Coursework Report

Chenrui Li 2760414L

Q1

1.1 How many questions and options are there in each split? [1 mark]

	Training Data	Validation Data	Test Data
total_questions	741	103	202
total_options	2964	412	808

1.2 What is the average number of tokens per question in the training set? [1 mark]

- According to the project output, the training set had an average of 6.2726 tokens in the problem

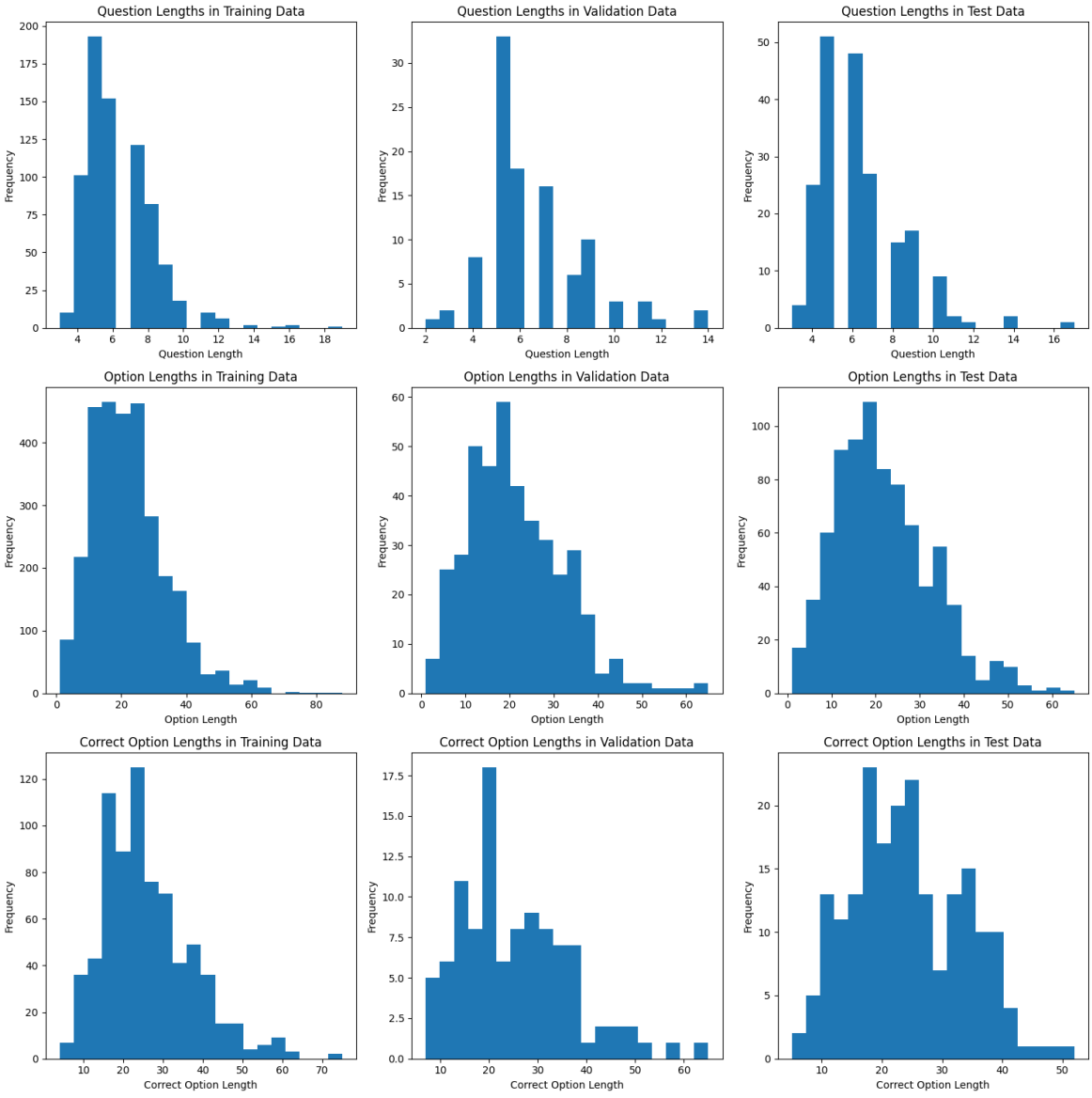
1.3 What is the average number of tokens per choice in the training set? [1 mark]

