NATURAL LANGUAGE INFERENCE

Team NAGin

Amal Sunny - 2021121011
Nukit Tailor - 2020114012
George Paul - 2021121006

Overview

Topics

Problem Statement

Describes the problem of Textual Entailment

Datasets

Description of the Datasets utilized

Models

Approaches taken and the models generated

Results

Results and observations from all the models

Textual Entailment

Problem

There are two sentences given, a premise and a hypothesis . We have to detect if the hypothesis can be entailed from the hypothesis. An example can be given as:

Premise: A soccer game with multiple males playing.
Hypothesis: Some men are playing a sport.
Verdict: Entailment!

Essentially, a 3 class, classification problem.

History

This has been a rather prominent and famous problem, with RTE(recognizing textual entailment) since 2008. Many new and novel techniques have been devised to tackle the problem - which can be seen on the leaderboard of the SNLI dataset page.

Datasets Used

SNLI - The Stanford Natural Language Inference (SNLI) Corpus

The Stanford Natural Language Inference (SNLI) corpus is a collection of **570k** human-written English sentence pairs manually labeled for balanced classification with the labels **entailment**, **contradiction**, and **neutral**. It was meant to serve as both a benchmark for evaluating representational systems for text as well as for developing NLP models of any kind.

MultiNLI

The Multi-Genre Natural Language Inference (MultiNLI) corpus is a crowd-sourced collection of **433k** sentence pairs annotated with textual entailment information. The corpus is modeled on the SNLI corpus, but differs in that covers a range of genres of spoken and written text, and supports a distinctive cross-genre generalization evaluation.

Models

SimpleRNN(Baseline)

A Seq-to-Seq Model reinforced with GloVe embeddings and fine tuned with various layers after encoding.
Sigmoid, ReLU and softmax are used as activation functions for different models under this

LSTM

Following up on the baseline model, an LSTM layer is used in the translation/encoding layer after initializing with the GloVe embeddings

GRU

Similar to the LSTM model, the translation layer is replaced with a GRU (Gated Recurrent Unit) layer.

BERT

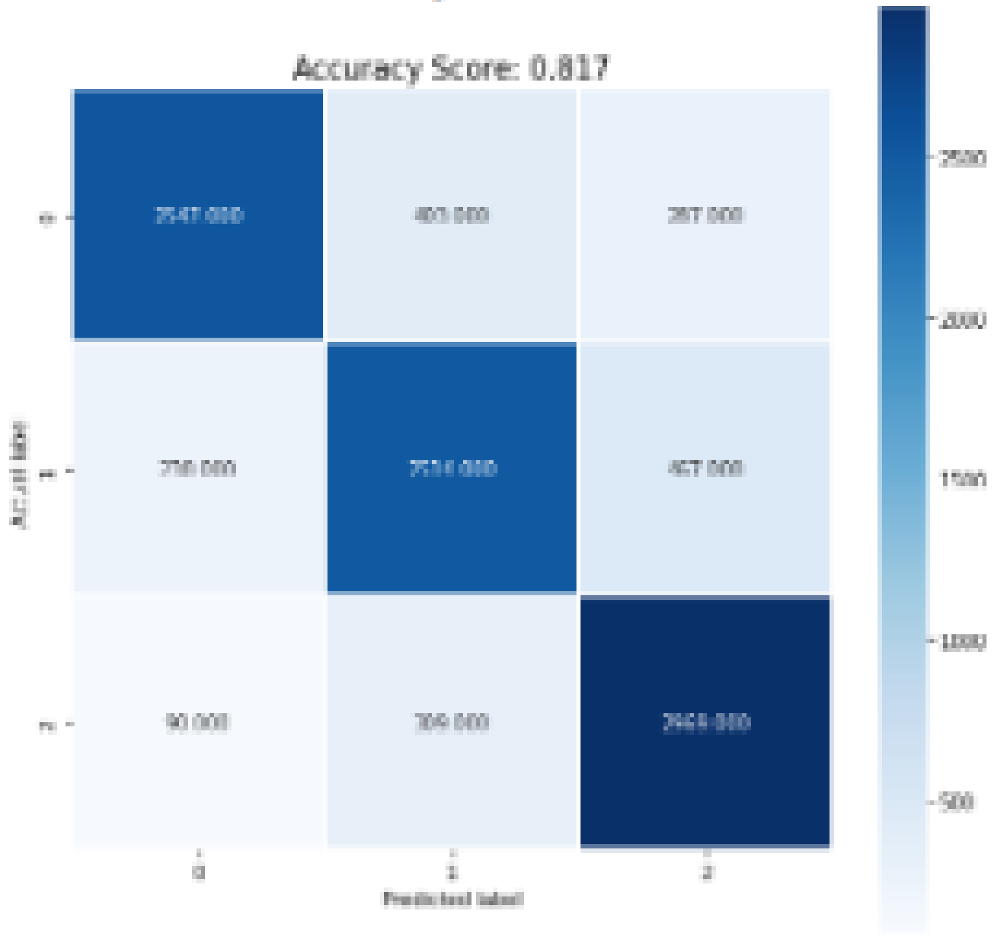
Applied a SoTA approach to the problem, by using a pretrained BERT model for tokenizing and encoding.

