Shannon Barretto

Evidence Detection Task

Yash Mahajan

The Task: Given a claim and a piece of evidence, determine if the evidence is relevant to the claim.

Training Data: 21K claim-evidence pairs
Validation Data: 6K pairs
Format: (Claim, Evidence, Label)

Deep Learning with Transformers (Approach C)

Model Description

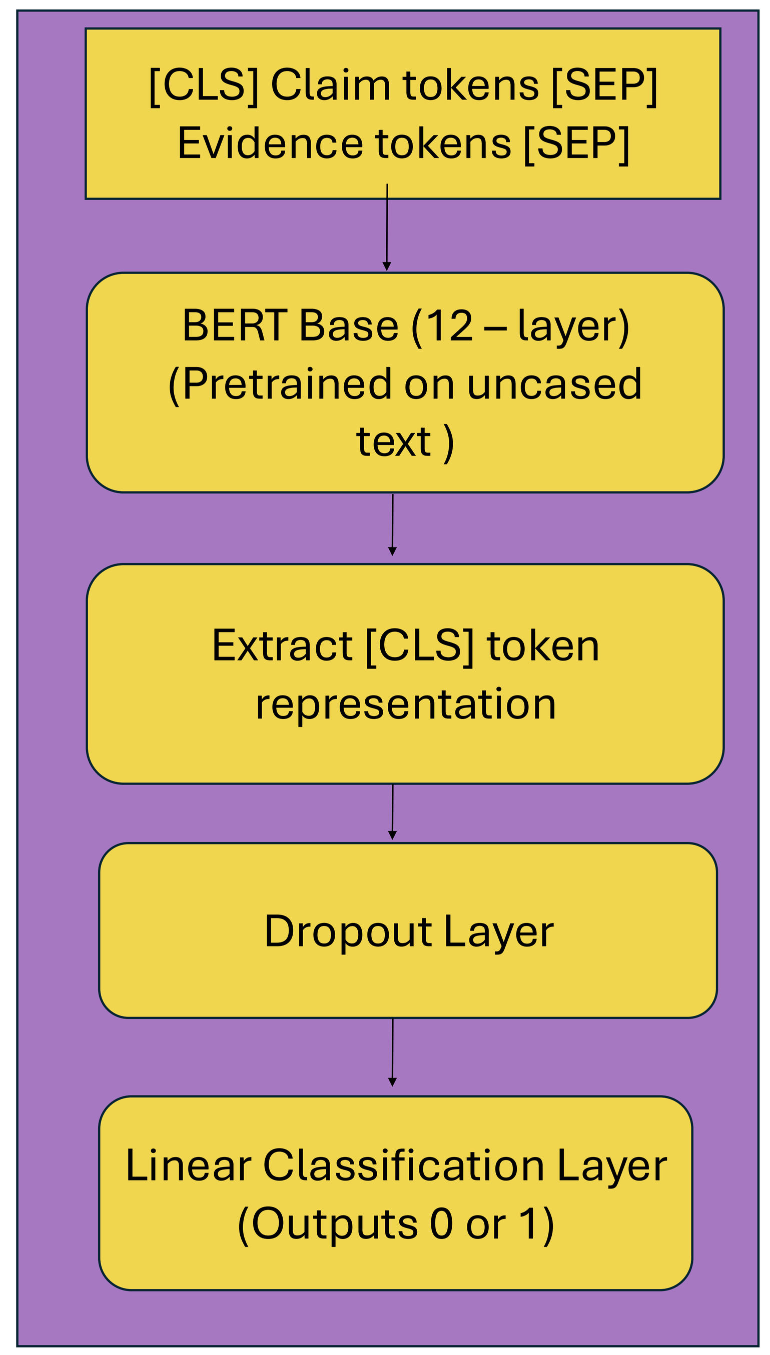
This is a model based on the **Bidirectional Encoder Representations from Transformers (BERT)**

It processes the input sequence formatted as:
[CLS] Claim tokens [SEP] Evidence Tokens [SEP]

[CLS] Token: Added at the beginning, its final hidden state serves as an aggregate representation of the entire input.

[SEP] Token: Separates the claim and evidence, helping the model understand their boundary.

To improve the performance, we added **dropout regularisation** and a **fully connected layer** on top of the existing BERT architecture. This helped in reducing overfitting to the training data.

