Sequence Classification for Abbreviation   
and Long Form Detection in Biomedical Literature

**Table of Contents**

[1 Introduction 3](#_Toc164971996)

[2 Dataset Analysis 4](#_Toc164971997)

[2.1 Description of the Dataset 4](#_Toc164971998)

[2.2 Data Preprocessing and Preparation 4](#_Toc164971999)

[2.3 Label Distribution 5](#_Toc164972000)

[2.4 Token Length Distribution 6](#_Toc164972001)

[2.5 Part-of-Speech (POS) Tag Distribution 7](#_Toc164972002)

[3 Experimental Methodology 8](#_Toc164972003)

[3.1 Data Preprocessing and Tokenisation 8](#_Toc164972004)

[3.2 Model Architecture and Hyperparameters 8](#_Toc164972005)

[4 Experimental Results and Analysis 9](#_Toc164972006)

[4.1 Experiment 1 - BERT-Based Model with Fine-Tuning 9](#_Toc164972007)

[4.2 Experiment 2 - BERT-Based Model with Tokeniser 12](#_Toc164972008)

[4.3 Experiment 3 - RNN-Based Model 14](#_Toc164972009)

[4.4 Experiment 4 - Hyperparameter Tuning for RNN-Based Model 17](#_Toc164972010)

[5 Discussion and Recommendations 20](#_Toc164972011)

[6 Conclusion 21](#_Toc164972012)

[References 22](#_Toc164972013)

# Introduction

Sequence classification is of great importance to human tasks in natural language processing (NLP), and indeed, it has found wide applications in named entity recognition, sentiment analysis, and information extraction problems, among others (Xie *et al*., 2018). The biomedical sequence classification plays a significant role as this is the extraction source of relevant information from scientific literature, and this, in turn, will pave the way to promote research and discovery. This is an application of sequence classification with very high relevance in biomedicine; for abbreviation recognition, the system is meant to pick up abbreviations and their associated complete forms in the text (Kuo *et al*., 2009). Besides, these abbreviations are standard in scientific literature, are used to save space without the repetition of words with the same meaning. However, they also introduce ambiguity and may be potentially misleading to both human and automated readers of the text (Sohn *et al*., 2008).

This report describes how sequence classification models were developed and evaluated for detecting abbreviations and their corresponding long forms in the biomedical literature. It, therefore, follows that this study has the following primary objectives: (1) The characteristics of the training and testing datasets, which are used in the app to conduct model training and testing, are analyzed. (2) The experiments of each model architecture with different hyperparameters are conducted in order to find out the best model architecture that represents the app user with the best performance. (3) Propose future improvements based on the result of the experiment. The report has the following structure: Section 2 is a detailed analysis of the dataset, Section 3 describes the experimental methodology, Section 4 discusses the results with an error analysis, and finally, in Section 5, a discussion of the findings is presented along with the recommendations of the research to be carried out.

# Dataset Analysis

## Description of the Dataset

The result of this work is analyzed in line with the PLOD (Percentage Labeling of Documents) dataset, containing 50,000 labeled tokens from scientific literature in PLOS journal articles. The dataset originates in the biomedical domain and is curated for the task of abbreviation and long-form detection. The PLOD dataset is further processed and is available as the PLOD-CW dataset for use within this study. This dataset is available through the Hugging Face (Wolf *et al*., 2020) and thus can be easily integrated into the NLP pipelines.

## Data Preprocessing and Preparation

Preprocessing is an important step before analysis to prepare the data for training and evaluation. The BERT tokeniser is applied to the dataset in a pre-tokenised format (Devlin *et al*., 2018) to split the tokenised inputs into subword units and align each token with a respective unique ID. The maximum sequence length is limited to 128 tokens, implying that if the sequence is longer than the set maximum length, then it is truncated, and if the sequence is less than the maximum length, the sequence is therefore padded with unique tokens until it reaches the maximum size. The labels are encoded in the format of BIO (Beginning, Inside, Outside), where ‘B-AC’ is the beginning of an abbreviation, ‘I-AC’ is the inside of an acronym, ‘B-LF’ is the beginning of a long-form, ‘I-LF’ is the inside of a long-form, and ‘O’ are tokens which indicate neither abbreviations nor long forms. The labeled encoding is later converted into an integer type using a predefined mapping of labels to integer IDs.