- According to the project output, the training set had an average of 22.3381 tokens in the options

1.4 What is the average number of tokens per correct choice in the training set? [1 mark]

- According to the project output, the training set had an average of 26.0324 tokens in the correct options

1.5 Perform any additional exploration of the data that you feel would be helpful for this multiple-choice question-answering task. Briefly describe what you found. [4 marks]



```
-------- Training Data --------
Most common words in questions: [('what', 392), ('is', 304), ('the', 254), ('in', 130), ('how', 116), ('who', 101), ('of', 94), ('a', 91), ('when', 81), ('are', 72)]
Most common words in options: [('the', 5020), ('of', 2577), ('and', 2078), ('in', 1754), ('a', 1705), ('is', 1331), ('to', 1086), ('as', 715), ('by', 614), ('or', 497)]
Most common words in correct options: [('the', 1366), ('of', 736), ('and', 604), ('a', 571), ('in', 528), ('is', 458), ('to', 285), ('by', 210), ('as', 197), ('or', 180)]
-------- Validation Data --------
Most common words in questions: [('what', 60), ('the', 50), ('is', 40), ('are', 16), ('of', 16), ('in', 15), ('was', 13), ('on', 12), ('who', 12), ('where', 12)]
Most common words in options: [('the', 641), ('of', 310), ('and', 261), ('in', 251), ('a', 216), ('is', 166), ('to', 141), ('as', 102), ('was', 83), ('by', 70)]
Most common words in correct options: [('the', 183), ('of', 91), ('a', 78), ('and', 72), ('in', 69), ('is', 50), ('to', 34), ('by', 30), ('as', 27), ('or', 21)]
-------- Test Data --------
Most common words in questions: [('what', 107), ('is', 83), ('the', 78), ('of', 40), ('how', 35), ('a', 31), ('in', 31), ('who', 31), ('was', 24), ('where', 20)]
Most common words in options: [('the', 1316), ('of', 657), ('and', 525), ('in', 469), ('a', 461), ('is', 348), ('to', 262), ('as', 204), ('by', 186), ('or', 155)]
Most common words in correct options: [('the', 328), ('of', 189), ('a', 149), ('and', 140), ('is', 125), ('in', 124), ('or', 65), ('to', 61), ('by', 59), ('for', 44)]
```

- From the output above, we can see the length distribution of questions and options in each dataset.
- We can see that most of the length distributions of questions are below 10 tokens, most of the length distributions of options are below 40 tokens, and most of the length distributions of correct options are below 50 tokens.
- This information is useful for limiting the length of inputs when we design our models.
- We can also see the most common words in each dataset.
- These words may be deactivated words, and we may consider removing them in the preprocessing step.
- In addition, we can see the most common words in the correct options, which may be keywords that our model needs to focus on.

Q2

2.1 Report the performance of each similarity measure (overlap coefficient, Sorensen-Dice & Jaccard) on the training and validation sets by measuring accuracy. [6 marks]

	Overlap Coefficient	Sorensen-Dice	Jaccard
Training	52.362%	42.915%	42.915%
Validation	46.602%	35.922%	35.922%

2.2 For each similarity measure, how many times was the score of the most similar answer tied with another answer? When there was a tied score among the top answers, how did you choose which to select? Why? [4 marks]

- In the training set, the overlap factor for this case was 246 times, 20 times for Sorensen-Dice and 20 times for Jaccard.
- In the validation set, the overlap coefficients for this case are 29 times, 4 times for Sorensen-Dice and 4 times for Jaccard.
- When there are multiple options with the same score, I select the first option with the highest score. This is because our model is index-based.
- This choice is based on the assumption that all options are equiprobable, so in the absence of other information, I choose the first maximum.
- If there is additional information that can be used to distinguish between tied maxima, we can modify this selection strategy.

Q3

3.1 Report the performance of the training and validation sets by measuring accuracy. Discuss how they compare with the set similarity measures from Q2. [6 marks]

- The accuracy of the training set is 37.922% and the accuracy of the validation set is 36.893%.
- In Q3, I used Term Frequency (TF) vectors and cosine similarity to measure the similarity between questions and options.
  - A TF vector is a method of converting text into a vector where each word is weighted by the number of times it appears in the text.
  - Cosine similarity is a measure of similarity between two vectors, which calculates the cosine of the angle between the two vectors.
- In Q2, I used the overlap coefficient, the Sorensen-Dice coefficient, and the Jaccard coefficient to measure the similarity between questions and options.
  - These are set-based similarity measures that compute the proportion of intersections and concatenations between the sets of word blocks of questions and options.

