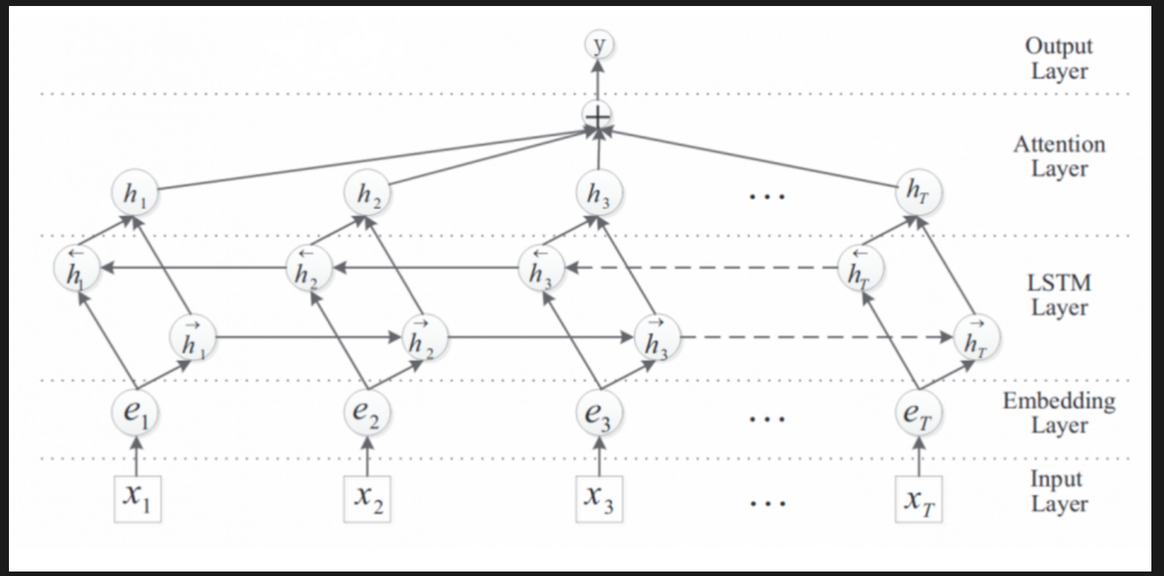
Problem Statement

“Detection of Sexist Statements Against Women”

Original Model

Automatic Detection of Sexist Statements Commonly Used at the Workplace

Bidirectional LSTM + Attention



Our Proposed Updates

Attempt to integrate an SVM classifier with the LSTM model, try GRU layers instead of LSTM, experiment with adding a CNN layer before the LSTM layers.

Use a bigger dataset to achieve better results

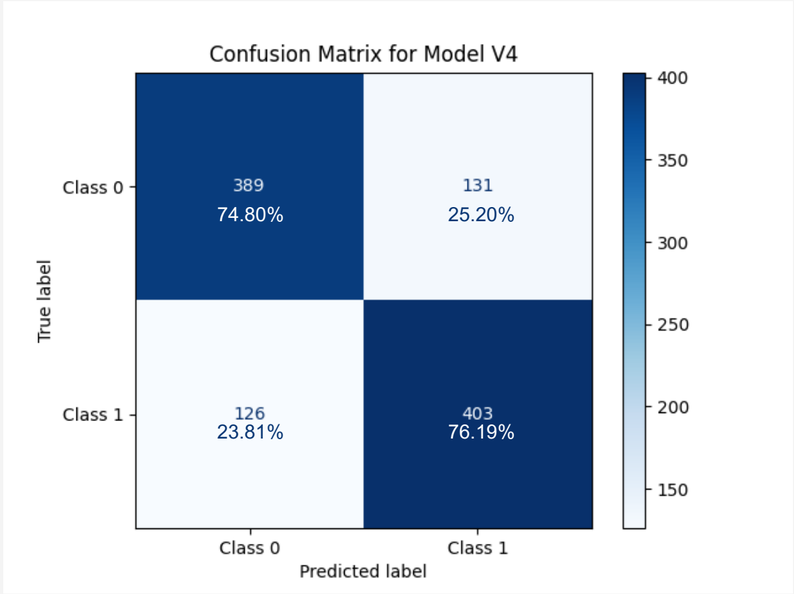
Work with dataset by modifying the model to enhance the performance

Use DataAugumentation and add user feedback loop

Using fine-tuning by training the model on general hate speech text and then training it on sexist speech