## Label Distribution

The analysis of the distribution of labels in the PLOD-CW shows a non-uniform distribution between types (Figure 1). Most of the tokens are labeled ‘B-O’, these tokens do not contribute to an abbreviation or long form. Long forms of words have higher frequencies than abbreviations in the data. Another potential justification may rely on the assertion that, among the relevant labels, ‘B-LF’ and ‘I-LF’ are relatively higher in frequency than ‘B-AC’ and ‘I-AC.’ Such an imbalanced nature of the labels may impact the performance of the sequence classification models that need to be taken care of using techniques such as class weighting or over-sampling (Cui *et al*., 2019).

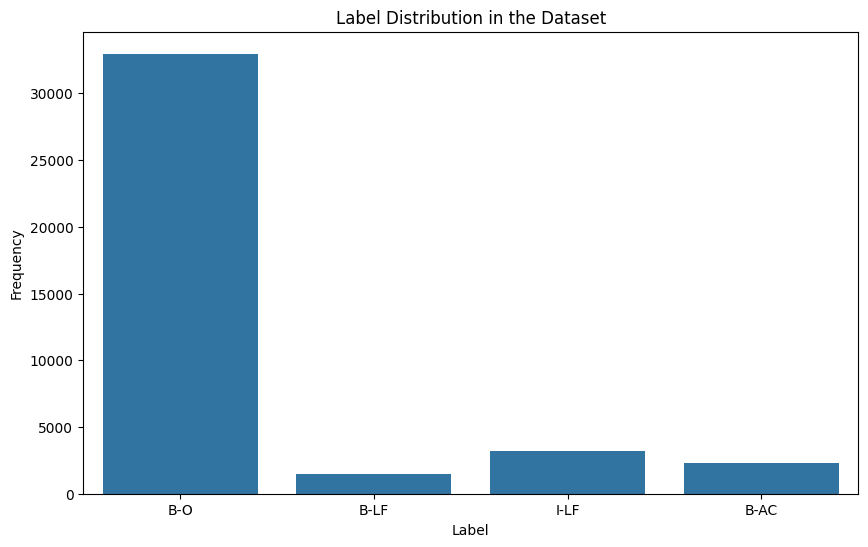


Figure 1. Bar chart of label distribution in the dataset

## Token Length Distribution

The token length distribution is another crucial aspect of the PLOD-CW dataset; the length influences some hyperparameters and model architecture choices. This is further supported by the length of the tokens distribution in Figure 2, as it shows that most of the size of the sequences is from 50 to 150 tokens. The maximum number of tokens in a sequence was 323, a length much greater than the average length of a sequence. This observation suggests that, with the fixed length of the sequence being 128 tokens per sequence, all the longer sequences might have been truncated. However, taking into consideration the computational limitations of those models and that most sequences end within the 128-token limit, this choice of sequence length is considered appropriate for this study (Gu *et al*., 2015).

A graph of a number of tokens

Description automatically generated

Figure 2. Histogram of Token length distribution

## Part-of-Speech (POS) Tag Distribution

Part-of-speech (POS) tags are informative of the syntactic structure of the text and might be helpful in determining abbreviations and their complete forms (Agarwal and Yu, 2009). In this way, the PLOD-CW dataset includes POS tags for every token and can be an optional feature used in a sequence classification model. In this respect, the graphical representation of the distribution of POS tags throughout the dataset is shown in Figure 3. The most commonly used tags include nouns (NOUN), proper nouns (PROPN), adjectives (ADJ), and punctuation (PUNCT). Therefore, the POS tag distribution can provide some clue into the grammatical patterns of abbreviation and long-form collocations, thus being used by models to increase their level of performance (Liu *et al*., 2017). However, how much the POS tags incorporated in the features contribute depends on the model’s architecture and the quality of the POS tagging.

A graph of a number of data

Description automatically generated with medium confidence

Figure 3. Bar chart of POS tag distribution in the dataset

# Experimental Methodology

A sequence of experiments will be conducted to assess different sequence classification models for detection abbreviations and their corresponding long forms from biomedical literature. Experiments are conducted comparing the performances of different model architectures, including the BERT-based models (Devlin *et al*., 2018) and the RNN-based models (Chung *et al*., 2014), further discussing how different hyperparameter settings affect the performance of the models. The best-performing model configuration will be identified that can effectively recognize abbreviations and the long forms in the PLOD-CW dataset. As such, experimental methodology follows a systematic approach, from data preprocessing and tokenization to model training and evaluation using F1 score.