Results and Conclusion

Performance of the model

A. Simple RNN with ReLU on SNLI

Total Accuracy = 81.74 %



	Precision	Recall	F1 Score
Entailment	0.89	0.79	0.83
Contradiction	0.78	0.78	0.78
Neutral	0.80	0.88	0.84

Simple RNN

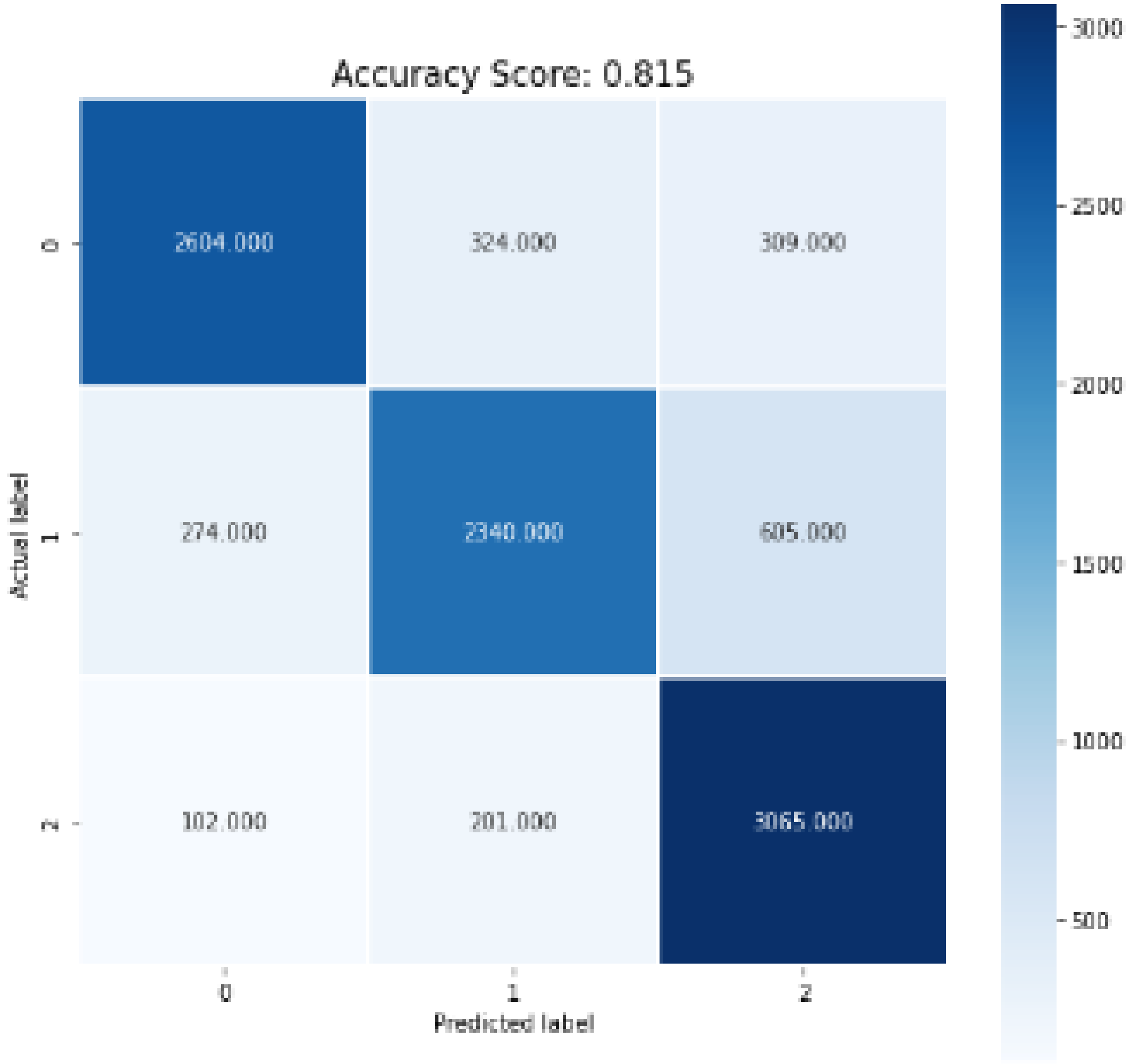
There were 6 total models, varying by the activation function used for each: sigmoid, ReLU, softmax.