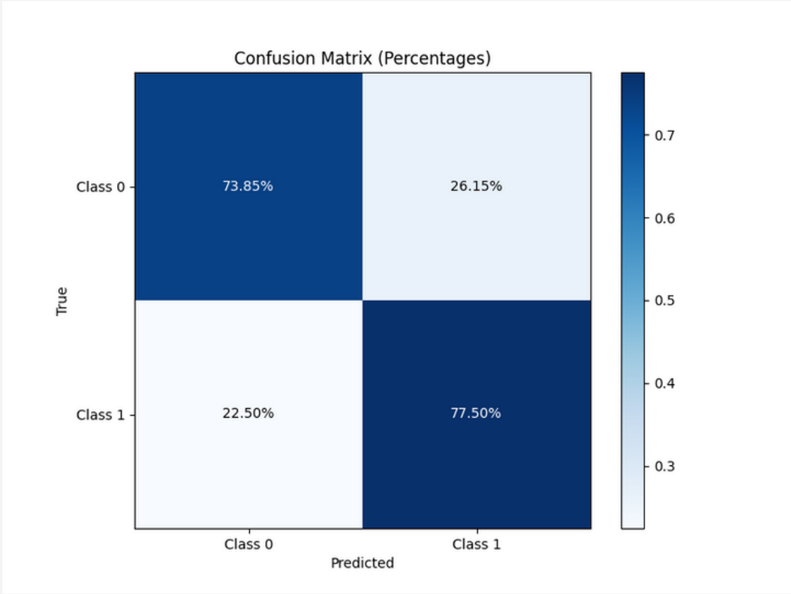
First Milestone

Original Model

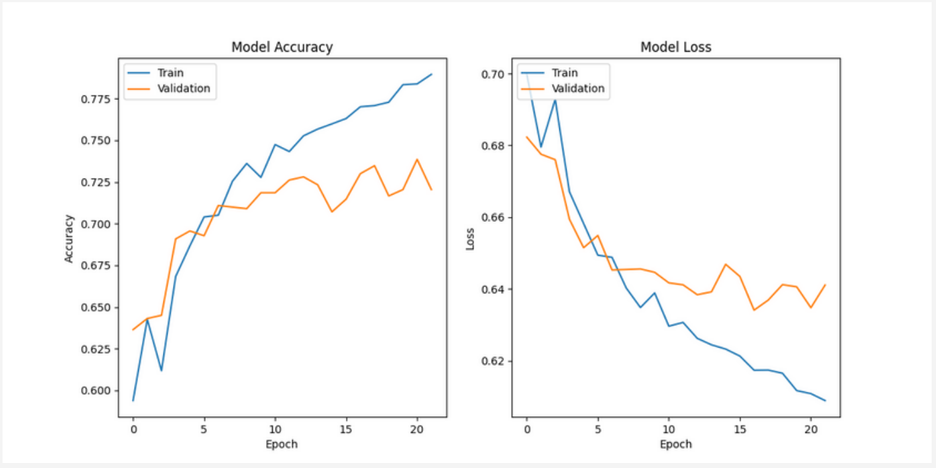


Bidirectional LSTM using 128, 128 neurons

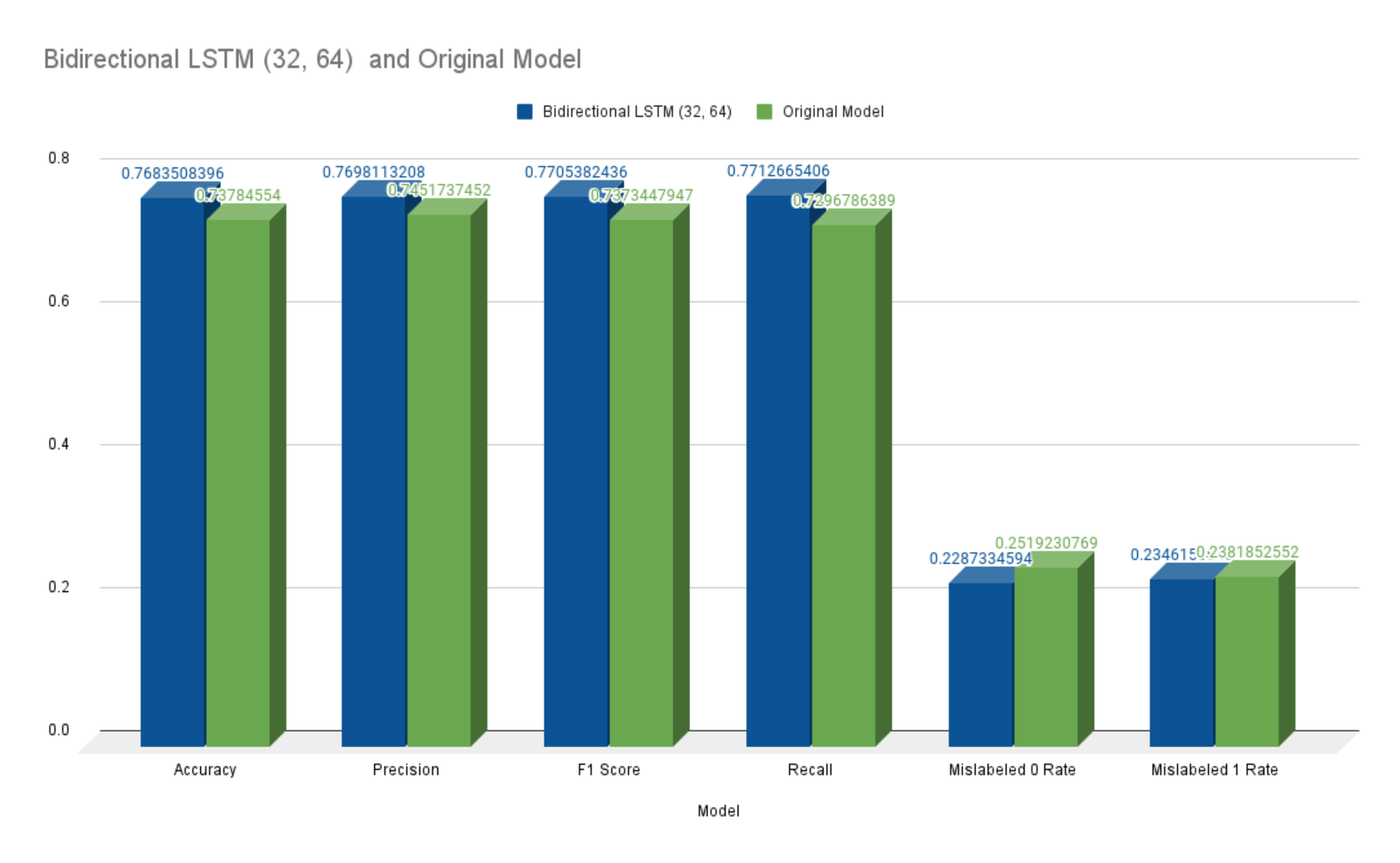
Best Model



Bidirectional LSTM using 32, 64 neurons

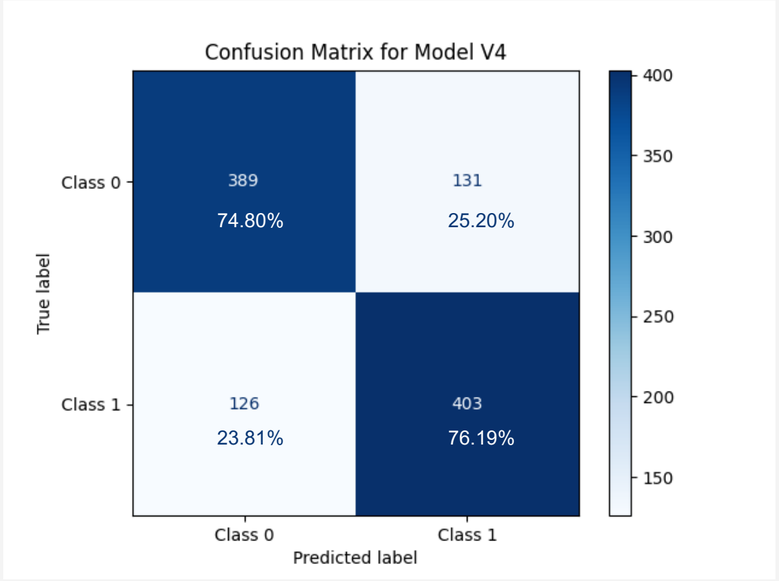


Other Metrics

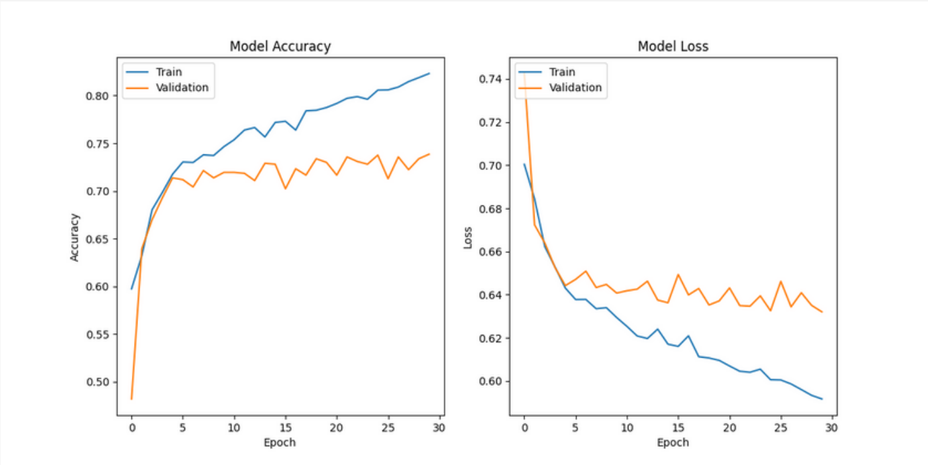


Second Milestone

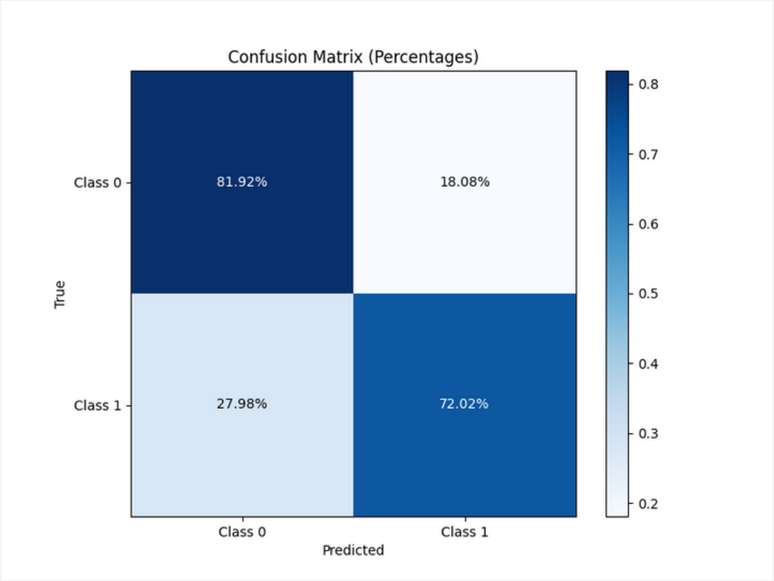
Original Model



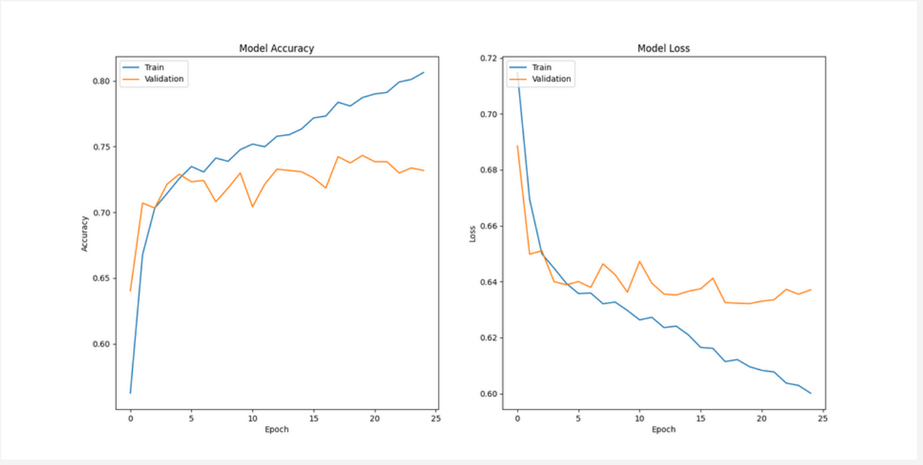
Bidirectional LSTM using 128, 128 neurons