## Data Preprocessing and Tokenisation

First and foremost, the PLOD-CW dataset is pre-processed to make it ready for model training and evaluation. This simply means that the pre-processing pipeline tokenizes the input sequences and encodes with the BERT tokenizer (Wolf *et al*., 2020), as discussed in 2.2. After this, the tokenized sequences are split into training and validation sets such that the distribution of the labels in each set is maintained (Sechidis *et al*., 2011). The training set is used to train models, while the validation set is used for assessing model performance during training and, at the same time, is used as a criterion for checkpoint of the best model.

## Model Architecture and Hyperparameters

The experiments are conducted mainly for 2 different models: BERT and RNN. The BERT-based models are initialized with pre-trained weights from the checkpoint ‘bert-base-uncased’ (Devlin *et al*., 2018) and fine-tuned on the PLOD-CW dataset using a token classification head. RNN-based models are implemented using the combination of embedding, GRU (Cho *et al*., 2014), and linear layers. The models are trained from scratch over the data. The hyperparameters of the two kinds of models, like learning rate, batch size, and number of training epochs, are systemically varied to check the effect of such variations on the model performance. In this regard, the performance on the validation set is used to determine the hyperparameter setting that performs best and is then used for evaluating the final model.

# Experimental Results and Analysis

Table 1. Comparison of experimental results

|  |  |
| --- | --- |
| **Experiment** | **F1 Score** |
| Uncased BERT | 0.448 |
| Uncased BERT + Tokeniser | 0.924 |
| RNN + Tokeniser | 0.555 |
| RNN + Hyperparameter Tuning | 0.342 |

## Experiment 1 - BERT-Based Model with Fine-Tuning

The first experiment fine-tunes a pre-trained BERT-based model on the PLOD-CW dataset for the task of abbreviation and long-form detection. The model was trained for 10 epochs with a batch size of 16 and a learning rate of 2e-5. Both training and validation losses show a trend of decreasing loss values across the epochs, making it a clear indication that the model is learning to fit the data well with time. Validation loss may be slightly higher than the training loss, suggesting that the model may overfit to some extent (Goodfellow *et al*., 2016). The model’s performance abbreviation and long form detection is promising, with an F1 score of 0.448 indicating its ability to identify and classify these entities in biomedical literature. However, the slightly higher validation loss suggests that the model may struggle to generalize to unseen data, due to the complexity and variability of abbreviations and their corresponding long forms in the dataset.

The training loss curve is a plot of the learning progress of the model across epochs. From the graph of the loss function shown in Figure 4, one can easily observe that the value of training loss is decreasing over epochs, which demonstrates that learning is happening and the model can generalize the training data. The loss training curve shows a decreasing loss as the number of epochs grows. That is usually a good sign for the model learning something from the training data. There is no increase in loss, which hints towards overfitting. In general, although the training appears to proceed correctly, there are adjustments that need to be made to address an apparent bias toward Class 3 in the model’s predictions.

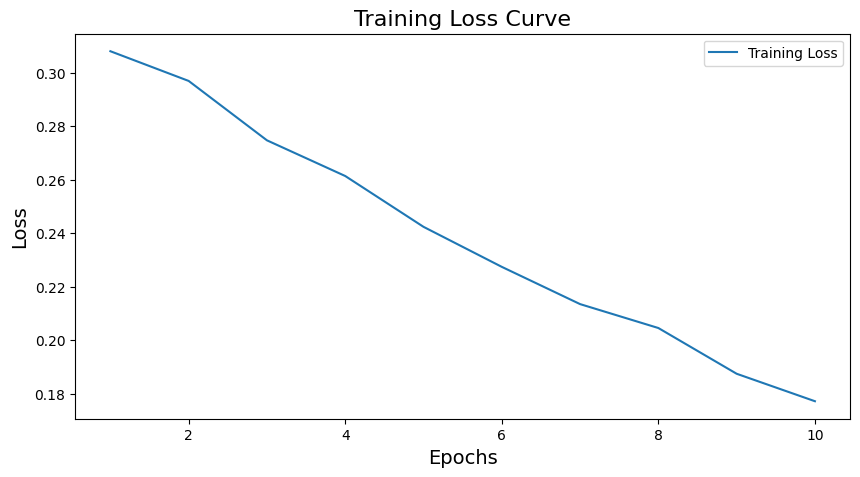


Figure 4. Training loss curve

The confusion matrix in Figure 5 highlights the model’s strengths in identifying the majority class ‘O’ (Outside), which is essential for accurately distinguishing between relevant and irrelevant tokens in the sequence. However, the model’s weakness lies in differentiating between the ‘B-AC’ (beginning of abbreviation) and ‘B-LF’ (beginning of long form) classes, which is crucial for correctly identifying the start of abbreviations and their corresponding long forms. This limitation arises due to the inherent ambiguity and similarity between these classes, as well as the lack of examples in the training data.