- These two approaches have their own advantages and disadvantages.
  - The TF vector and cosine similarity approach takes into account word frequency information, so it captures the effect of differences in word frequency on text similarity.
    - However, it does not take into account word order information, and thus may miss some semantic information.
  - On the other hand, set-based similarity measures only consider the presence or absence of word chunks without taking word frequency and word order information into account, and thus may miss out some important information.
    - However, they are usually computed faster than TF vector and cosine similarity methods.
- From the accuracy point of view, the model in Q3 performs less well compared to the similarity metric in Q2.
- This suggests that the BoW model may not be suitable for this task, a potential reason being that it ignores word order and context.
- In addition, the BoW model ignores the semantic information of the words, which may cause the model to have difficulty in selecting the correct answer.

3.2 Propose, motivate, and evaluate one modification to this process to improve this method. Report the performance on the training and development sets and compare them with the unmodified version. [6 marks]

- Due to the limitations of the BoW model, we need to try more sophisticated models to solve this task.
- Here, I tried the Word2Vec model, which is an unsupervised word embedding model that maps words into a low-dimensional continuous vector space so that similar words are mapped to neighbouring positions.
- Since the Word2Vec model is unstable each time it is trained, it needs to be averaged over several training runs.
- The averages for a particular 10 training runs I did on one occasion was
  - Training set: 38.834 per cent
  - Validation set: 37.183 per cent
- This suggests that the Word2Vec model is performing roughly the same as the BoW model, with a slight improvement.
- The potential reason for this is that the Word2Vec model takes into account the contextual information of the words and therefore it is able to capture the semantic and syntactic relationships between words. This may help to improve the performance of the model.
- However, the Word2Vec model has some limitations. For example, it ignores the order information of words, and thus may miss some important information.
- To further improve the performance of the model, we can try more sophisticated models, such as BERT or GPT-3. These models are Transformer-based models, which are able to capture the complex semantic and syntactic relationships between words, and thus may be more suitable for this task than the Word2Vec model.

Q4

4.1 Report the performance of the training and validation sets by measuring accuracy. [8 marks]

- The accuracy of the training set is 14.305% and the accuracy of the validation set is 20.388%.

4.2 What are the limitations of the set similarity and cosine similarity methods used in Q2, Q3 and Q4? [4 marks]

- Limitations of set similarity:
  - Ignores the importance of elements, all elements are treated as equally important, which is not always the case in natural language.
  - Punching lacks contextual information and focuses only on the presence or absence of elements, ignoring the relationships and semantics between them.
  - Considers only the presence or absence of words and ignores word frequency, which can be very important for text similarity.
  - Disregarding the order of words, which is crucial for understanding the context and semantics of a sentence.
- Limitations of cosine similarity:
  - Ignores differences in size and focuses only on the direction of the vectors, ignoring the size of the vectors.
  - Similar to set similarity, cosine similarity also ignores frequency and order information of words, as well as contextual information. This information may be important for text similarity.

Q5

5.1 Report the accuracy, precision, recall and F1 score of the predictions on the question-option pairs representation of the training and validation sets [10 marks]

	Accuracy	Precision	Recall	F1 Score
Training	96.862%	92.188%	95.547%	93.837%
Validation	81.331%	63.830%	58.252%	60.914%

5.2 Report the accuracy for this method for selecting the correct answer on the training and validation sets of this model. Note this is different from the value in part (a). To enable this, select the

option for each question with the highest output logit value for the positive class of the model. [6 marks]

- The accuracy of the training set is 92.713% and the accuracy of the validation set is 56.331%.