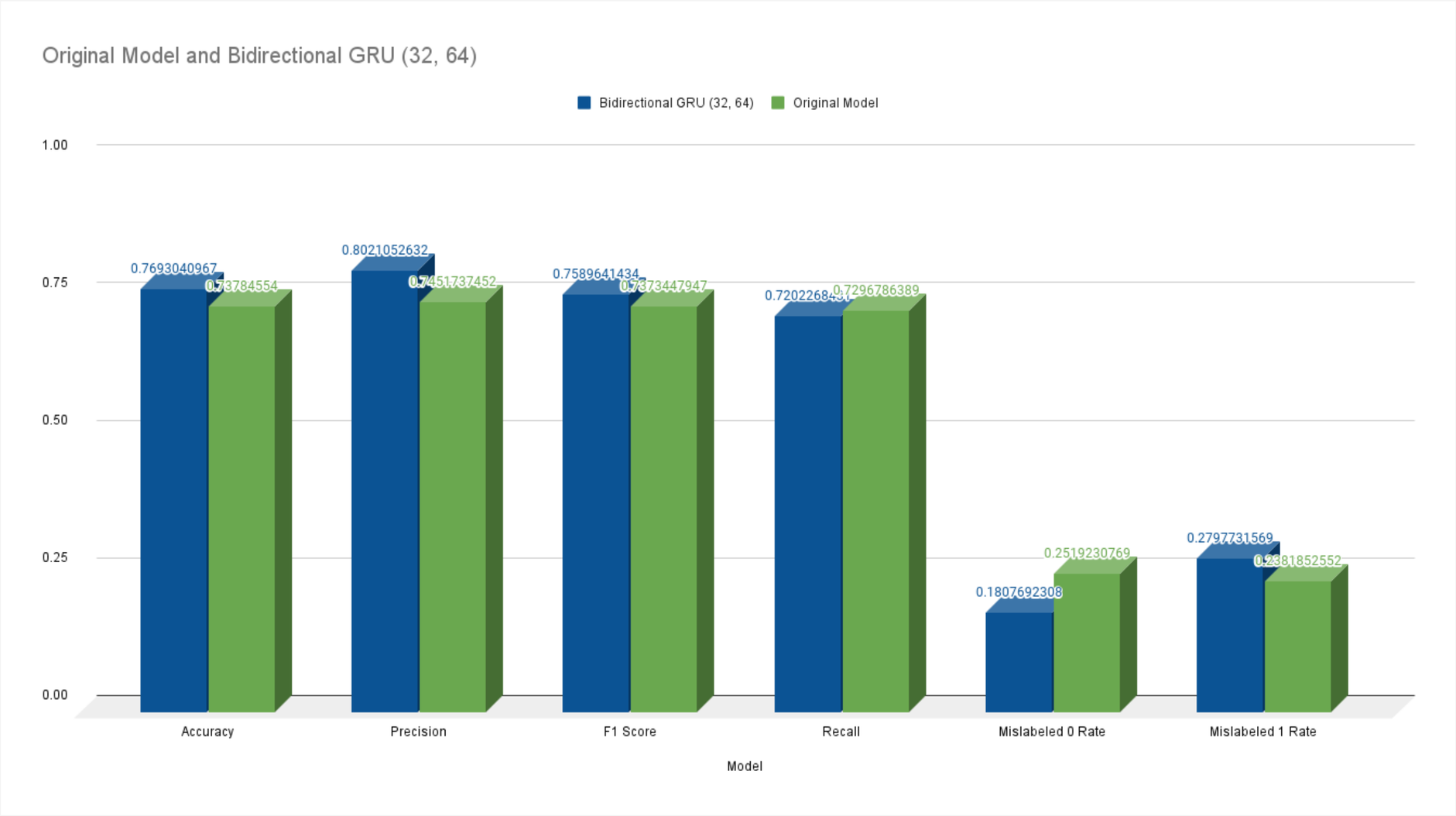
Best Model



Bidirectional GRU using 32, 64 neurons



Other Metrics



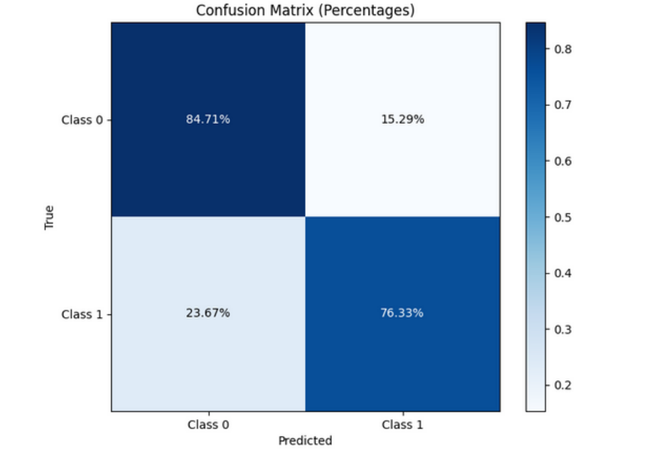
Final Milestone

Milestone 3 progress

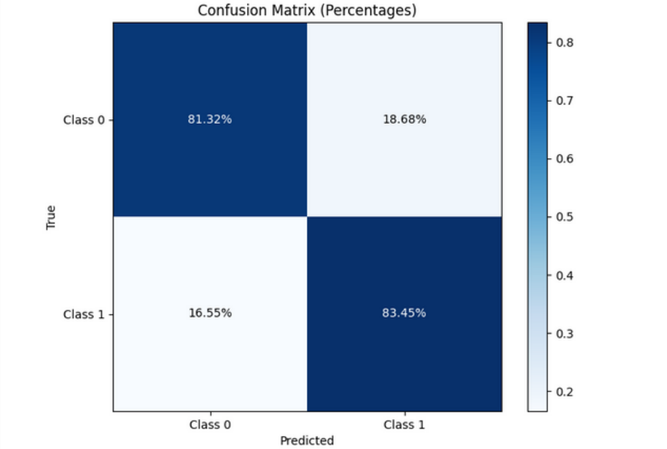
- Part 1:
  - Used data augmentation as a form of regularization to increase the accuracy of the model:
    - Increased the dataset from 10,000 statements to almost 40,000 statements.
    - Trained the model using first hate speech, then using the augmented dataset to fine-tune the model.
    - Trained the model without fine-tuning while adjusting number of epochs and batch sizes
- Part 2:
  - Added user feedback loop where when the user is testing the model, they can input the correct output and if the model mispredicts, it appends the mispredicted statements to the dataset and retrains.

Confusion Matrix

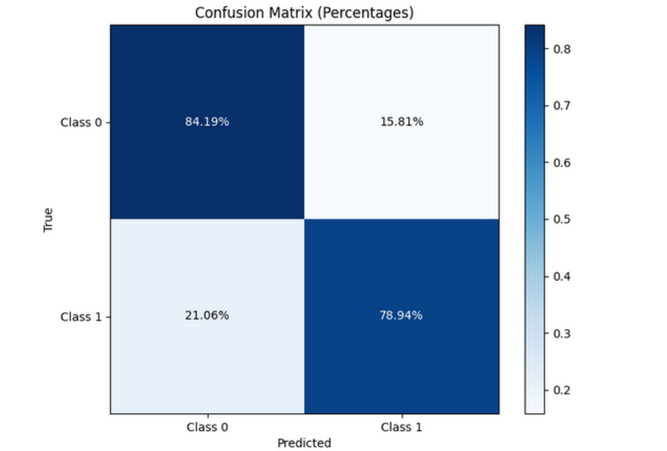
GRU model, fine-tuning, 30 epoch, early stopping, batch size 32, fine-tuning