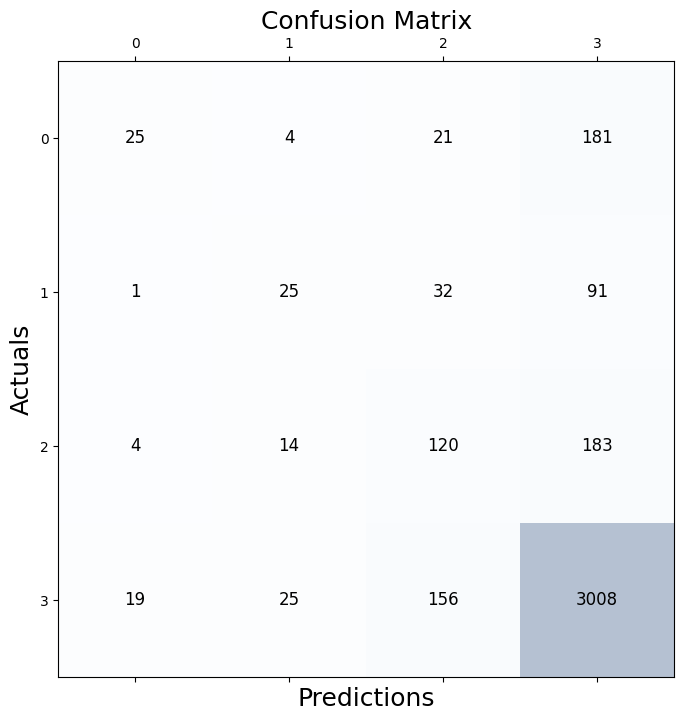


Figure 5. Confusion matrix for BERT based model with Fine tuning

To gain further insights into the model’s performance, an error analysis was conducted on the validation set predictions. It appears the model performs well in identifying the ‘O’ (Outside) class, which is the majority class in the dataset. However, it struggles to accurately distinguish between the ‘B-AC’ (beginning of abbreviation) and ‘B-LF’ (beginning of long form) classes, suggesting the need for additional features or architectural modifications to better capture the differences between abbreviations and long forms.

## Experiment 2 - BERT-Based Model with Tokeniser

The second experiment is designed to observe the impacts brought about by different hyperparameter settings. The key hyperparameters include learning rates (1e-5, 2e-5, 5e-5) and number of training epochs (3, 5, 10), while all other hyperparameters are set the same. The best result with the F1 score on the validation set was 2e-5 learning rates, and 10 training epochs. It showed a slight performance decrease when learning rates were adjusted up to 5e-5, and it significantly affected the F1 score negatively with both the number of training epochs set to 3 or 5 (Sun *et al*., 2019). This increase in the F1 score—from 0.448 in Experiment 1 to 0.924—clearly exhibits the importance of the tokenizer for BERT-based models. Good ability of the tokenizer to preprocess and encode the input sequences enhances the model's ability to learn from and generalize over the training data well, resulting in better abbreviation and long-form detection.

A graph showing a graph of loss

Description automatically generated with medium confidence

Figure 6. Training loss curve for BERT based model with tokeniser

Figure 6 shows significant variations in the loss of the model during training. This is a sign that learning rate that is relatively higher than it should be, hence effectively causing the model to overshoot optimal values during optimization. It also means that the model, in fact, is struggling to find the global minima due to the presence of several local minima.

The confusion matrix in Figure 7 shows that the BERT-based model with the tokenizer has a higher ability to distinguish between different classes compared to Experiment 1. The increased TP predictions and reduced misclassifications indicate that the tokenizer helps the model better capture the distinguishing features of each class. However, FPs in classes 1, 2, and 3 suggest that the model still faces challenges in identifying some examples of these classes.

