Natural Language Inference
RoBERTA & BILSTM Approach

The University of Manchester

Group 54
Alan Prophett
Zac Curtis

Abstarct

- We tackled the Natural Language Inference task using two models: a BiLSTM and a fine-tuned transformer.
- Both were trained to determine if a hypothesis logically follows a given premise.
- The BiLSTM achieved over 70% F1, while the transformer exceeded 90%, highlighting the effectiveness of deep pretrained models.

Introduction

WHAT

- NII is a task that determines whether a hypothesis logically follows from a given premise.
- It involves classifying the relationship as entailment, neutral, or contradiction (binary in our case).

WHY

- Comparing a simpler BiLSTM and a deep pretrained transformer helps evaluate trade-offs between efficiency and accuracy.
- This dual approach shows how different architectures understand logical relationships in text.

HOW

Deep Learning (Non-Transformer)

- Captures word order and context using sequential processing
- Will struggle with long dependencies

Deep Learning

- Uses self-attention to understand full sentence context in parallel
- Handling long-range dependencies better.

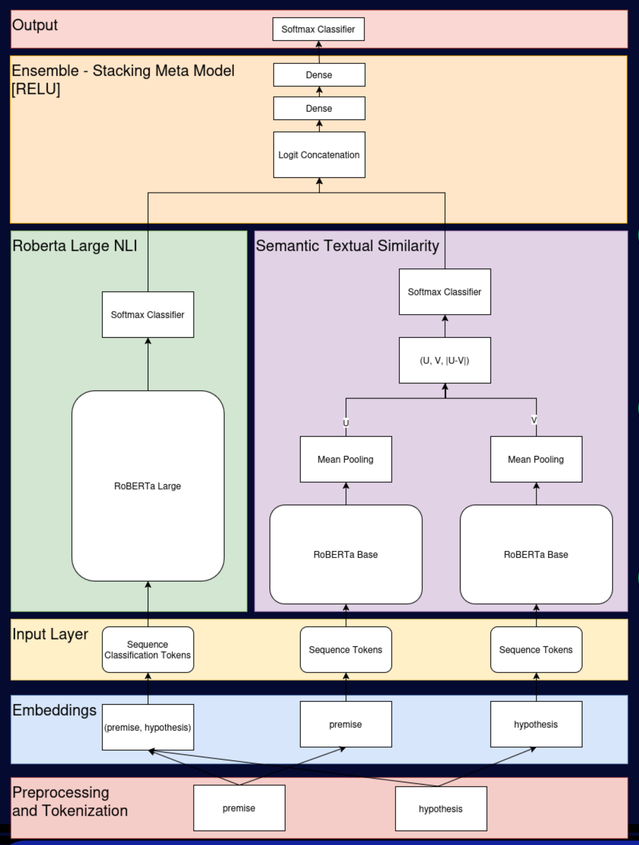
References

[1] Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., & Inkpen, D. (2017) 'Enhanced LSTM for Natural Language Inference', arXiv preprint arXiv:1609.06038v3. Available at https://arxiv.org/abs/1609.06038v3 : (Accessed: 10 April 2025).

[2] Reimers, N. and Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

DEEP Learning Transformer



INPUT DATA

- Minimal preprocessing: maintain contextual integrity, with stop words and punctuation deliberately retained.

SEMANTIC TEXTUAL SIMILARITY

- Compare the similarity of semantically meaningful sentence embeddings [2].