GRU model, 30 epoch, early stopping, batch size 32

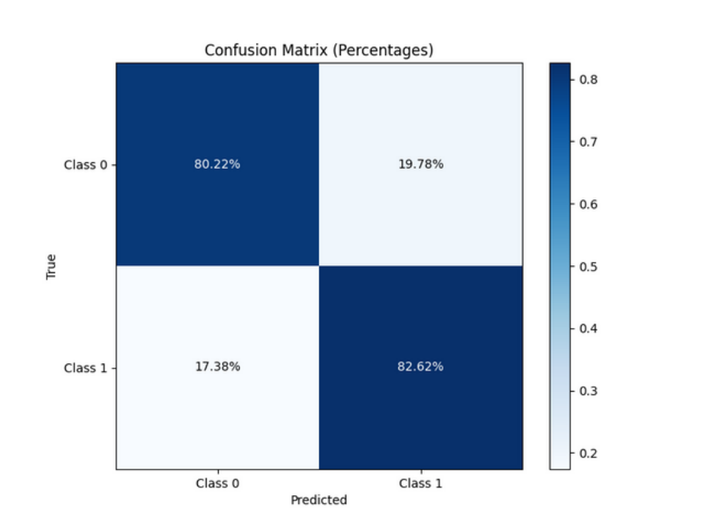


GRU model, 100 epoch, early stopping, batch size 8

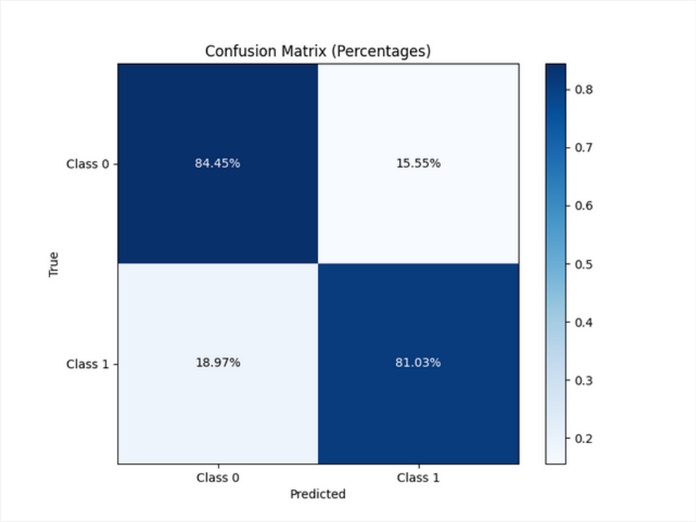


Confusion Matrix

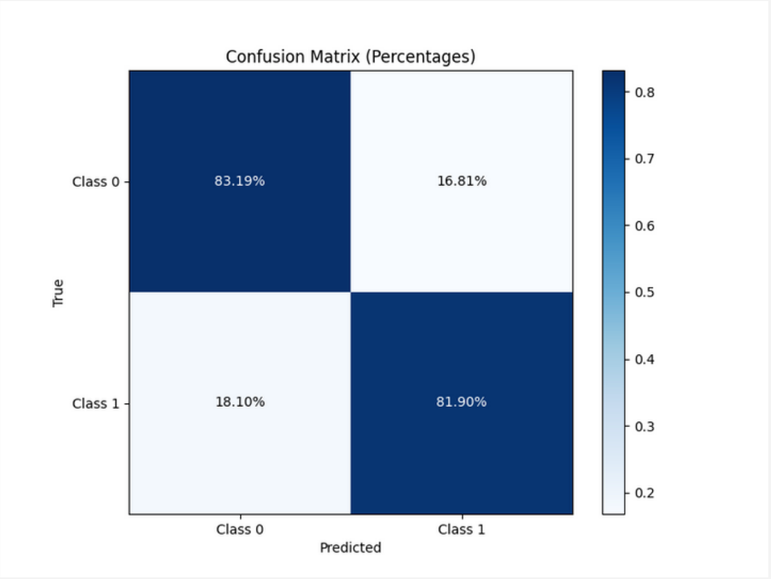
GRU model, 100 epoch, early stopping, batch size 16



GRU model, 100 epoch, early stopping, batch size 32

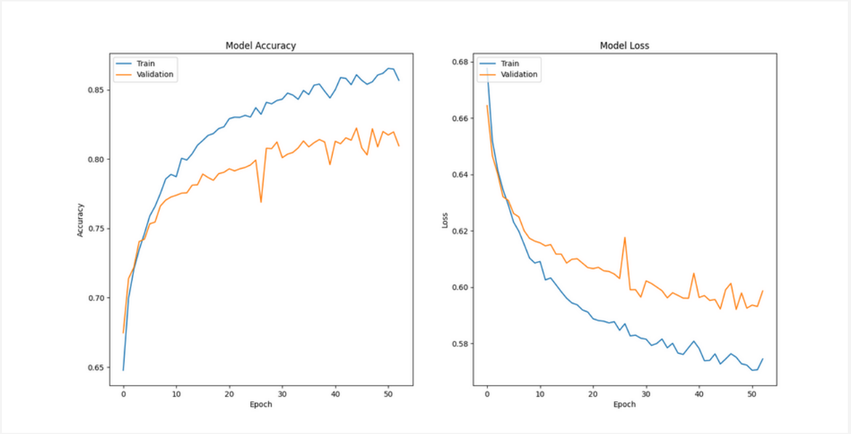


GRU model, 100 epoch, early stopping, batch size 64

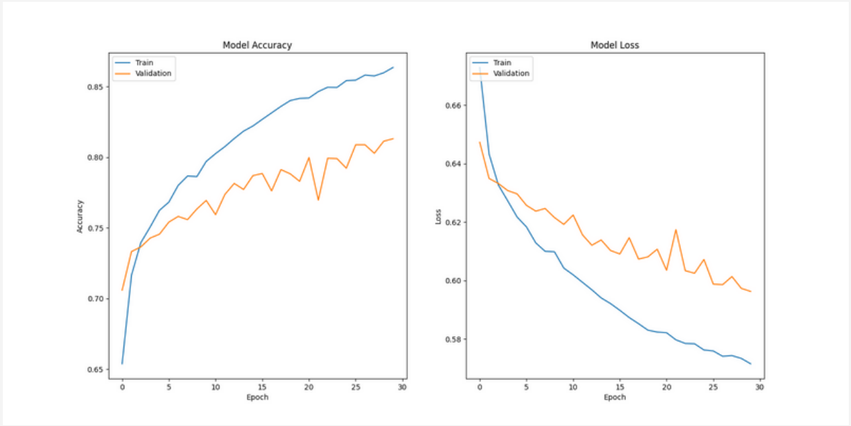


Accuracy & Loss

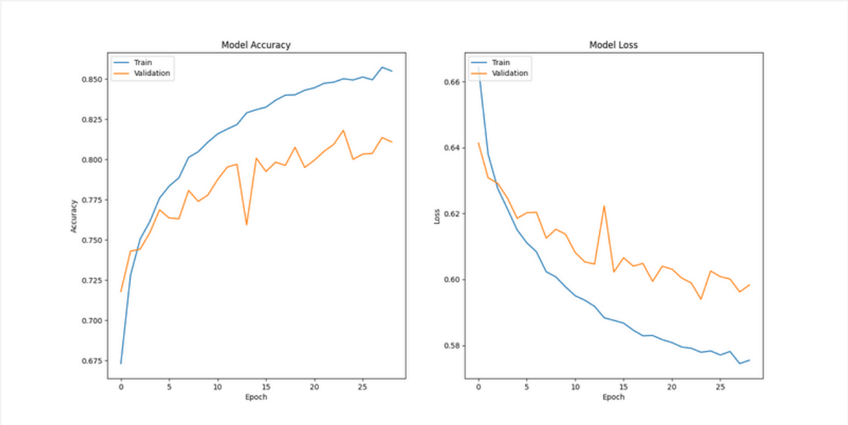
GRU model, fine-tuning, 30 epoch, early stopping, batch size 32, fine-tuning