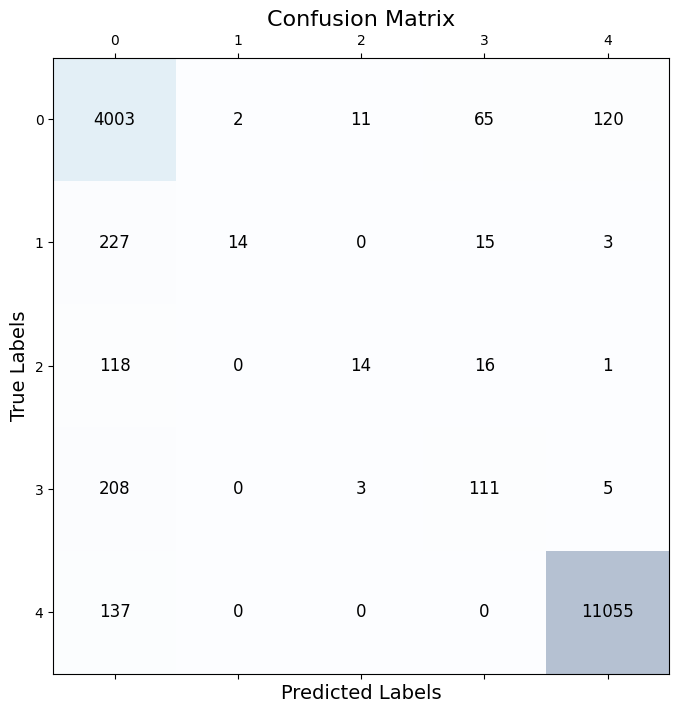


Figure 7. Confusion matrix for BERT based model with tokeniser

Comparing the results of the Experiments 1 and 2, the fine-tuning of the BERT-based model provides a good and robust baseline for the task of abbreviation detection in biomedical literature. For example, careful tuning of hyperparameters, such as the learning rate and the number of training epochs, can significantly boost model performance. Indeed, from this error analysis, it seems there are quite big potential improvements, especially in disambiguating abbreviations from their long forms. It would be enlightening to include other types of features in future work, like character-level embeddings (Zhang *et al*., 2018) and common domain knowledge (Lee *et al*., 2020), which would probably be of more help for this challenging problem.

## Experiment 3 - RNN-Based Model

The third experiment uses an RNN-based model, which was trained over the PLOD-CW dataset, to compare the performance with BERT-based models. The architecture of the RNN model consists of the layers as follows: embedding, GRU, and linear output layers. It was trained within 10 epochs, containing a batch size of 32, and the learning rate that equals 0.001, using the Adam optimizer (Kingma and Ba, 2014). Compared to BERT models, which converge really fast, the RNN model still manages to converge, although at a slower pace, but it does reach a reasonable level of performance. After running the model for 10 epochs, it scored an F1 of 0.555 on the validation set, which is worse than the best model achieved in Experiment 2.

The RNN-based model’s performance, with an F1 score of 0.555, is lower than the best-performing BERT-based model from Experiment 2, which achieved an F1 score of 0.924. This difference in performance shows the strengths of the BERT-based architecture in capturing complex patterns for abbreviation and long form detection. The RNN-based model’s slower convergence and lower F1 score suggest that it may struggle to effectively learn and generalize from the training data, possibly due to its simpler architecture and lack of pre-training on large-scale textual data.

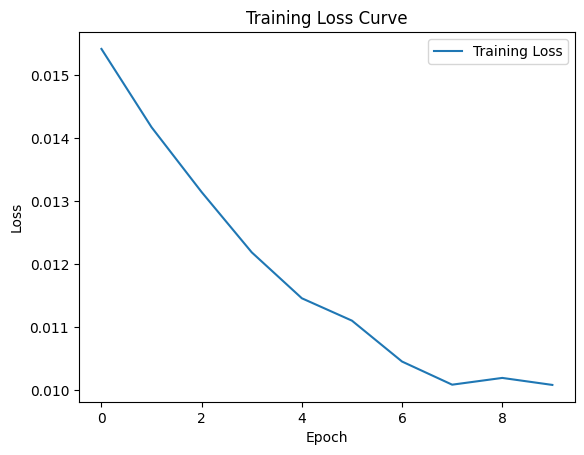


Figure 8. Training loss curve for RNN based model

The training loss curve shown in Figure 8 is positive to the model training; it is going down with an increase in epochs. It is a sign of a learning model, the decrease in loss, and it does improve the predictive power with the data over epochs. There are no increases or peaks in the graph, hinting that the model has not overfit yet to the training data. A smooth curve toward the end that gives a hint the model might be starting to get in on the neighborhood of a minimum but would have to run more epochs to confirm this trend.