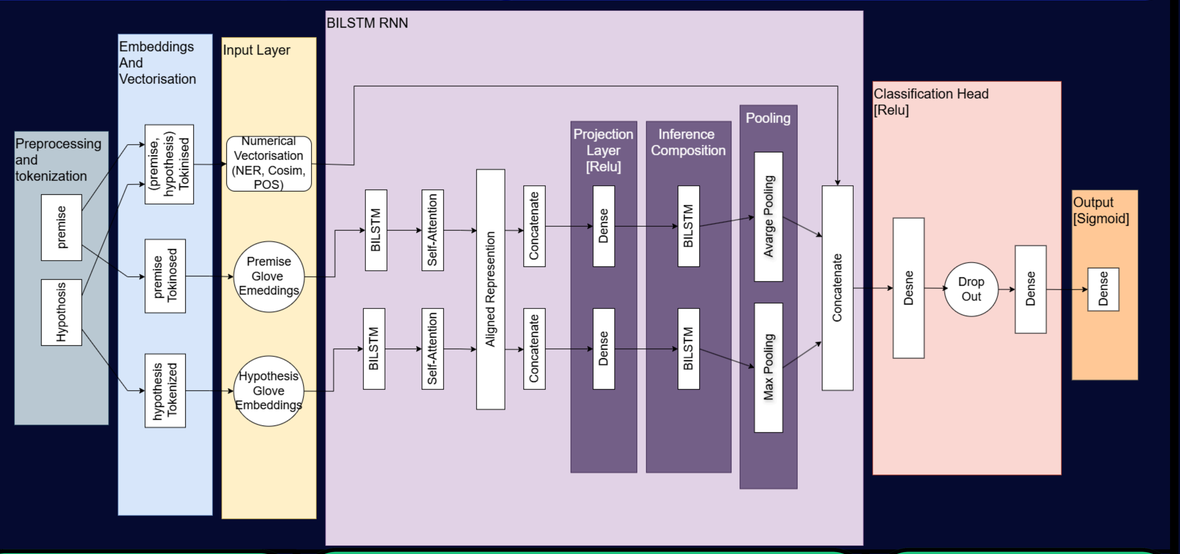
RoBERTa LARGE MODEL

- RoBERTa-large model fine-tuned in a pairwise classification setup [3].

ENSEMBLE META MODEL

- Leverages strengths of different architectures.
- Meta-classifier learns optimal weighting of individual model predictions.
- Improves robustness and generalization.

Deep Learning Non Tranformer



INPUT

- Text pairs (premise + hypothesis) cleaned and tokenized.
- Tokens are converted into Glove Emeddings
- (NER) and POS tagging are applied

BILSTM RNN ESIM

- Separate BiLSTMs encode each sentence.
- Intermediate layers refine embeddings through projection and composition.
- Inference is modeled by comparing encoded sentence representations.
- Max pooling aggregates important features across time steps.

CLASSIFIER

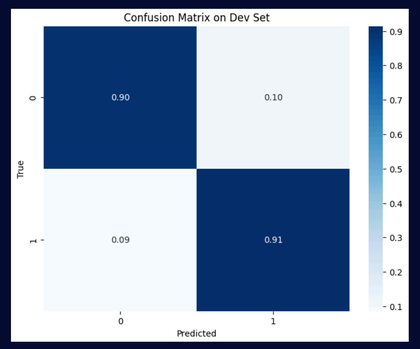
- Final vector passed through dense layers with dropout.
- Output layer uses sigmoid to predict class label (entailment or not).

Results

TRANSFORMER MODEL PERFORMANCE

Ensemble Model Component Test: Accuracy Comparison

Model	Train	Dev	Test
RoBERTa Large	99.41%	90.88%	91.09%
Similarity Model	97.07%	78.29%	78.03%
Ensemble Model	99.61%	90.87%	91.02%



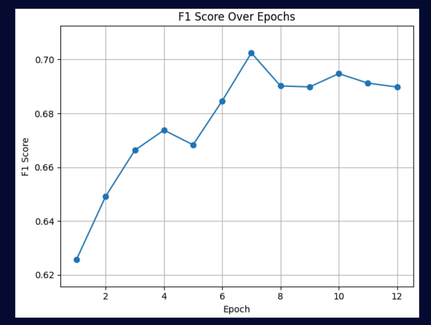
BILSTM RNN MODEL PERFORMANCE

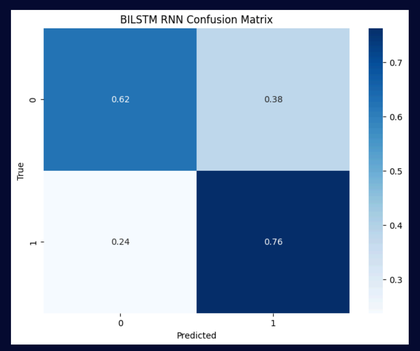
BILSTM RNN Performance on Dev Set

MCC	39.9%
Weighted Precision	69.9%
Weighted Recall	69.9%
Weighted F1-score	69.9%
Accuracy	69.9%

Ensemble Performance on Dev Set

MCC	81.8%
Weighted Precision	90.9%
Weighted Recall	90.9%
Weighted F1-score	90.9%
Accuracy	90.9%





Conclusions

- BiLSTM for NLI is lightweight and easy to train
- Transformer model offers higher robustness and better overall performance.

FUTER WORK

BOTH

- Data Set augmentation.

BILSTM

- More hyper Parameter tuning
- FInetuning encoder as well as model

Transformer

- Explore additional models to ensemble, to improve robustness.