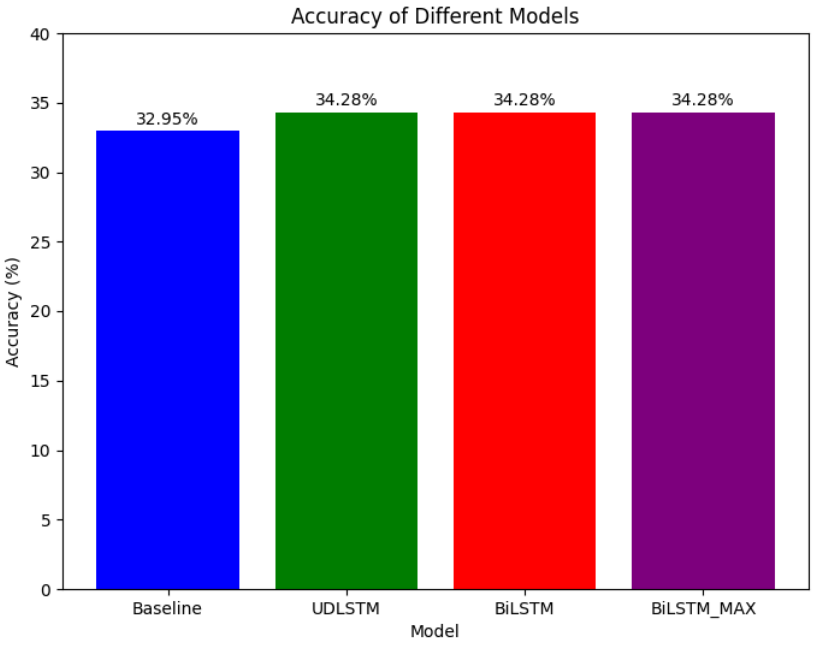
Results
As can be seen in the notebook, the models predict "entailment" regardless of the input. This means that our models do not score higher than random chance. This is reflected in the Accuracy of Different Models figure presented below.



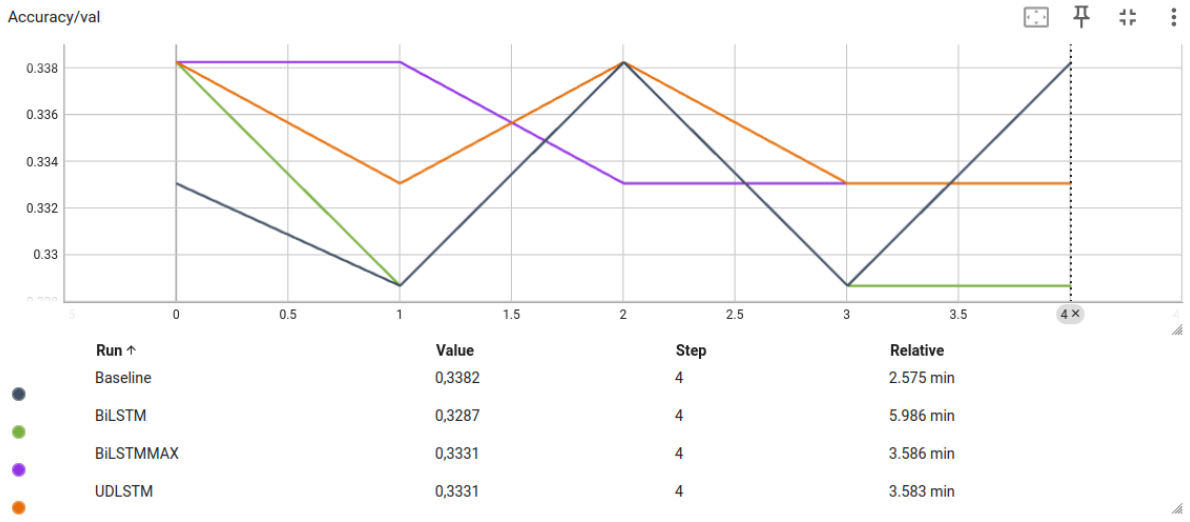
The minor variation in accuracy observed in the baseline model could potentially stem from how the models treated outliers, particularly instances with labels of -1.