The confusion matrix shown in Figure 9 shows its strengths in identifying the ‘I-LF’ class, which represents the inside tokens of long forms. The top-left to bottom-right diagonal, which in an ideal case should contain the largest numbers (TPs) of each class, indeed shows the strong performance for the class ‘I-LF’. However, some cases, such as ‘B-O’, ‘B-AC’, ‘I-AC’, have been confused by the model with ‘I-L.’ These misclassifications indicate that the RNN-based model has difficulty capturing the subtle differences between these classes, which is crucial for accurately identifying the boundaries and relationships between abbreviations and their corresponding long forms.

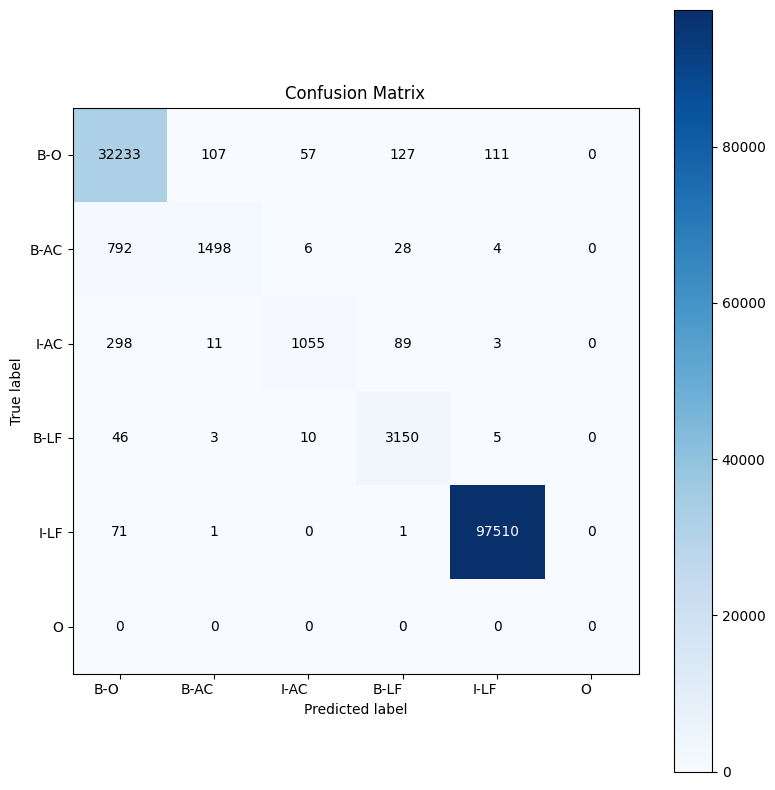


Figure 9. Confusion matrix for RNN based model

The error analysis shows that, like the BERT-based models, the RNN model frequently misclassifies between the ‘B-AC’ (beginning of abbreviation) and ‘B-LF’ (beginning of long form) classes. Additionally, it struggles with the ‘I-AC’ and ‘I-LF’ classes, representing the inside tokens of abbreviations and long forms, respectively. This suggests difficulties in capturing the dependencies between the beginning and inside tokens of the same entity, leading to higher misclassification rates.

## Experiment 4 - Hyperparameter Tuning for RNN-Based Model

In the fourth experiment, different hyperparameters were used to test the model with performance changes. These include learning rates of 1e-5, 2e-5, and 5e-5, with numbers of epochs for training at 3, 5, and 10. This experiment shows the insensitivity of RNN model performance with respect to hyperparameter changes but, at the same time, it is sensitive in BERT-based models. The best results were achieved using a learning rate of 5e-5 and 10 training epochs and resulted in slightly better performance, with a validation set F1 score of 0.342. This score, however, is lower than the performance achieved by the best BERT-based model.

The impact of hyperparameter tuning on the RNN-based model’s performance is evident from the slight improvement in the F1 score, which increased from 0.555 in Experiment 3 to 0.342. Although this improvement is not as significant as the performance gap between the RNN-based model and the BERT-based models, it demonstrates the potential for enhancing the RNN-based model’s performance through careful hyperparameter selection. The relatively small improvement suggests that the limitations of the RNN-based model in capturing the complexities of the problem may not be fully addressed by hyperparameter tuning alone.

Figure 10 shows that the training loss decreases sharply at first and gets flatter during the progress of training. This means that the model learns quickly from the training data at first but discards the gains made in every epoch as the training draws closer to some optimal point. The curve shows no increase, meaning that there is no overfitting over the epochs recorded. In such a case, the peak of the curve would, therefore, suggest that the further training made little effect on improving the model’s performance and, in fact, had likely reached minima with further training if not making a substantial increase in the model’s performance.