5.3 Why would you expect this approach to outperform the use of [CLS] vectors described in Q4? [2 marks]

- In Q4, we use a context vector of [CLS] tokens to represent the entire question-option pair. This approach relies on the representational power of the [CLS] tokens in a pre-trained BERT model that is designed to capture the overall meaning of the entire input sequence (question + option). However, this approach has several potential limitations:
  - Information loss: the [CLS] vector is a representation of the entire sequence, which may lose some important information between questions and options, such as the order of words and contextual information.
  - Task-specific optimisation: [CLS] vectors are learnt in the pre-training phase and may not always be optimally tuned to reflect the semantic relations of a particular task (e.g., a multiple-choice quizzing task), especially if the fine-tuning is not sufficient.
  - Contextual generalisation ability: although the [CLS] vector is pre-trained to capture the overall meaning of the sequence, it may not always be optimally tuned to reflect the semantic relations of a particular task (e.g., a multiple-choice quiz task), especially when fine-tuning is inadequate.
  - Missing fine-grained information: the direct use of [CLS] vectors may not adequately capture the nuances between the question and the individual options, which are crucial for selecting the best answer. This approach may ignore specific, fine-grained interaction information between the question text and the options.
- In contrast, the approach described in Q5 directly targets the sequence classification task (in this case the task of choosing the correct answer) via the BERT model, and is expected to perform better for a number of reasons, including:
  - Contextual information is taken into account: through Fine-tuning, the BERT model is better able to capture the contextual information between questions and options, thus improving the model's performance.
  - Task-specific optimisation: with Fine-tuning, all layers of the BERT model are optimised for a specific task, rather than just relying on the [CLS] vectors from pre-training. This means that the model is able to learn representations that are better suited to specific questions and answer selection tasks.
  - Better semantic understanding: Fine-tuning enables the model to better understand the semantic relationships between questions and options because it is trained on task-specific data and is able to capture the nuances that are critical to choosing the correct answer.
  - Exploitation of fine-grained interaction information: by training on task-specific data, the Fine-tuned model is able to exploit fine-grained interaction information between questions and options, which may help to identify correct answers more accurately.
  - In addition, we used the output logit values of the BERT model to select the best answer for each question, rather than simply selecting the best answer for the entire question-option pair. This approach allows for better differentiation of each option, which improves the performance of the model.
- Therefore, we expect this approach to outperform the approach in Q4. The results show that this method is indeed better than the method in Q4.
- In summary, Fine-tuning the entire BERT model and applying it directly to the task of choosing the best answer provides deeper semantic understanding and task-specific optimisation, which is expected to achieve higher performance, compared to methods that only use [CLS] vectors.

Q6

6.1 Report the accuracy using your best method on the test set. Use the performance on the validation set to select the best method. [2 marks]

- After the above analysis, we have chosen the method in Q5 as the best method.
- Using the method in Q5, the accuracy of the test set is 55.941%.

6.2 Discuss whether the achieved accuracy would be sufficient for deployment [2 marks]

- An accuracy of 55.45 per cent may not be sufficient for deployment. This depends on the following:
  - The nature of the task:
    - Some applications may tolerate lower accuracy rates, especially if the system's purpose is to provide initial screening or recommendations to experts, while the final decision is still made by humans; or if the model is used only for academic research or educational purposes.
    - However, for applications that rely heavily on automated precision results, such as medical diagnostics or safety-related systems, such an accuracy rate may not be sufficient; or if the model is used for commercial applications, such as automated customer service or online exams, then an accuracy rate of 55.45% may not be sufficient either.
  - Benchmark Comparison: If there is no ready-made solution for the task at hand or the accuracy of the existing solution is less than 55.45%, then this model may be good enough. On the other hand, if better known methods exist, then this accuracy may not be sufficient to justify deployment.
  - The cost of error: the potential impact or cost of misclassification needs to be assessed. In some cases, an incorrect prediction may lead to serious consequences, while in other cases the error may be only a minor inconvenience.
  - User expectations: Users' expectations of accuracy can affect their trust in the system and their willingness to use it. If users have high expectations of accuracy, then 55.45% may lead to dissatisfaction.

    - Potential for improvement: If there is room for further optimisation of the model performance, one may choose not to deploy it for the time being and re-evaluate it after the model has been improved. For example, alternative model architectures, data enhancements, more data, or different preprocessing strategies could be tried.
    - Learning after deployment: if this is a system that can continue to learn and improve, it may be possible to accept the current accuracy and expect to improve performance over time through online learning or regular model updates.
    - In addition, we need to consider aspects such as robustness, interpretability and fairness of the model to ensure that it will perform well in practice.
- In summary, the 55.45% accuracy rate may be acceptable for some deployment scenarios, but may not be sufficient for applications that require high accuracy. Decisions need to be made based on specific application scenarios, risk assessment, and user expectations. All possibilities for improving model performance should also be explored before making deployment decisions.