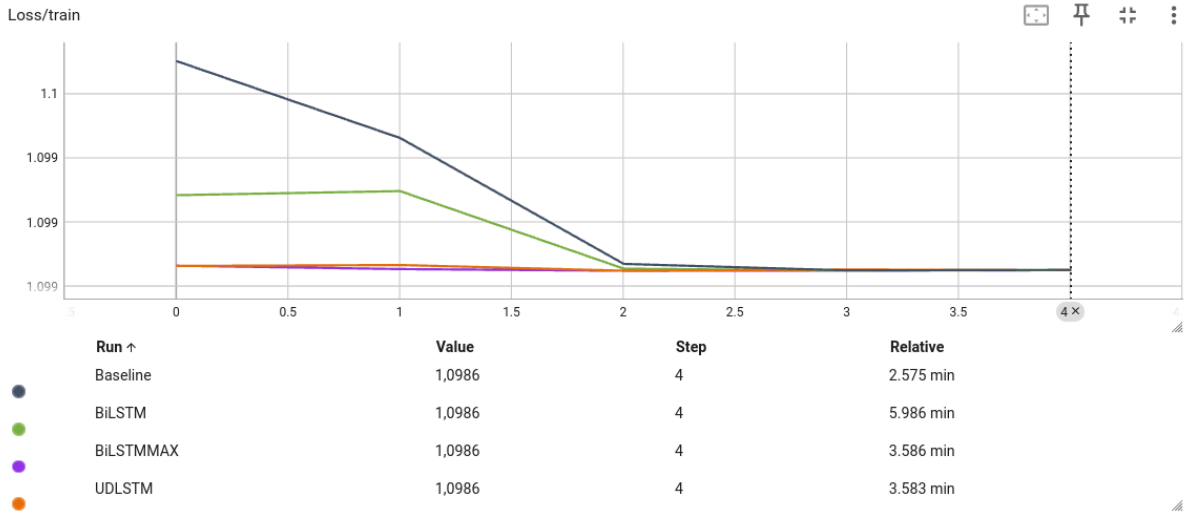
The uniform prediction of "entailment" across all models poses a significant challenge for comparison. The absence of the EvalSent implementation is due to time constraints as the primary focus was achieving good model performance. Without reliable model performance, any comparisons with SNLI or EvalSent results would lack significance and provide an unreliable basis for analysis and conclusions. However, unfortunately we have to acknowledge the presence of a bug affecting affected all four models. Despite this setback, we'll continue to analyse the models based on theoretical expectations.

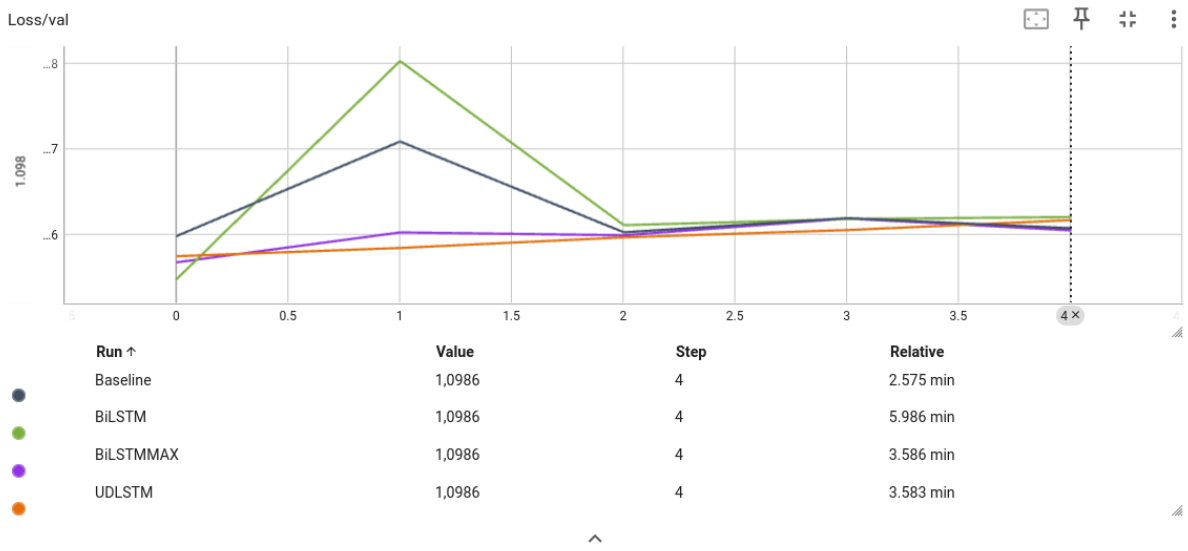
If all models had functioned as intended, we anticipate that the baseline would have shown the poorest performance due to its simplicity and lack of consideration for word order, syntactic structure, and word importance. This model specifically has trouble capturing meaning that relies on word order and/or context.
We believe the Undirectional LSTM would have performed better than the baseline as it can capture some sequential information based from the beginning to the end of a sentence. However, because of the models inability to process sentences backwards, it cannot work well with context that's given later in the sentence. This leads us to believe that this model specifically has trouble capturing long-range dependencies.
The Bidirectional LSTM would likely have performed even better by processing sentences in both directions, enabling it to capture past and future context. However, even though this model can capture context in both directions, it may still struggle with broader contexts.
Finally, we expect that the Bidirectional LSTM with Max Pooling would outperform all other models. This is because it has all the advantages mentioned for the bidirectional LSTM (it

