**Section 4(in the report):**

Experiment 1:

-Consider incorporating a wider range of hyperparameter tuning or a learning rate scheduler to potentially enhance model generalization.

-Documentation on data preprocessing, handling class imbalance, and augmentation techniques. A detailed description of how the dataset was split into training and validation sets is necessary.

-Provide a rationale for the specific choice of learning rate and batch size. Experiment with different optimizer algorithms or weight initialization strategies for potential performance gains.

-Incorporate a validation loss curve alongside the training loss curve to visually compare and detect overfitting. For the confusion matrix, it would be beneficial to also report precision, recall, and F1 score for each class to assess performance thoroughly.

-Since there's a noticeable bias towards the majority class in the confusion matrix, consider techniques like SMOTE for oversampling minority classes and visually displaying the improved balance.

Experiment 2:

-Explore a wider range of hyperparameters, possibly using grid search or random search methodologies. Additionally, provide more context on why specific values were chosen for learning rates and epochs.

-It might be good to include a more detailed description of the experimental setup, such as the preprocessing steps, the tokenizer configuration, and the rationale behind the specific BERT model chosen.

-Trade-offs associated with each hyperparameter setting, such as computational costs versus performance benefits. Detail the implementation process more thoroughly, including any challenges faced and how they were overcome.

-Include additional metrics such as precision, recall, and specificity alongside the F1 score to provide a fuller picture of model performance. Offer a deeper analysis of the confusion matrix, explaining the possible reasons for the misclassifications observed.

-Consider including other forms of visualization, like ROC curves or Precision-Recall curves, for a more nuanced view of model performance.

Experiment 3:

-The rationale behind choosing specific hyperparameters and their expected impact on the results.

-The setup could also experiment with different optimizers or learning rate schedules to justify the choice of Adam and a static learning rate.

-The visuals could be improved by including additional metrics, such as precision, recall, and possibly a ROC curve, to give a more holistic view of the model's performance. Also, ensure that the explanation of the test results includes potential reasons for the observed performance and discuss any unexpected findings.

-Include more detailed analysis of the confusion matrix to understand the kinds of errors made. Additional visuals, such as precision-recall curves for each class, might also provide deeper insights.

Experiment 4:

-The report should discuss the rationale behind choosing the specific range and values of hyperparameters tested.

-Explanation in detail of the data preprocessing, the architecture of the RNN, the training procedure, and the evaluation metrics. Each step should be reproducible based on the information provided.

-The setup, including the learning rates and epochs, shows a good way of the hyperparameter tuning process. The implementation's success is evident from the improvements in F1 score, even though they were modest. The justification for the setup would be more convincing if it included a discussion on why these particular values and ranges were chosen.

-Include statistical tests if possible to confirm the significance of the observed improvements. Ensure that each visual has a descriptive analysis in the text that not only presents the results but also explains why they are significant.

-To further enhance the picture of model performance, you could include more metrics such as precision, recall, and a detailed breakdown of the F1 score by class. Additionally, visualizing the learning rate against performance could further illustrate the sensitivity of the model to hyperparameter changes.

-The confusion matrix indicates that model has issues with certain class distinctions. This suggests a need for improvement in the model architecture or training data. Consider adding a discussion about possible reasons for misclassification and potential ways to address them.

**Section 2-3(in the report):**

Figure 1(Label Distribution): Shows a clear imbalance in the dataset with a majority of 'B-O' labels and fewer instances of 'B-AC' and 'I-AC'. This could affect the model's ability to learn less frequent labels.

-Consider using a logarithmic scale to make differences in smaller categories more noticeable. Also, include the actual frequency values on the bars for more precise interpretation.

Figure 2 (Token Length Distribution): Depicts the distribution of tokens' length in a histogram overlaid with a curve.

-Ensure that the overlaying curve accurately represents the data distribution. If it's a normal distribution fit, this might be misleading if the data isn't normally distributed. Instead, a kernel density estimation curve could be more appropriate.

Figure 3 (POS Tag Distribution): Shows the frequency of different POS tags in the dataset. Nouns and proper nouns dominate, which could be indicative of the scientific nature of the text.

-Include a pie chart for a complementary view of the proportional distribution of POS tags.

-Justify the use of certain models over others based on the data (e.g., BERT might be more suitable given the predominance of specific POS tags).

-Explain how the label imbalance might be addressed in training.

-Offer direct connections between the visual data representations and the implications for the modeling process. For example, how might the token length distribution affect the choice of BERT's maximum sequence length?

-Address any issues with data quality, such as missing values, outliers, or inconsistencies, and how these were handled.

-Label Distribution: Provide more detailed statistics such as the mean, median, and standard deviation of label frequencies, as well as discussing potential strategies for dealing with the imbalance (e.g., SMOTE for oversampling).

-Perform a preliminary error analysis, which could be insightful for understanding which types of tokens or labels are more challenging for models to handle.

**Section 5-6(in the report):**

-Provide more specific examples or case studies where the suggested improvements (like character-level embeddings and domain-specific knowledge) have been applied successfully. Discuss potential challenges or considerations in implementing these improvements, such as data availability, computational overhead, or integration with existing systems.

-Offer a deeper dive into why specific improvements (e.g., data augmentation techniques like synonym replacement) might enhance model performance.

-Discuss potential mitigation strategies for balancing these trade-offs, such as implementing lighter versions of BERT or hybrid models that could offer a middle ground.

-Highlight specific instances or qualitative examples from data where the model successfully identified complex cases or failed, thus providing insights into both strengths and areas for improvement.

-Include a section on future research directions that could potentially address the remaining challenges highlighted by the experiments, such as improving the model's ability to distinguish between similar abbreviations.

Question 5: Evaluation of the overall attempt, and great explanation on the choice between the most accurate against the most effective solution(in our case not in general). Explanation on how the original problem is solvable and how the experiments helped to understand this.

In the paragraph of section 6, it is right the F1 score written there?

Question 4: How to improve performance in each experiment done(not in general) and Some explanation of the results achieved.