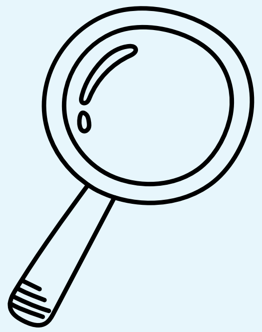
Saksham Goel
Kushagra Srivastava

Evidence Detection: Pairwise Sequence Classification

The University of Manchester

1. Abstract

This project employs a BI-LSTM and a RoBERTa transformer to tackle evidence detection, using a designated dataset. The poster compares the efficiency and depth of analysis offered by both models, showing their performance and potential in natural language understanding tasks.

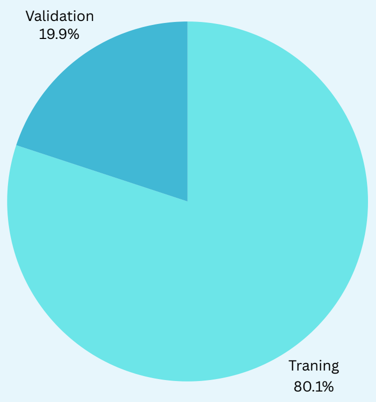


2. Introduction

In natural language understanding, accurately assessing evidence relevance is important. This project presents two models: a BI-LSTM and a RoBERTa transformer, tailored for textual data processing. Developed in a closed dataset framework, their efficacy in evidence detection is evaluated, providing insights for tasks like misinformation detection and automated fact-checking.

3. Dataset

The labelled dataset consists of 23,703 pairs of claim-evidence pairs for training and 5,902 claim-evidence pairs for validation



4. BiLSTM

Method

- **Data Preparation**: Combined claim and evidence texts, tokenized them, and padded sequences for uniform length.
- **Model Architecture**: Utilized a Bidirectional LSTM with dense layers for contextual understanding and classification.
- **Hyperparameter Tuning**: Explored various combinations of embedding dimension, LSTM units, and dropout rates for optimal performance.
- **Training and Validation**: Trained the model on the training data and evaluated its performance using validation data, optimizing for accuracy.

Results

The model was evaluated on a combination of evaluation metrics:

- **Accuracy**: 80%
- **Precision**: 100%
- **Recall**: 67%
- **F1-Score**: 80%

These metrics indicate the model is very accurate, minimizes false positives and captures a significant portion of relevant evidence instances

Model Architecture

Embedding Layer

Bidirectional Layer

Dense Layer

Dropout Layer

Dense Layer

Batch Normalization

Dense Layer

Comparison

This model outperforms the baseline model on EvalAI on all the metrics.A major reason for this could be the model architecture that is able to properly understand the relationship between the claim and evidence. Another reason could be the hyperparameter selection

5. RoBERTa

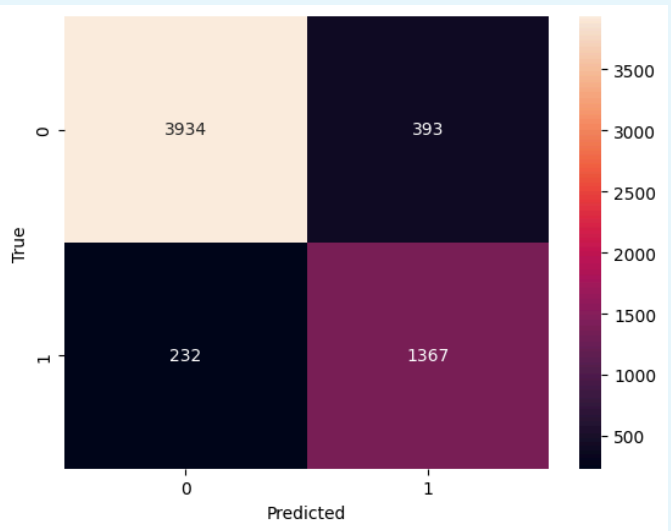
Method

- RoBERTa is a Robustly optimised BERT pretraining approach, developed by Facebook.
- **Fine-tuning**: Conducted on 23,703 claim-evidence pairs
- **Hyerparameter tuning:** Iterating through learning rates, selected the best via validation accuracy
- **Training strategy**: Split data 80:20 into training and validation sets, fine-tuned over 3 epochs

Reasoning

- Proven success of transformer models in NLP through many studies
- BERT based models are state-of-the-art
- Pretraining helps the model have an existing understanding of language, helping it generalise well to unseen data.

Confusion Matrix for RoBERTa



Results

- Validation Performance: **89.5%** accuracy with a best learning rate of 1e-5**.**
- Precision & Recall: Balanced precision (**89.9%**) and recall (**89.4%**).
- F1 Score: Strong at **89.6%**
- MCC: High at **0.74**, showing a strong correlation between predictions and actual labels

Conclusion

We looked at evidence detection through an BiLSTM and RoBERTa model. The evaluation metrics showed that these models performed well. The BiLSTM had a perfect precision, but moderate recall. This meant that positive predictions were always correct, but all positive instances weren't always identified.

On the other hand, the RoBERTa had a balanced precision and recall, showing a more reliable performance. RoBERTa performed better regarding all metrics except precision. This shows the superiority of transformer models on this task.