For SNLI, best SimpleRNN model: **ReLU**
For MultiNLI: **ReLU**

LSTM model

G. LSTM on SNLI

Total Accuracy = 81.52 %

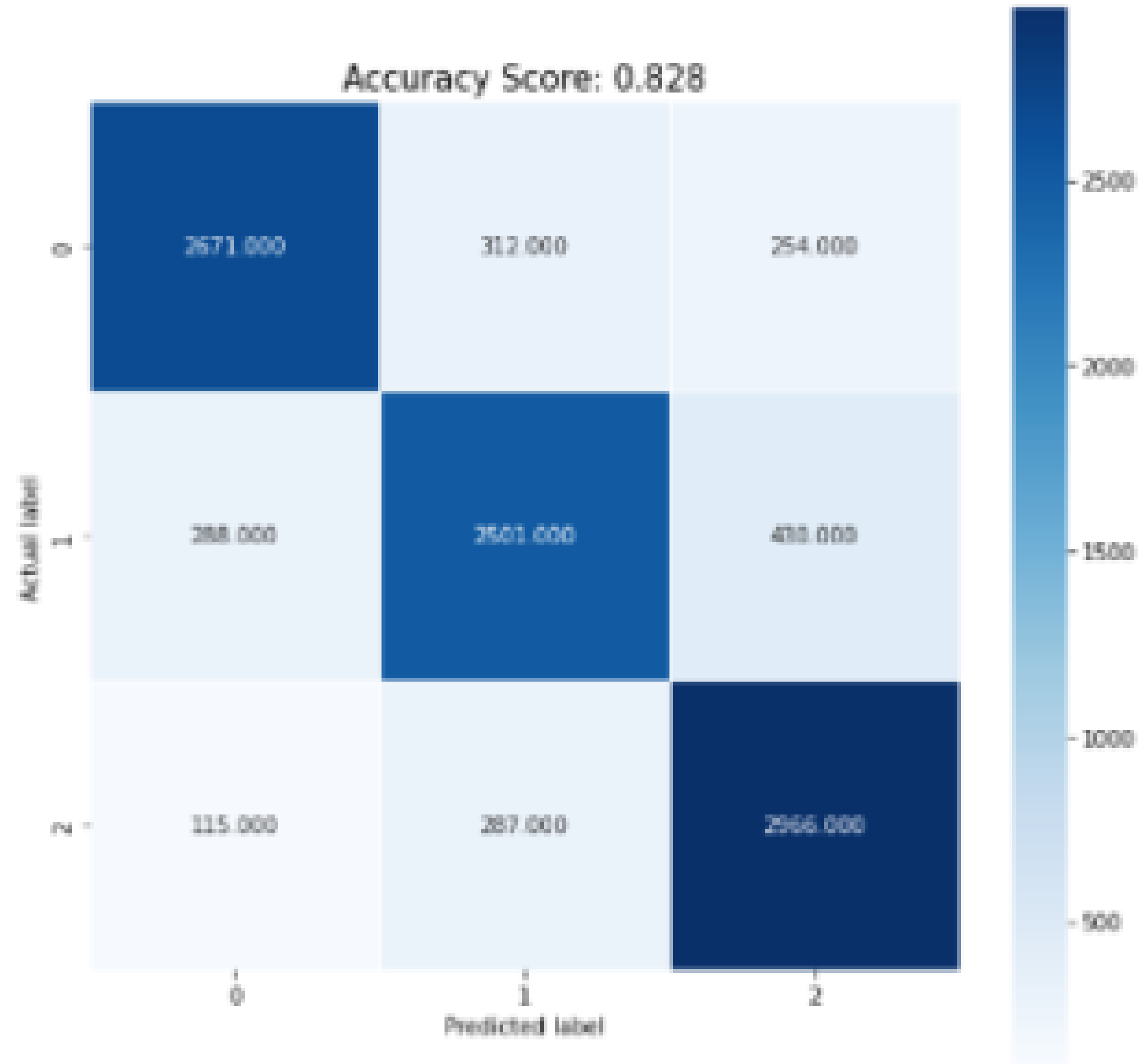


	Precision	Recall	F1 Score
Entailment	0.87	0.80	0.84
Contradiction	0.82	0.73	0.77
Neutral	0.77	0.91	0.83