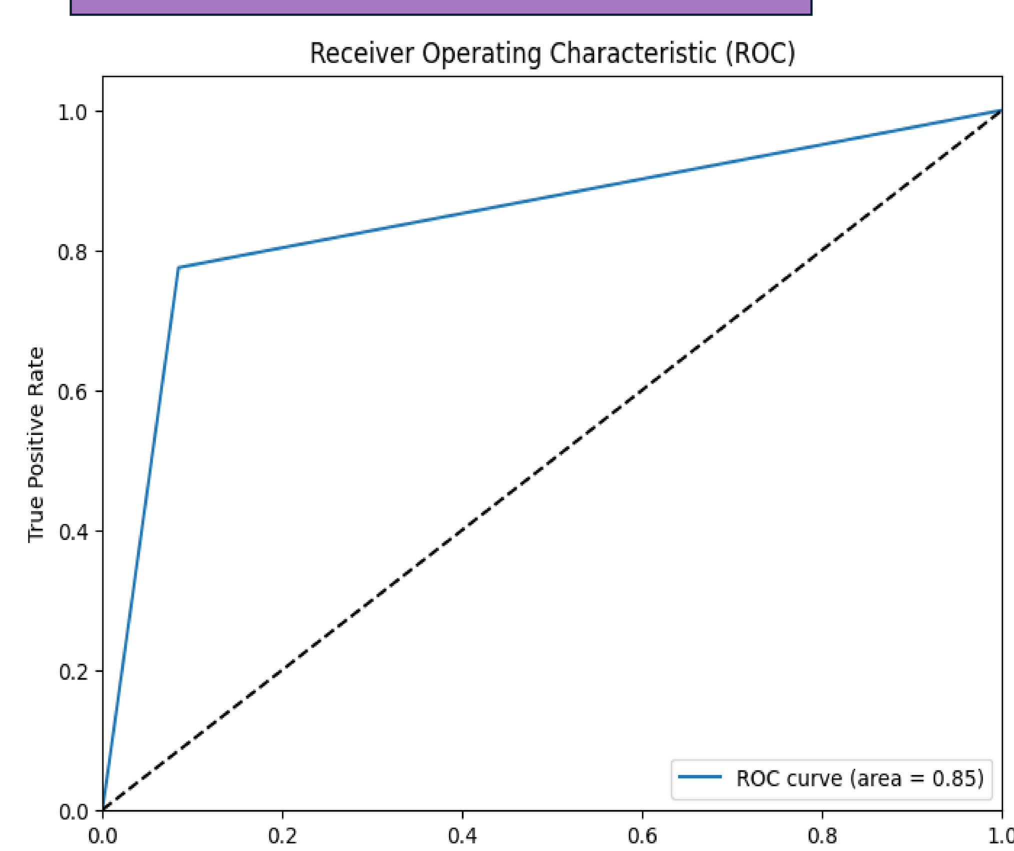


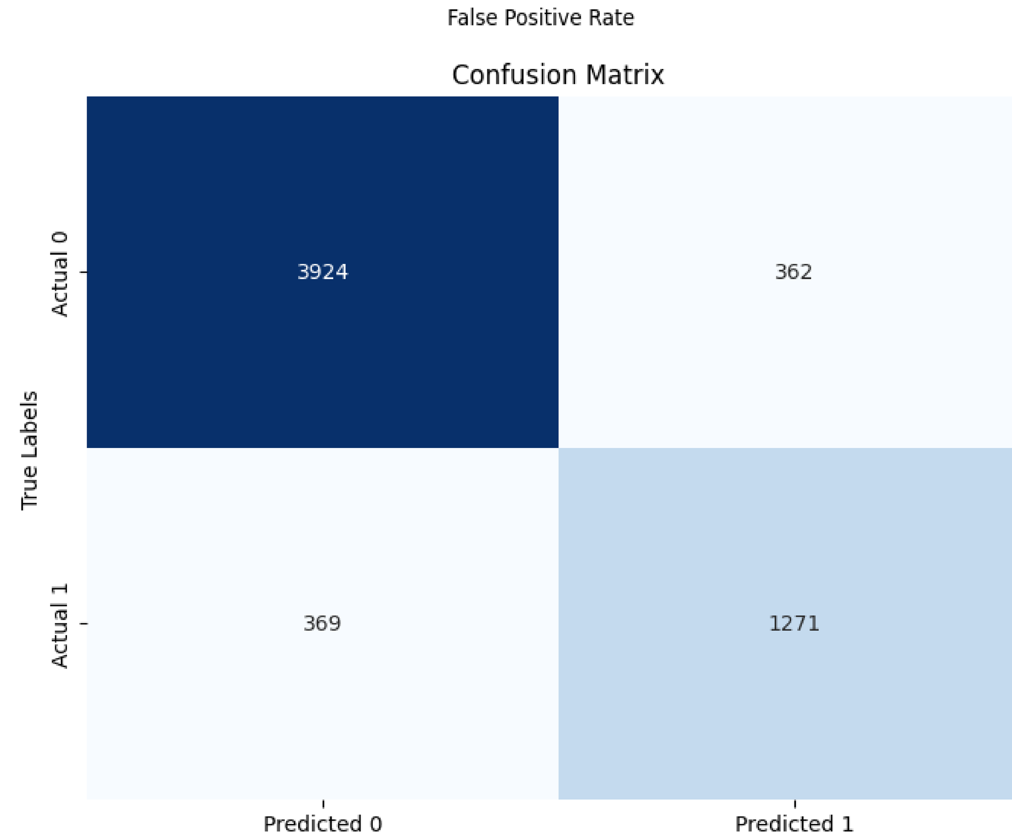
Results

The model achieved an accuracy of 87.65% which is similar to the baseline BERT model (87%)

We also plot the receiver operating characteristic (ROC). The final model achieved an AUC score of 0.85, suggesting that it can distinguish between relevant and irrelevant evidence

The model correctly predicts a large proportion of **true negatives** (3924) and **true positives** (1271). There are **362 false positives** and **369 false negatives**, which are relatively balanced. This suggests the model is not heavily biased toward predicting one class over the other.