GRU model, 30 epoch, early stopping, batch size 32

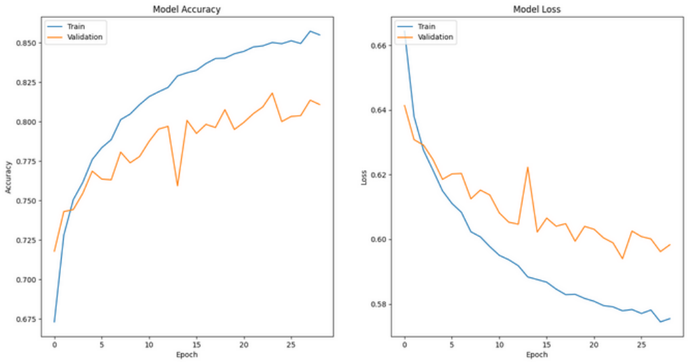


GRU model, 100 epoch, early stopping, batch size 8

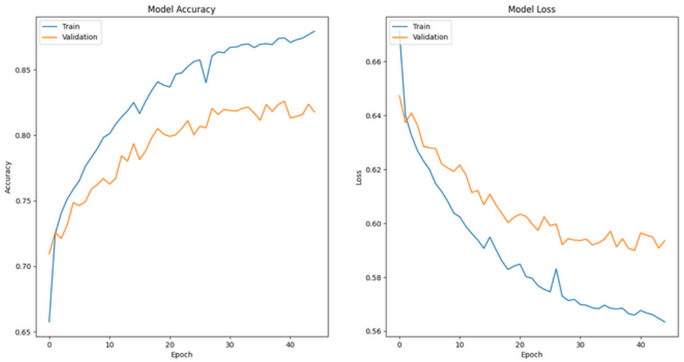


Accuracy & Loss

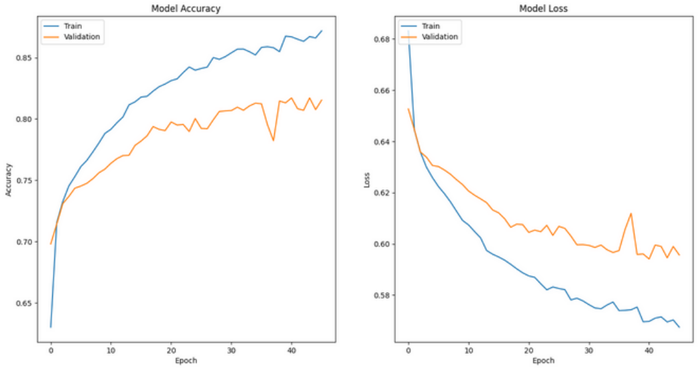
GRU model, fine-tuning, 100 epoch, early stopping, batch size 16

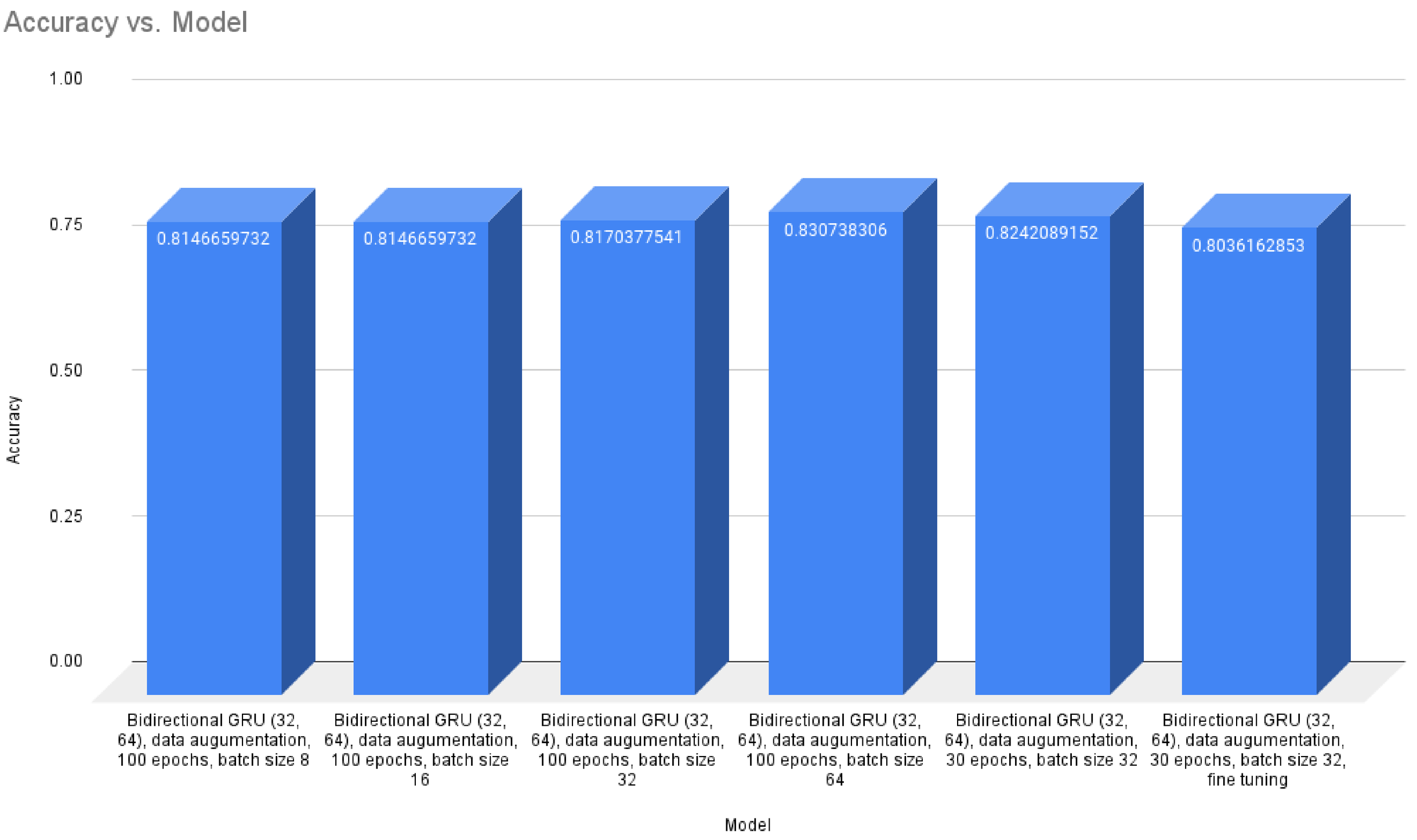


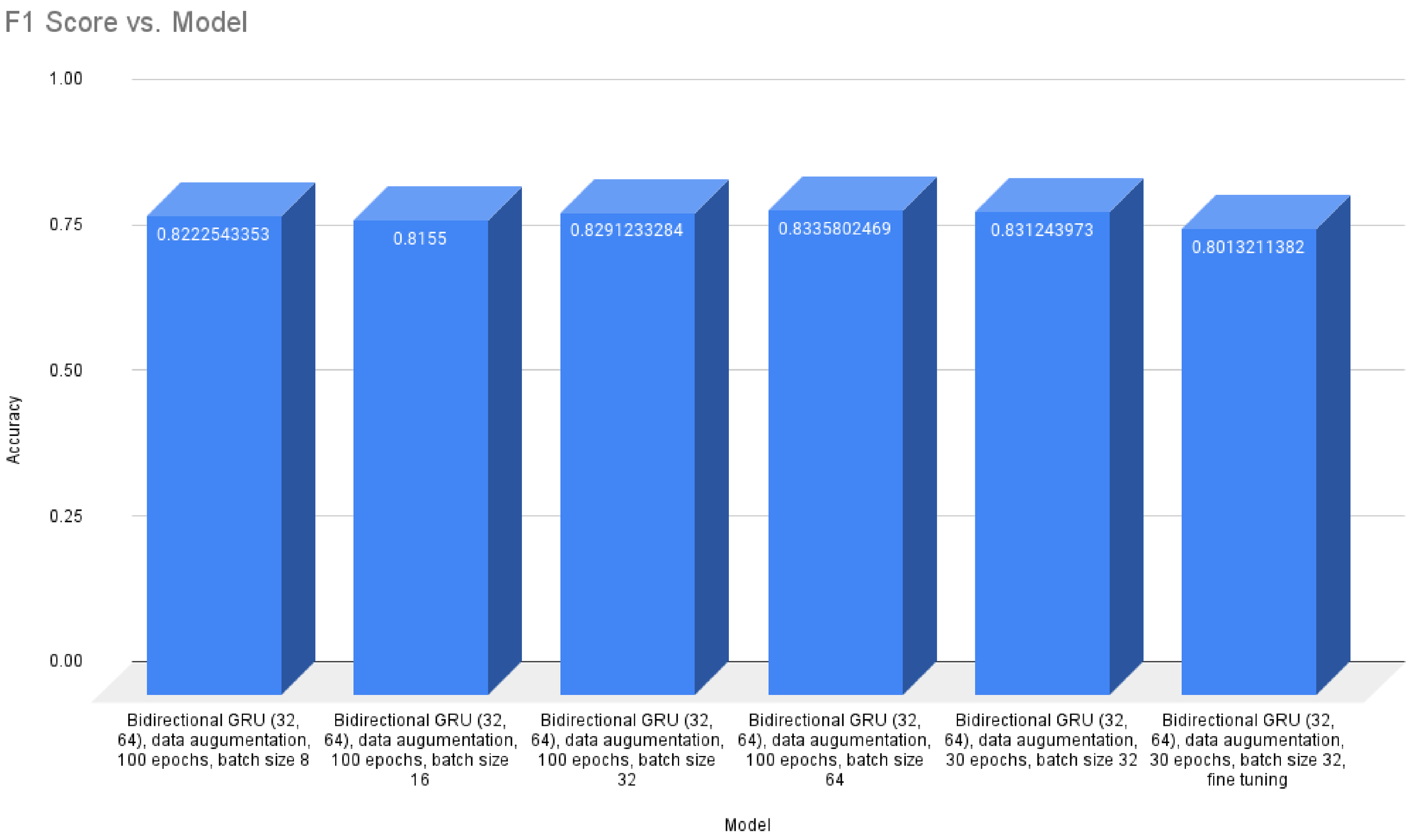
GRU model, 100 epoch, early stopping, batch size 32