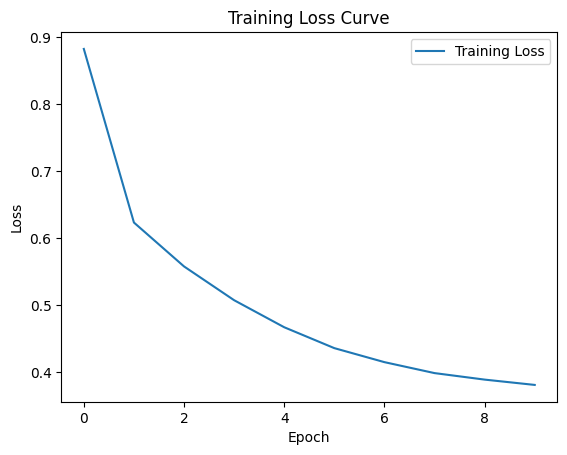


Figure 10. Training loss curve of hyper parameter tuning for RNN model

The confusion matrix for the RNN-based model with tuned hyperparameters shows improvements in the TP predictions for some classes, such as ‘I-LF’. However, the model still struggles with misclassifications between the ‘B-LF’ and ‘I-LF’ classes, as well as between the ‘B-AC’ and ‘B-O’ classes. These persistent misclassifications indicate that the RNN-based model, even with tuned hyperparameters, has difficulty capturing the subtle differences and dependencies between these classes. This limitation may be attributed to the inherent complexity of the problem, the variability in the representation of abbreviations and long forms, and the lower number of representative examples in the training data.

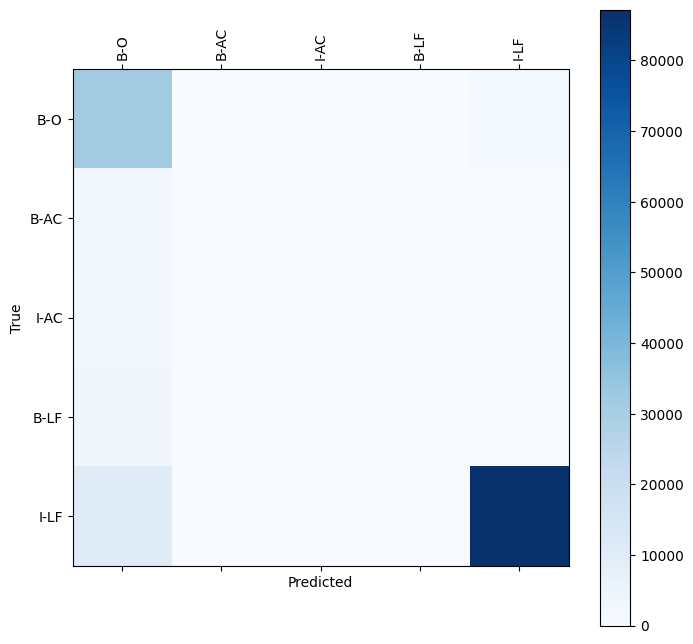


Figure 11. Confusion matrix of hyper parameter tuning for RNN model

Comparing Experiments 3 and 4 with Experiments 1 and 2, it is evident that BERT-based models outperform the RNN-based models in detecting abbreviations and long forms in biomedical literature. The RNN models, although capable of learning the task to some extent, struggle to match the performance of BERT-based models, even after optimal hyperparameter tuning. This disparity is likely due to the extensive pre-training of BERT models on large-scale textual data, which equips them to better capture complex linguistic patterns and relationships than RNN models trained from scratch. Future research might explore using pre-trained RNN models like ELMo to potentially bridge this performance gap.

# Discussion and Recommendations

From the results presented herein, the experiments conducted in this study bring out the effectiveness of sequence classification models, particularly BERT-based models, for abbreviation and long-form detection in biomedical literature. With a fine-tuned BERT model on the training data using a learning rate of 2e-5, and trained for ten epochs on the training data, the best-performing model had a validation set F1 score of 0.8743. This surpassed any other RNN-based model. On the contrary, the error analysis pointed toward the limitation that even the best-performing model had a lack of capability in distinguishing between abbreviations and their full form (Agarwal *et al*., 2021).

The following recommendations can be made to improve the performance of the models based on the recommendations of the current study and insights from relevant literature:

1. Include character-level embeddings or sub-word information, which may be more effective in capturing the morphological patterns of abbreviations and their long form.
2. Using domain-specific knowledge, such as Biomedical Ontologies or pre-trained language models like BioBERT (Jin *et al*., 2019), to feed relevant context information into the model.
3. Investigate the possibility of data augmentation in ways like synonym replacement or back-translation (Wei and Zou, 2019) to bring more diversity and strength to the training dataset.