Conclusion and Future Work

Our BERT-based evidence classifier achieved strong performance with an accuracy of 87.6%, demonstrating its effectiveness in distinguishing between relevant and irrelevant evidence.

There is a class imbalance in the training data with most examples (**72%**) labelled 0 (not relevant). Expanding the dataset with more **diverse claim-evidence pairs** may further improve performance and help address any residual issues related to class imbalance.

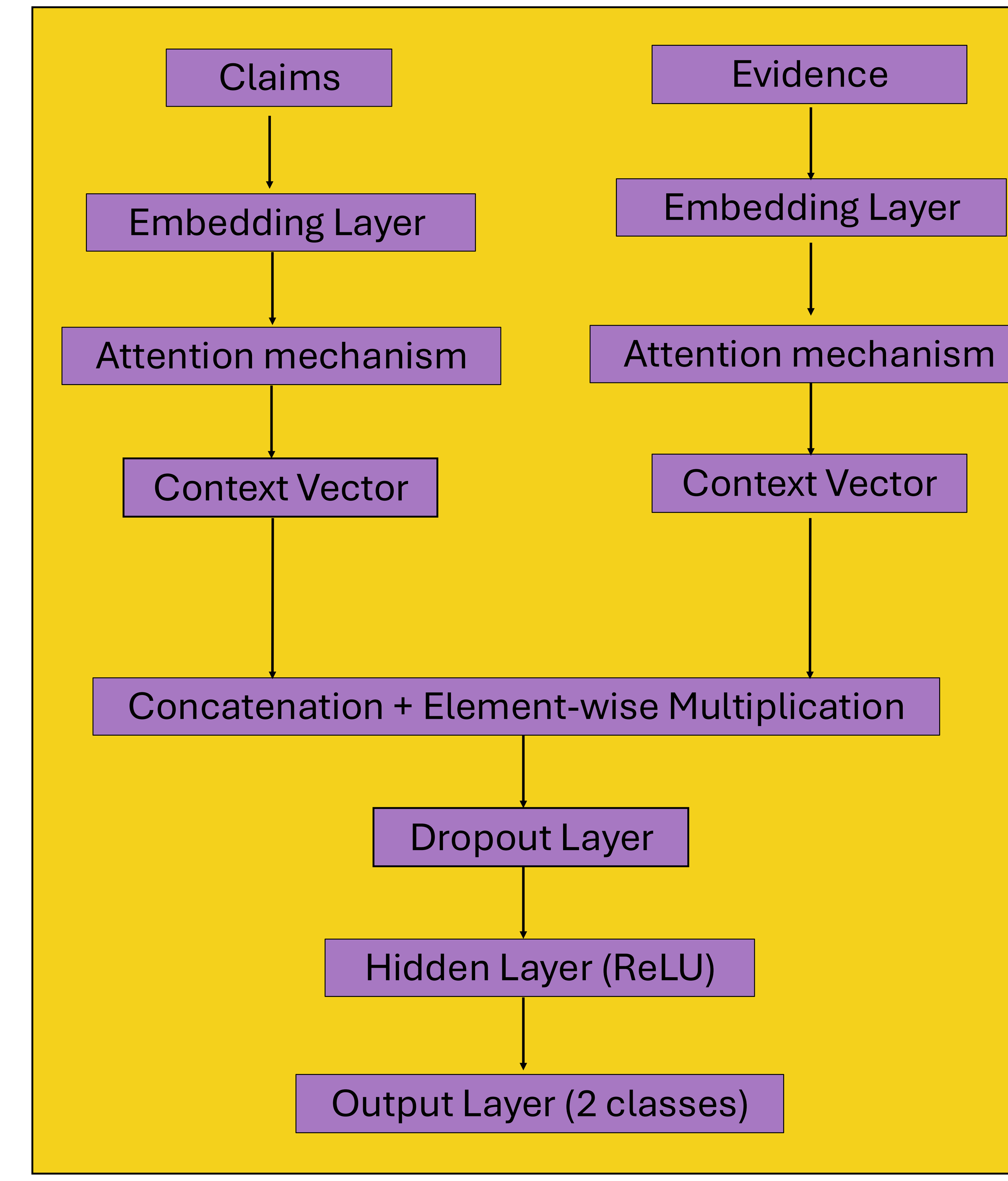
Deep Learning W/O Transformers (Approach B)