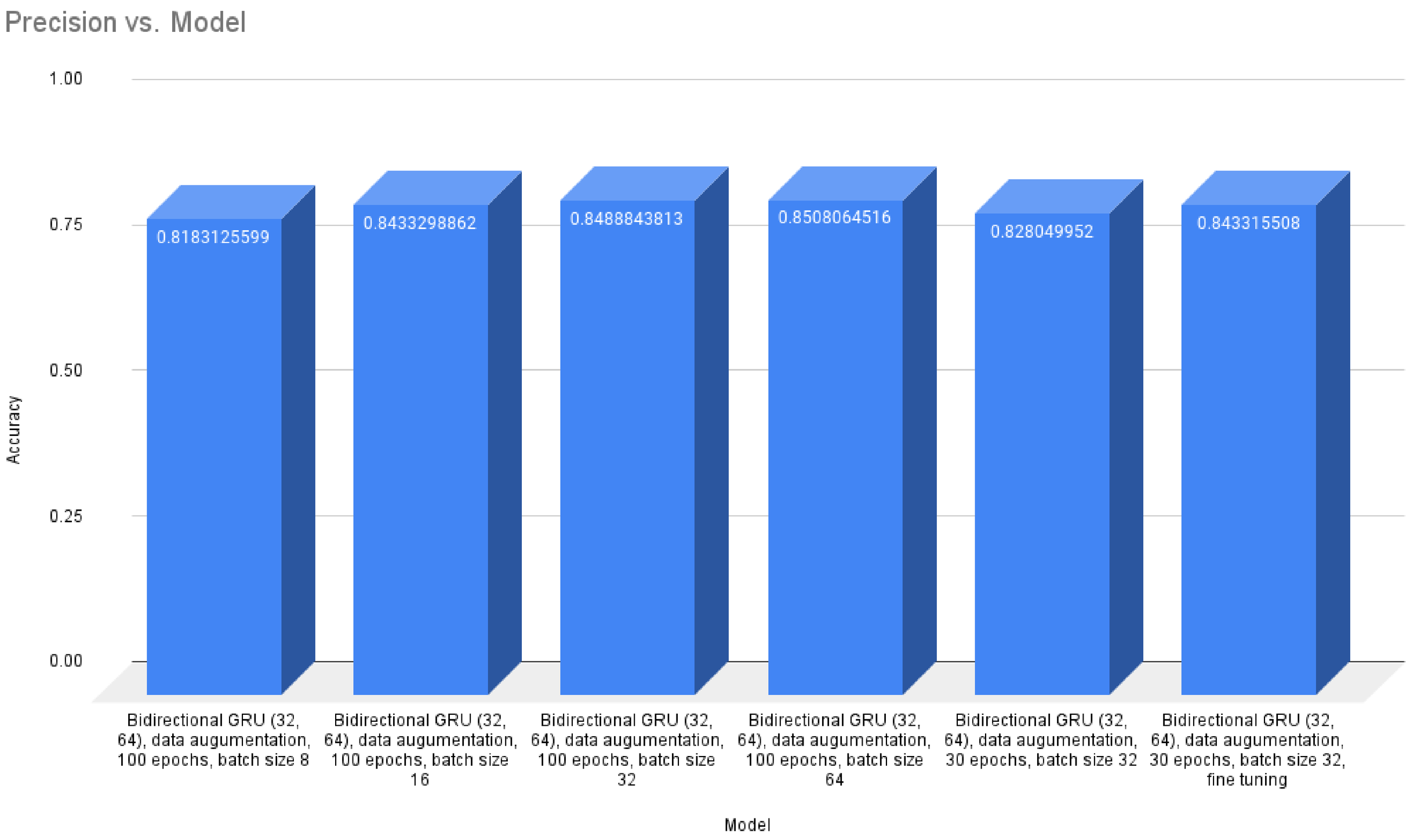


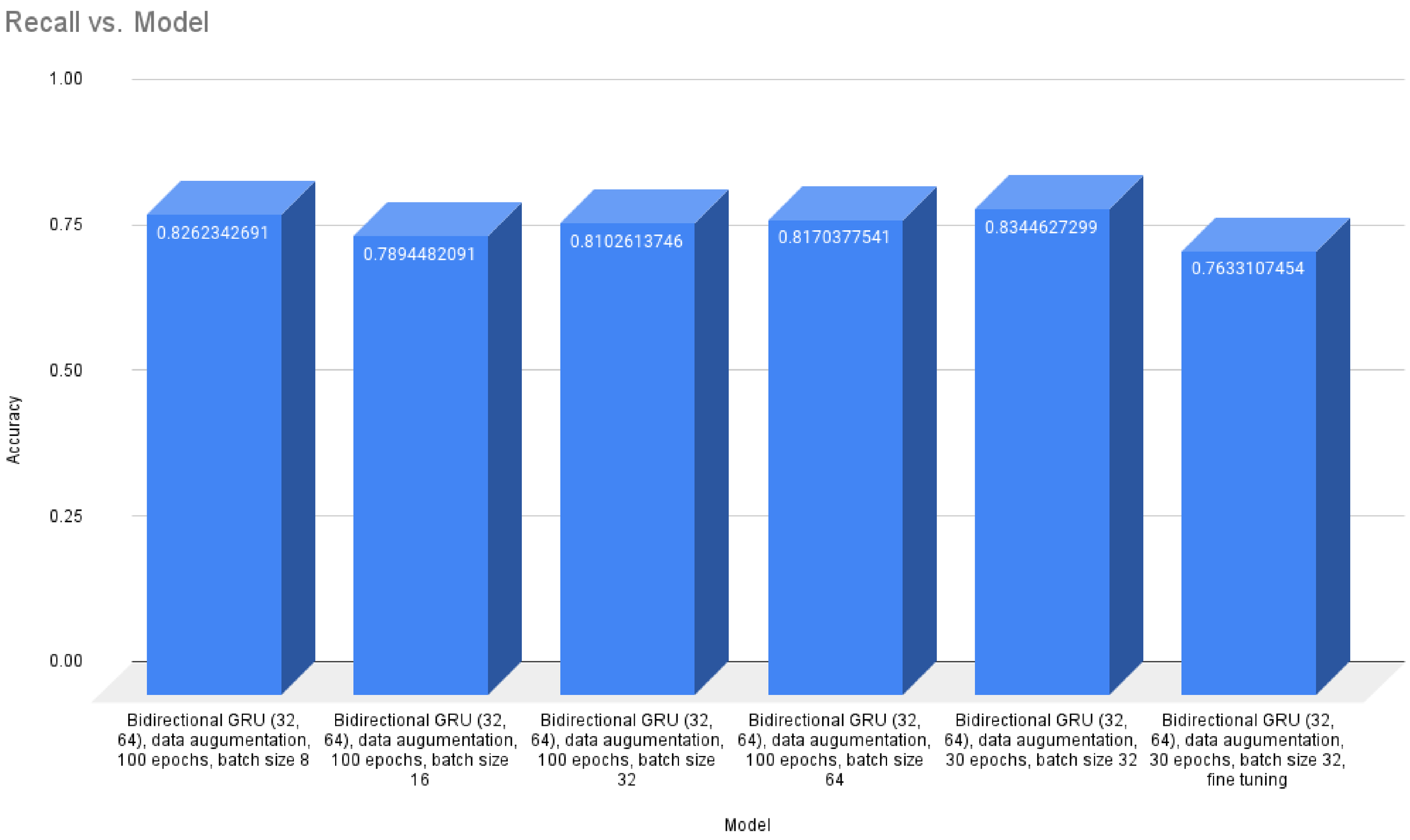
GRU model, 100 epoch, early stopping, batch size 64