GRU model

H. GRU on SNLI

Total Accuracy = 82.84 %

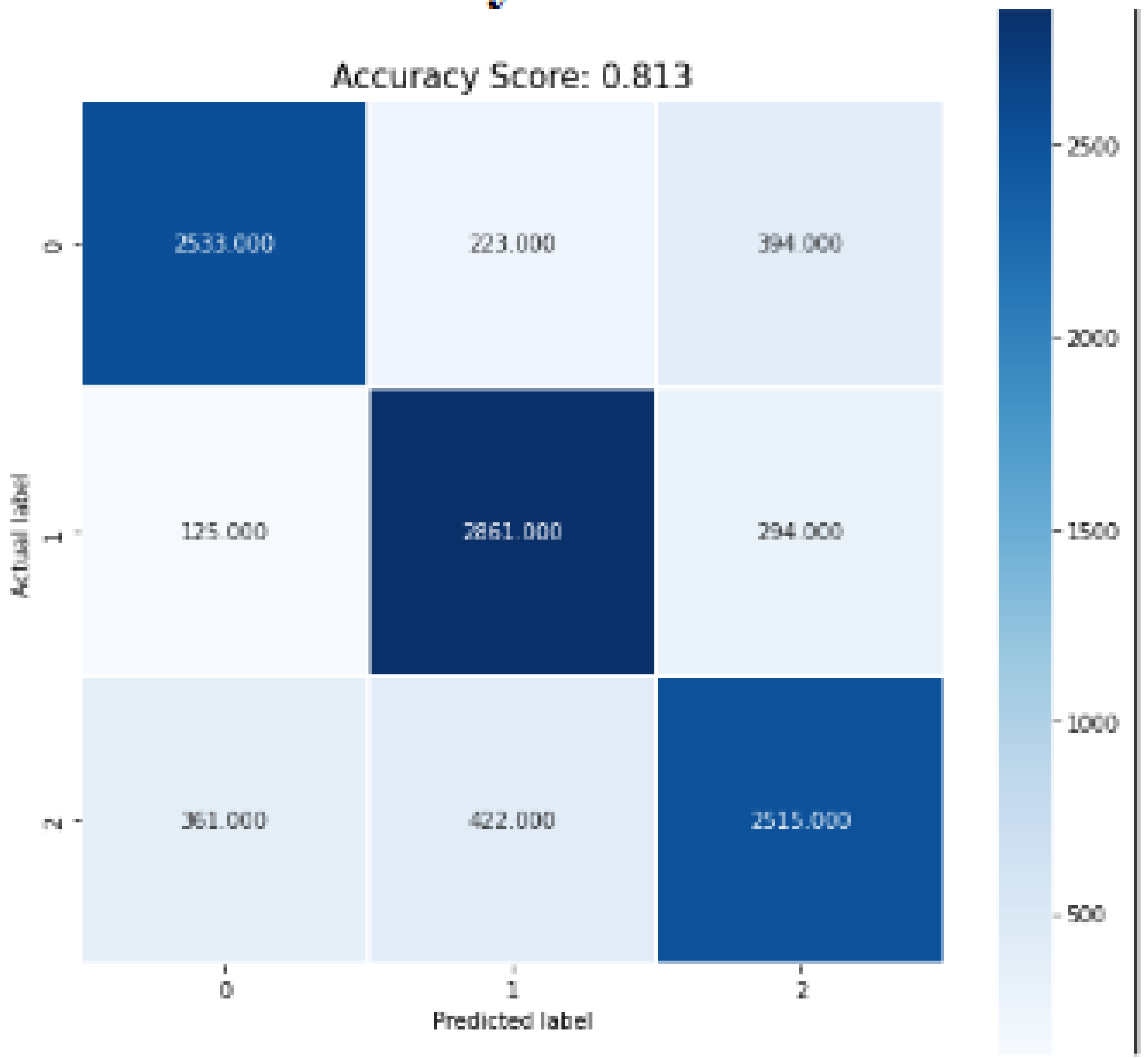


	Precision	Recall	F1 Score
Entailment	0.87	0.80	0.85
Contradiction	0.82	0.78	0.79
Neutral	0.81	0.88	0.85

BERT model

I. BERT on SNLI

Total Accuracy = 81.30 %



	Precision	Recall	F1 Score
Entailment	0.84	0.80	0.82
Contradiction	0.82	0.87	0.84
Neutral	0.79	0.76	0.77

Observations

F1-scores

The general trend of F1-scores among all models seems to favour entailment and neutral rather well, while doing relatively poorly for contradiction. However, the BERT model stands as an outlier, while not doing any worse for the other two.

Accuracy Scores

It is surprising to see the more complicated models like BERT, LSTM and GRU have similar scores to our baseline model

Performance-to-cost ratio

The larger, more complex models hence provide a lower performance-to-cost ratio when taking into account the memory and time needed to train these models

Thank you!

:)