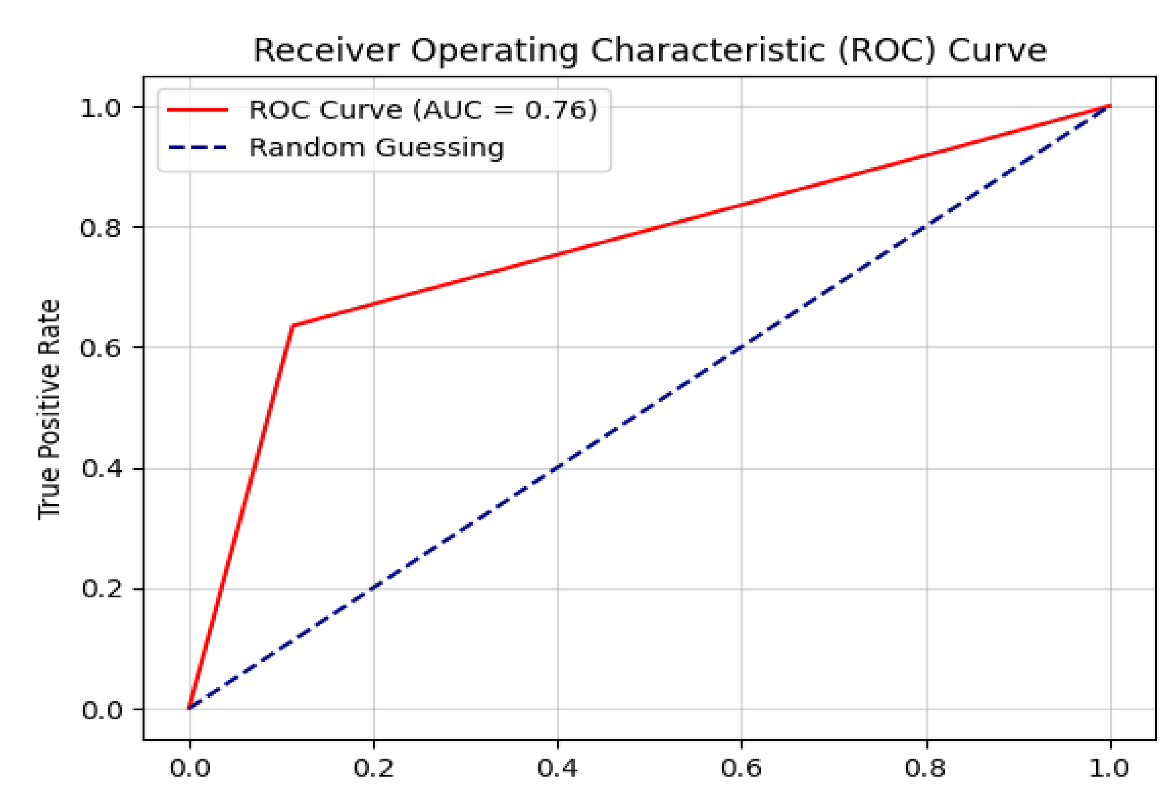
Neural Network Model

This model uses a **Bidirectional Long short-term memory (LSTM) with Attention** mechanism.
We initialized our model with pre-trained **GloVe 300d** embeddings. While training, we used just **1 layer** of LSTM, as it was over fitting and our training loss was decreasing after the first epoch itself when more layers were added to the model.