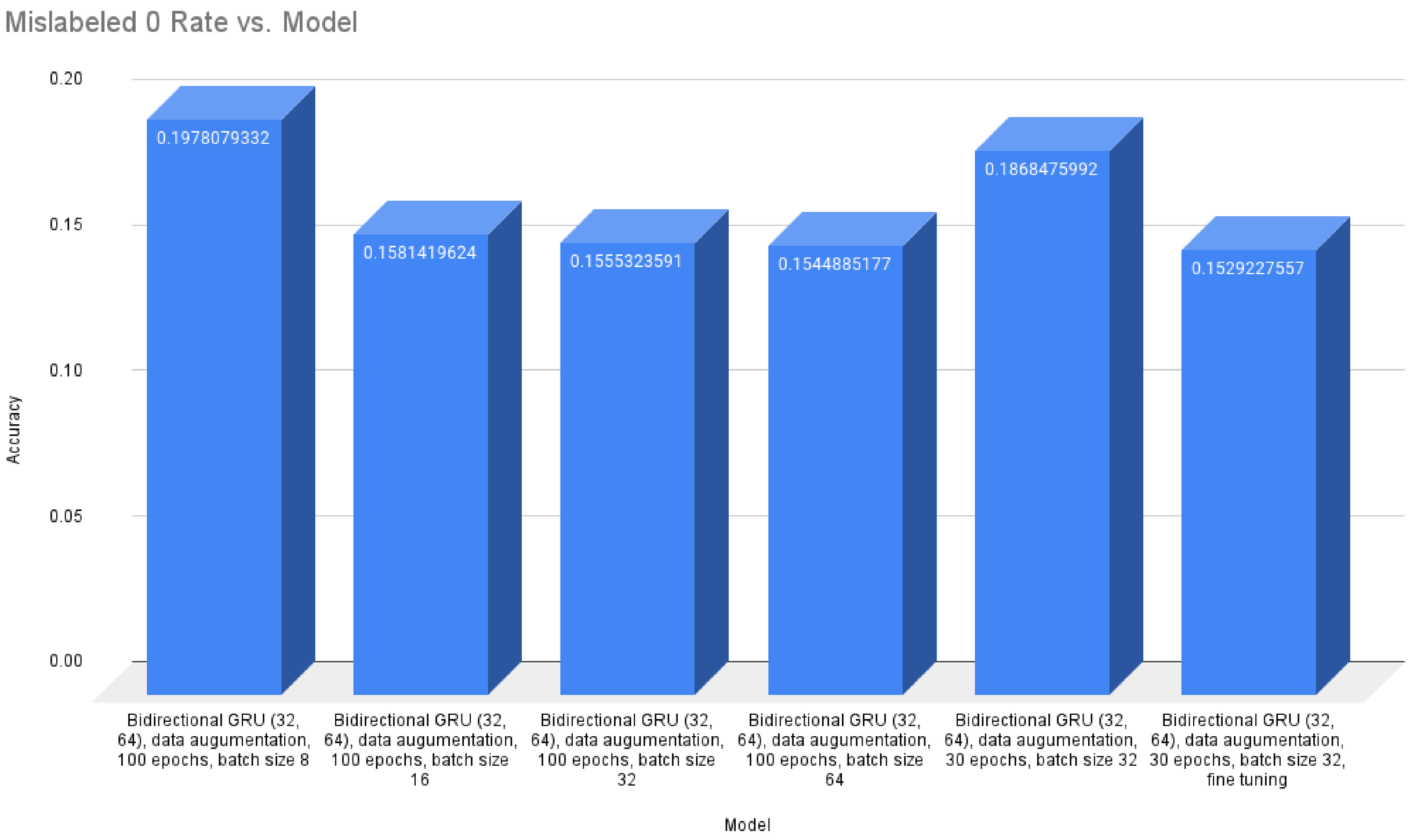


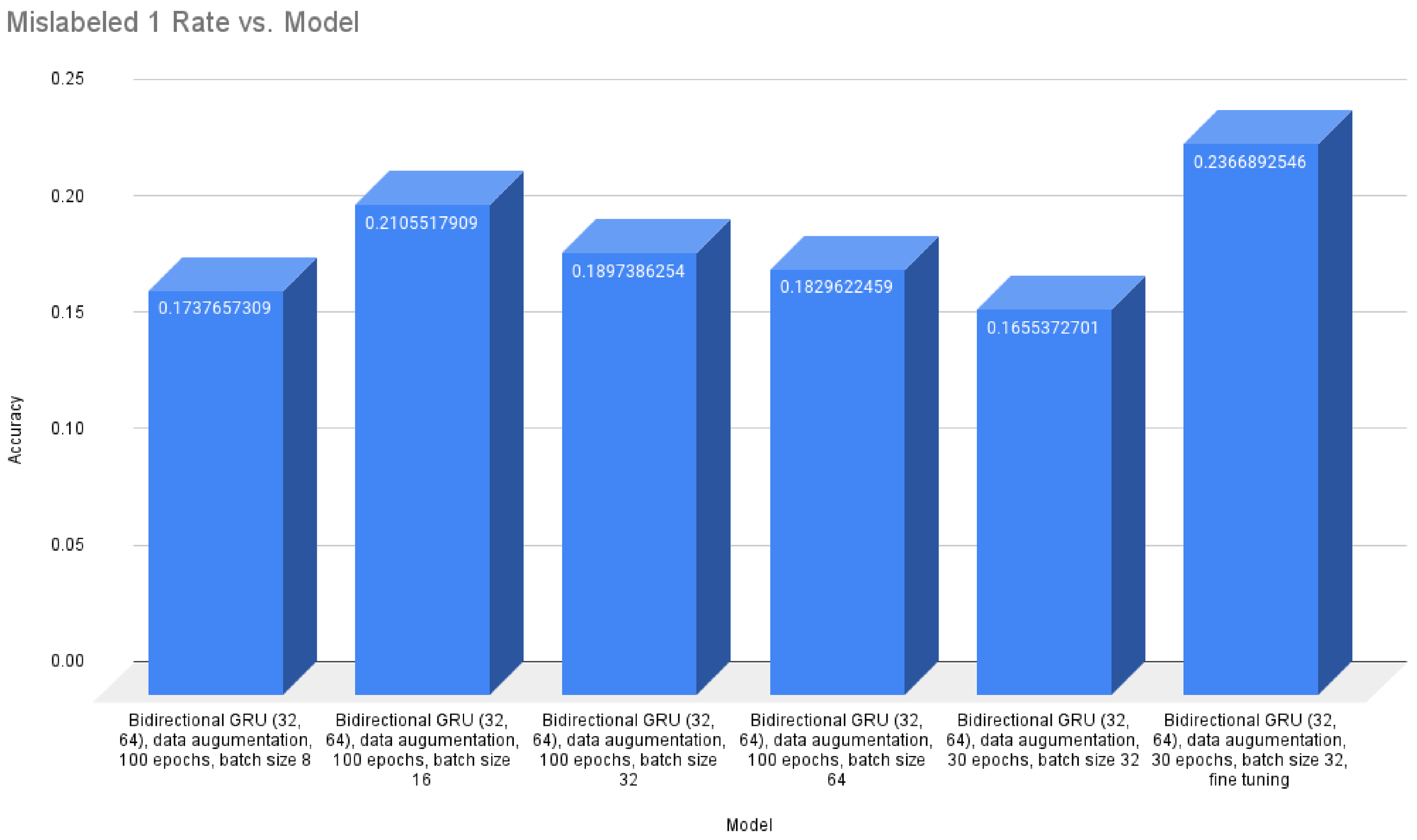

Accuracy vs. Model
1.00
0.75
0.50
0.25
0.00
Accuracy
0.8146659732
0.8146659732
0.8170377541
0.830738306
0.8242089152
0.8036162853
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 8
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 16
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 64
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32, fine tuning
Model


F1 Score vs. Model
1.00
0.75
0.50
0.25
0.00
Accuracy
0.8222543353
0.8155
0.8291233284
0.8335802469
0.831243973
0.8013211382
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 8
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 16
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 64
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32, fine tuning
Model


Precision vs. Model
1.00
0.75
0.50
0.25
0.00
Accuracy
0.8183125599
0.8433298862
0.8488843813
0.8508064516
0.828049952
0.843315508
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 8
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 16
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 64
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32, fine tuning
Model