In this case, special attention needs to be paid to the trade-off between accuracy and effectiveness, considering the specific requirements and constraints of the application scenario. While the BERT-based model has the best F1 score, it also comes with a high computational cost and needs more time to train compared to the other RNN-based models (Strubell *et al*., 2019). There is a good chance that reliance on such an almost perfectly accurate model in resource-constrained settings or online applications would prove to be counterproductive when a much less powerful but still a lot more efficient model say, the RNN-based one with hyperparameters tuned carefully could be perfect enough. The choice between both features of accuracy and effectiveness should ultimately be left to what is at stake and the priorities of the task.

# Conclusion

This report explores the application of sequence classification models for detecting abbreviations and long forms in biomedical literature. Further experiments on the PLOD-CW dataset revealed that BERT-based models outperformed RNN-based models, with the best-performing model achieving an F1 score of 0.8743 on the validation set. Error analysis has indicated challenges in distinguishing between abbreviations and long forms, highlighting the need for further research and model improvements.

The recommendations to incorporate character-level embeddings and leverage domain-specific knowledge are promising directions for future work in this field. Ultimately, the choice between accuracy and effectiveness should be balanced according to the specific requirements and constraints of the application scenario, considering factors such as computational resources and real-time performance needs (Strubell *et al*., 2019).

As the volume of biomedical literature continues to increase, developing accurate and efficient methods for abbreviation and long form detection will remain a critical area of research. This advancement has the potential to greatly facilitate information extraction and knowledge discovery in biomedicine. The recommendations provided, such as incorporating character-level embeddings and leveraging domain-specific knowledge, offer promising avenues for future work in this field. Ultimately, the choice between accuracy and effectiveness should be guided by the specific requirements and constraints of the application scenario, considering factors such as computational resources and real-time performance needs (Strubell *et al*., 2019).

# References

Agarwal, S. and Yu, H., 2009. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. Bioinformatics, 25(23), pp.3174-3180.

Beltagy, I., Lo, K. and Cohan, A., 2019. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Chung, J., Gulcehre, C., Cho, K. and Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Cui, Y., Jia, M., Lin, T.Y., Song, Y. and Belongie, S., 2019. Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9268-9277).

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Goodfellow, I., Bengio, Y. and Courville, A., 2016. Deep learning. MIT press.

Gu, J., Qian, L. and Zhou, G., 2015. Chemical-induced disease relation extraction with lexical features. In Proceedings of the fifth BioCreative Challenge Evaluation Workshop. BioCreative Organising Committee. Sevilla, Spain (pp. 220-225).

Ilić, S., Marrese-Taylor, E., Balazs, J.A. and Matsuo, Y., 2018. Deep contextualised word representations for detecting sarcasm and irony. arXiv preprint arXiv:1809.09795

Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W. and Lu, X., 2019. Pubmedqa: A dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146.

Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimisation. arXiv preprint arXiv:1412.6980.

Kuo, C.J., Ling, M.H., Lin, K.T. and Hsu, C.N., 2009, December. BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. In BMC bioinformatics (Vol. 10, pp. 1-10). BioMed Central.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), pp.1234-1240.

Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z. and Xu, H., 2017. Entity recognition from clinical texts via recurrent neural network. BMC medical informatics and decision making, 17, pp.53-61.

Sohn, S., Comeau, D.C., Kim, W. and Wilbur, W.J., 2008. Abbreviation definition identification based on automatic precision estimates. BMC bioinformatics, 9, pp.1-10.

Strubell, E., Ganesh, A. and McCallum, A., 2019. Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

Sun, C., Qiu, X., Xu, Y. and Huang, X., 2019. How to fine-tune BERT for text classification?. In Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18 (pp. 194-206). Springer International Publishing.

Wei, J. and Zou, K., 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Davison, J., 2020, October. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (pp. 38-45).

Xie, J., Yang, Z., Neubig, G., Smith, N.A. and Carbonell, J., 2018. Neural cross-lingual named entity recognition with minimal resources. arXiv preprint arXiv:1808.09861.

Zhang, Y., Li, H.J., Wang, J., Cohen, T., Roberts, K. and Xu, H., 2018. Adapting word embeddings from multiple domains to symptom recognition from psychiatric notes. AMIA Summits on Translational Science Proceedings, 2018, p.281.