We used special token in our model; **<PAD> Padding** and **<UNK> Unknown** to handle text processing challenges:

- **<PAD> token (index 0):** used to make variable-length texts uniform by padding shorter sequences to match the batch's longest input.
- **<UNK> token (index 1):** Represents out-of-vocabulary words not seen during training or below our frequency threshold (3).

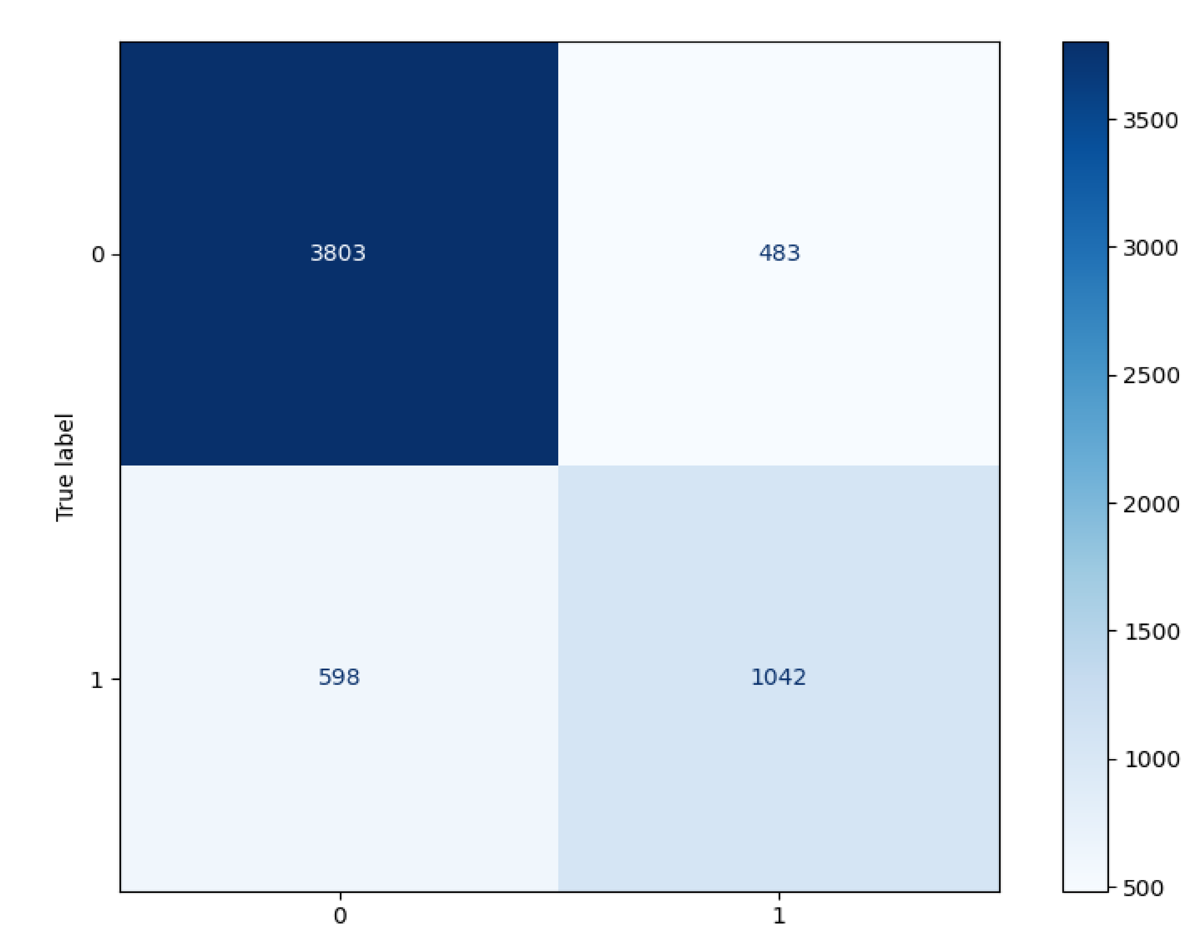




Results

The model achieved an accuracy of 81.76% on the dev set provided to us, which is 1% more compared to the baseline LSTM model (80.58%).

It even achieved an AUC of 0.76 on the dev set, indicating good discriminative ability between relevant and non-relevant claim-evidence pairs.



Conclusion and future work

Even though we experimented with different parameters, the accuracy gets capped at ~81-82%, likely due to limited training data. Despite the class imbalance(**72% non-relevant** vs **28% relevant**), the model correctly identifies 3803 true negatives and 1042 true positives.
Model performance could be improved by using different embeddings, which would be interesting to explore in future work.