Recall vs. Model
1.00
0.75
0.50
0.25
0.00
Accuracy
0.8262342691
0.7894482091
0.8102613746
0.8170377541
0.8344627299
0.7633107454
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 8
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 16
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 64
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32, fine tuning
Model


Mislabeled 0 Rate vs. Model
0.20
0.15
0.10
0.05
0.00
Accuracy
0.1978079332
0.1581419624
0.1555323591
0.1544885177
0.1868475992
0.1529227557
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 8
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 16
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 64
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32, fine tuning
Model


Mislabeled 1 Rate vs. Model
0.25
0.20
0.15
0.10
0.05
0.00
Accuracy
0.1737657309
0.2105517909
0.1897386254
0.1829622459
0.1655372701
0.2366892546
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 8
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 16
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 100 epochs, batch size 64
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32
Bidirectional GRU (32, 64), data augumentation, 30 epochs, batch size 32, fine tuning
Model

User Feedback Loop

User Feedback Loop Demonstration

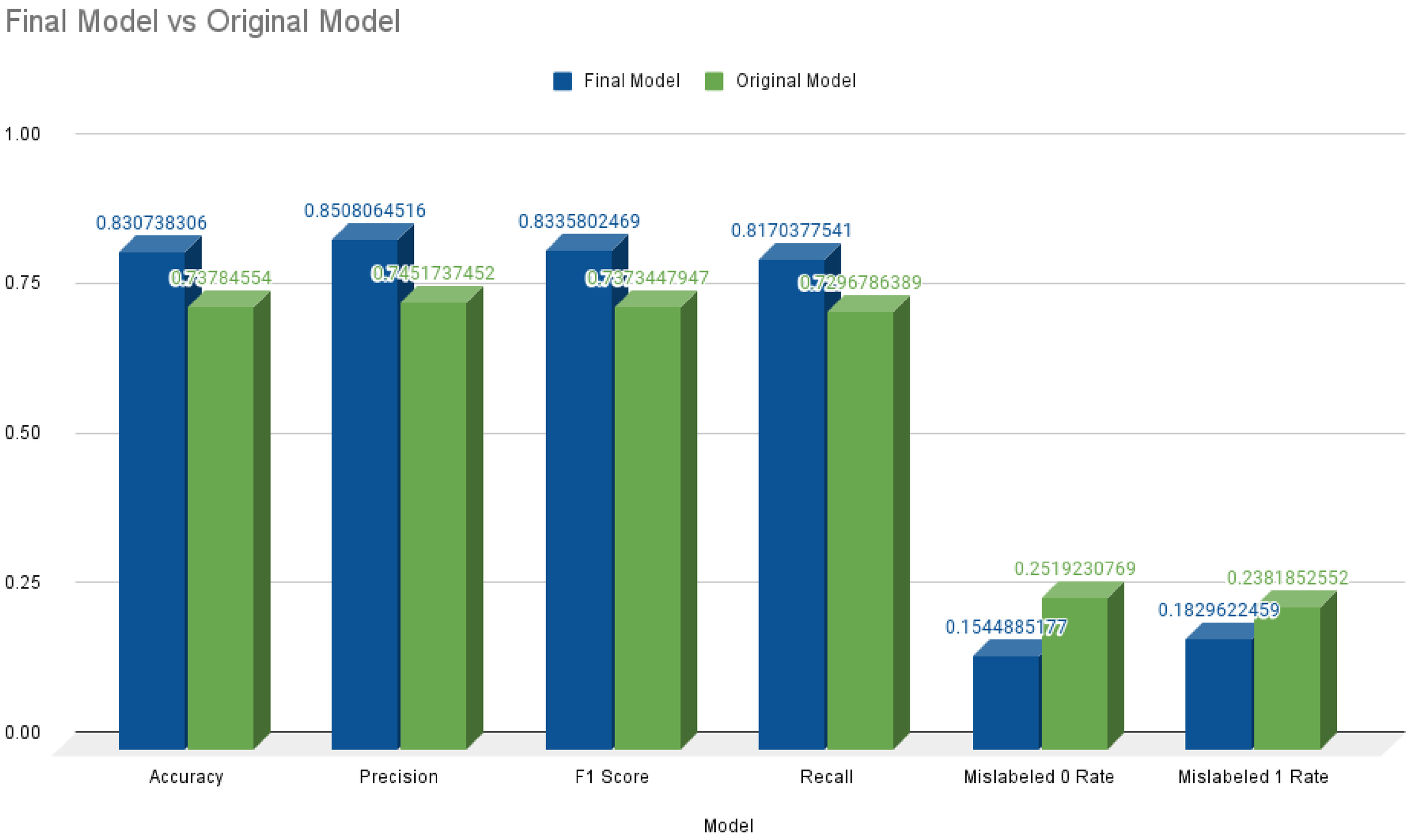
```
Enter a statement (or type 'done' to finish input): women do not have a y chromosome
Enter the correct label (0 or 1) for the statement: 0
Enter a statement (or type 'done' to finish input): women are not suited for leadership roles
Enter the correct label (0 or 1) for the statement: 1
Enter a statement (or type 'done' to finish input): I hate life
Enter the correct label (0 or 1) for the statement: 0
Enter a statement (or type 'done' to finish input): done
2023-11-29 20:21:33.239511: I tensorflow/compiler/xla/stream_executor/cuda/cuda_dnn.cc:442] Loaded cuDNN version 8906
1/1 [==============================] - 1s 1s/step
1/1 [==============================] - 0s 18ms/step
1/1 [==============================] - 0s 22ms/step
Model predictions:
Input: women do not have a y chromosome, Prediction: 1
Input: women are not suited for leadership roles, Prediction: 1
Input: I hate life, Prediction: 0
50
Epoch 1/100
2023-11-29 20:21:39.859593: I tensorflow/compiler/xla/service/service.cc:168] XLA service 0x7b2b6419e960 initialized for platform CUDA (this do
2023-11-29 20:21:39.859640: I tensorflow/compiler/xla/service/service.cc:176]   StreamExecutor device (0): Tesla T4, Compute Capability 7.5
2023-11-29 20:21:39.867248: I tensorflow/compiler/mlir/tensorflow/utils/dump_mlir_util.cc:269] disabling MLIR crash reproducer, set env var `ML
2023-11-29 20:21:40.027302: I ./tensorflow/compiler/jit/device_compiler.h:186] Compiled cluster using XLA!  This line is logged at most once fo
3982/3982 [==============================] - 49s 11ms/step - loss: 0.6606 - accuracy: 0.6812 - val_loss: 0.6401 - val_accuracy: 0.7247
Epoch 2/100
3982/3982 [==============================] - 40s 10ms/step - loss: 0.6345 - accuracy: 0.7354 - val_loss: 0.6438 - val_accuracy: 0.7187
```

```
Figure(1600x800)
125/125 [==============================] - 1s 4ms/step
Precision:  0.8698224852071006
Recall:  0.7826718296224588
F1 Score:  0.8239490445859872
Accuracy:  0.8264691109994977
Figure(800x600)
Model retrained!
```

User Feedback Loop Demonstration

```
Enter a statement (or type 'done' to finish input): women do not have a y chromosome
Enter the correct label (0 or 1) for the statement: 0
Enter a statement (or type 'done' to finish input): done
2023-11-29 21:29:03.609947: I tensorflow/compiler/xla/stream_executor/cuda/cuda_dnn.cc:442] Loaded cuDNN version 8906
1/1 [==============================] - 2s 2s/step
Model predictions:
Input: women do not have a y chromosome, Prediction: 0
No corrections needed!
```

Conclusion


Final Model vs Original Model
Final Model
Original Model
1.00
0.75
0.50
0.25
0.00
0.830738306
0.73784554
0.8508064516
0.7451737452
0.8335802469
0.7373447947
0.8170377541
0.7296786389
0.1544885177
0.2519230769
0.1829622459
0.2381852552
Accuracy
Precision
F1 Score
Recall
Mislabeled 0 Rate
Mislabeled 1 Rate
Model

Lessons Learned

- It is inappropriate to use only accuracy as an evaluation metric.
- The importance of data augmentation in improving the model's accuracy.
- The importance of fine-tuning the hyperparameters to achieve the best results.
- Checkpoints and callbacks are very important to save the model and to stop the training when the model stops improving.

Member Contribution

- Merna:
  - Worked on the data augmentation part
- Zein:
  - Worked on user feedback loop

Thank You