can learn from past and future context) but it can also utulise Max Pooling to help the model focus on the most relevant information. Giving more attention to important features of the input, can help when model learn during training, resulting in a better performance. However, by using max pooling we lose a lot of information that could potentially be important. Using this method might lead to oversimplifactation in the embeddings.

Based on the findings presented in Conneau et al.'s paper (https://aclanthology.org/D17-1070.pdf), we see that the Bidirectional LSTM with Max Pooling surpasses other models in performance. Given that our models were implemented following the methodology outlined in this paper, with identical hyperparameters, model types, datasets, and implementation details including early stopping, we anticipated similar outcomes. However, our training process yielded unexpected results. As depicted in the Tensorboard graphs below, even thoughthe loss consistently decreases over epochs, the validation accuracy remains stagnant at approximately 33.0% to 33.8%. This suggests that the models fail to learn effectively during training, leading to bad performances.







Run	Value	Step	Relative
Baseline	1,0986	4	2.575 min
BiLSTM	1,0986	4	5.986 min
BiLSTMMAX	1,0986	4	3.586 min
UDLSTM	1,0986	4	3.583 min

To access the complete codebase, including checkpoints, runs, logs, and other files, please visit our GitHub repository: https://github.com/MariekePop/ATiCS_practical1
For detailed instructions on how to utilise the code, we refer to the README file provided in the repository.

Please note that in our implementation, we made the deliberate choice to align the vocabulary with the SNLI trainingset . This decision was based on the substantial size of the trainingset, ensuring a diverse range of words for effective training. By limiting the vocabulary to words present in the training set, we effectively manage the vocabulary size, consequently, the inference time, as the model only needs to consider embeddings for a smaller set of words.

Additionally, aligning the vocabulary with the training set results in a cohesive and in-domain vocabulary. The words in the training set are representative of the language used in similar contexts, such as in the validation and test sets. This means we reduce the likelihood of encountering out-of-vocabulary words during training and testing.

Examples sentences in the notebook
In the practical1_notebook_demonstration.ipynb we can see what that the models predict "entailment" for the following two examples (as it does for all datapoints):

Premise - “Two men sitting in the sun”
Hypothesis - “Nobody is sitting in the shade”
Label - Neutral (likely predicts contradiction)

Premise - “A man is walking a dog”
Hypothesis - “No cat is outside”
Label - Neutral (likely predicts contradiction)

However, we would expect them to predict contradiction for both examples, even though the label is actually neutral. This might be because because it has difficulty capturing a broader context (contextual understanding). Just because two men are sitting in the sun doesn't mean that nobody is sitting in the shade, or just because there is a man walking a dog, this doesn't mean that there are no cats outside. The model might not realise that both claims could be true simultaneously. When the models struggle with contextual understanding, we see that they inferring a relationship between different parts of a text are be a challenge, this could be what we see in these examples.

A possible approach to enhance a models contextual understanding without drastically increasing model computation during inference is injecting text data during training which is representative of contextually-relevant context that will be seen at inference. (Sainath et al.) https://ieeexplore.ieee.org/abstract/document/10096287?casa_token=8W6xXaGv4w8AAAAA:q21l3ahbSDEqz2qu6AKZCnDMEabWazgXRDTI9mA074if7X4Jyi2fWYPh9qN5BeUslFzKok-gqU8