The University of Manchester

Peter Hamer

Vinayak Bhadoriya

NLI Shared Task

Task Definition: Given a premise and a hypothesis, determine if the hypothesis is true based on the premise.

Training Dataset: 26k premise-hypothesis pairs

Development Dataset: 6k premise-hypothesis pairs

Traditional Machine Learning

Ensemble Model

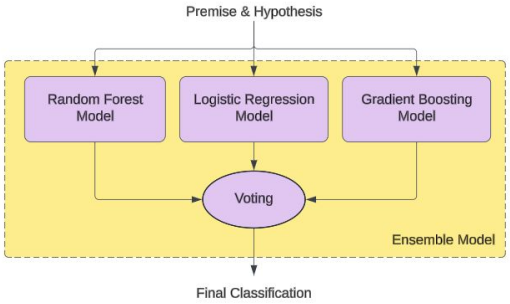
The final model uses an ensemble architecture with hard (majority) voting, combining three separately trained models:

- A Random Forest Model
- A Logistic Regression Model
- A Gradient Boosting Model

All three models use TF-IDF word embeddings.

Each model's hyperparameters were tuned using a cross-validated grid search, choosing the set with the highest accuracy.

Each model was trained using the same 26k pairs, before being combined into the ensemble mode.



Evaluation

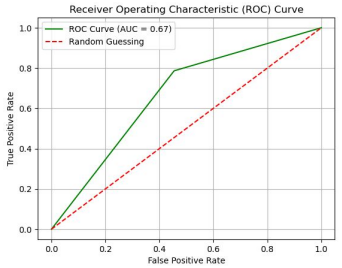
The final model was evaluated using the development set. The following metrics were observed:

- Precision, Recall & F1-score
- Cross-Validation Accuracy
- ROC Curve (AUC)

Results

Achieved a cross-validation accuracy of 64.78%, which is a significant improvement over the baseline SVM model (54.92%).

Final model tended to perform better for entailment pairs (1) than contradiction/neutral pairs (0), achieving F1-scores of 0.71 and 0.62 respectively.



TARGET / OUTPUT	Class0	Class1	SUM
Class0	1778 26.39%	1481 21.98%	3259 54.56% 45.44%
Class1	744 11.04%	2734 40.58%	3478 78.61% 21.39%
SUM	2522 70.50% 29.50%	4215 64.86% 35.14%	4512 / 6737 66.97% 33.03%

The performance of the model may have been improved if the type of word embeddings used was different across the different sub-models, as different characteristics would be captured. Future work could include an investigation into how different word embeddings effect the performance of the proposed model architecture.

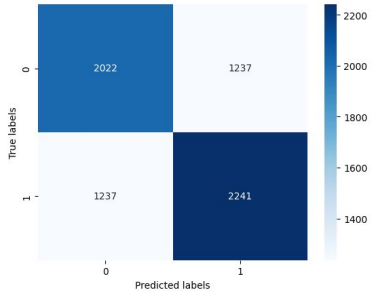
Additionally, it would be interesting to see how changing or adding sub-model architectures would effect the final models performance.

Deep Learning (w/o Transformers)

Neural Network Model

The model is based on the sentence embedding 'sum-of-words' approach proposed by Bowman et al.(2015) in the SNLI paper.
Our model has a few modifications: instead of the sum-of-words approach, we use the average of the word embeddings and after experimenting with different architectures, the following is what we propose:

Use GloVe embeddings to encode the premise and hypothesis, get the average of the embeddings instead of just summing them then and pass through our neural network, until our final 'softmax' layer - which converts the vectors into a probability distribution of the 2 possible outcomes (entailment or contradiction/neutrality). We can then simply choose the outcome with the greater probability, which will be our prediction.



Results

The model achieved an accuracy of 63% on the dev set provided to us, a definite step-up compared to the baseline model which used a BiLSTM (56%).

The model performed similarly on both entailment and contradiction/neutral pairs, achieving F-1 scores of 0.64 and 0.62 respectively.

Further work

Even though we experimented with different layers and different number of parameters, the accuracy seemed to get capped at ~64-65%.

That may be because of the amount of training data but it could also be a reflection of a lack of local inference modelling as proposed by Chen et al. in 2017, and to further improve models, that approach may pay the way forward.