Evaluation of Foundation Models for Biomedical Named Entity Recognition

By

Amin Mohamed Shire

Supervised by Dr Shyamasree Saha

Acknowledgements

I would like to express my deepest appreciation to my supervisor Dr Shyamasree Saha, whose support and feedback were vital in all stages of this project. Also, would like to thank Professor Conrad Bessant and the whole Queen Mary bioinformatics department for their guidance over the last year when I could not have imagined myself taking on all that I did. Special thanks to my classmates who I have grown a lot with over our time on the course, especially to Team Mint, together we won the best software project earlier this year and I am very proud to be a part of that. I can't forget my family who were very patient with me over the last few months and my friends who helped me more than I can express. And lastly, my cat who was there all those late nights when no one else was.

Contents page

Abstract

Biomedical Named Entity Recognition (BioNER) is a critical task in natural language processing, particularly for extracting meaningful information from vast biomedical literature. This study evaluates the performance of several foundation models, including BERT, BioBERT, BioMedBERT, RoBERTa, GPT-2, and Bioformer models 16L and 8L, on the 'Europe PMC annotated full-text corpus for gene/proteins, diseases and organisms', which is a fully human-annotated biomedical dataset comprised of 300 research articles. The models were fine-tuned and tested across various batch sizes and learning rates, with their performance assessed using precision, recall, and F1-score metrics. The results reveal that domain-specific models like BioBERT and BioMedBERT, pre-trained on extensive biomedical corpora, significantly outperform general-purpose models in BioNER tasks. BioBERT achieved the highest F1-score of 95.08% with a batch size of 8 and a learning rate of 1e-5, demonstrating its robustness in capturing the nuanced terminology of biomedical texts. BioMedBERT also performed exceptionally well, particularly with larger batch sizes, confirming the advantages of specialized pre-training for domain-specific tasks. In contrast, while still effective, general-purpose models like BERT and RoBERTa did not reach the same level of performance, underscoring the importance of domain-specific training. GPT-2, designed primarily for text generation, exhibited the lowest performance across the board, indicating its limitations in NER tasks without significant adaptation. The Bioformer models, although slightly trailing in F1 scores, proved to be the most computationally efficient, offering a balanced trade-off between performance and resource usage. These findings highlight the necessity of careful model selection and fine-tuning in BioNER tasks, particularly when dealing with domain-specific content. This study contributes valuable insights into the optimal use of these advanced models in the biomedical field, with implications for future research and practical applications in biomedical informatics.

Introduction

Named Entity Recognition

Named entity recognition (NER) is a critical natural language processing (NLP) method designed to extract and classify proper nouns from text into predefined categories such as person, location, organisation, time, date, and various numerical expressions. In the last three decades, NER has come a long way, from the first rule-based approaches to the current deep learning methods, there are countless models and methods to tackle NER tasks, so it is vital that they are evaluated to find the most efficient and effective models, this is even more important in specialised domains such as biomedical NER.

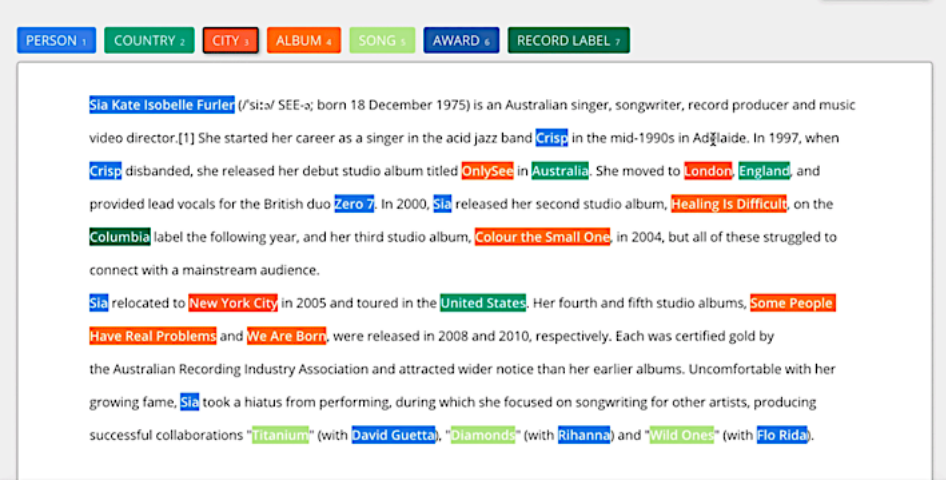


Figure 1: *An extract annotated by NER [1]*

History of NER

The term "named entity recognition" was first coined in the mid-1990s during the Sixth Message Understanding Conference (MUC-6) [2], where it was identified as a crucial task for extracting meaningful information from text. NER gained substantial traction in the late 1990s and early 2000s as researchers began employing machine learning and statistical methods to enhance the performance of NER systems [3]. By 2003, significant advancements were made with the introduction of Conditional Random Fields (CRFs) [4], a statistical modelling method that became a cornerstone for NER. CRFs allowed for the prediction of labels by considering the features of each word and their context, thus improving the accuracy of entity recognition.

One of the pivotal moments in the evolution of NER was the CoNLL-2003 shared task, which provided a standardised dataset for NER research [5]. This dataset became a benchmark for comparing models and introduced standard evaluation metrics such as precision, recall, and F1 score, which are still widely used in NER evaluation today. The availability of this dataset enabled researchers to make consistent and significant strides in improving NER systems.

Deep Learning emerges

The 2010s marked a transformative period for NER with the advent of deep learning and neural networks [6]. The increase in computational power and the development of sophisticated hardware like GPUs and TPUs facilitated the training of more complex and deeper neural networks. Neural networks (NNs) are inspired by the structure and functioning of the human brain, comprising interconnected neurons organised into layers, including input, hidden, and output layers [7].

Neural networks are structured with an input layer that receives the initial data, hidden layers where computations are performed to detect patterns and features in the data, and an output layer that produces the final prediction or classification. Each neuron in a neural network receives inputs, processes them, and passes the output to the next layer. These connections between neurons have weights that are adjusted during the training process to minimise prediction errors [8]. Activation functions such as ReLU (Rectified Linear Unit) [9], sigmoid, and tanh introduce non-linearity into the network, enabling it to learn complex patterns [10].

Deep learning builds upon the foundational concepts of neural networks by incorporating multiple hidden layers, often referred to as deep neural networks (DNNs). This depth allows deep learning models to capture and represent hierarchical features and complex patterns in data. Deep learning models excel with large amounts of data and significant computational resources, which is essential for tasks such as image and speech recognition, where high-dimensional data is common [11].

Several innovative architectures have been developed to address specific challenges in deep learning. Convolutional Neural Networks (CNNs), primarily used for image data, utilise convolutional layers to automatically and adaptively learn spatial hierarchies of features [12]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are

effective for sequential data like text and time series [13]. RNNs have connections that form cycles in the network, allowing information to persist, while LSTMs address issues like the vanishing gradient problem, enabling the learning of long-term dependencies [14].

Transformers

A significant breakthrough in the realm of deep learning was the development of the Transformer architecture, introduced in the seminal paper "Attention is All You Need" by Vaswani et al. [15]. Transformers revolutionised the processing of sequential data by employing self-attention mechanisms, which allow the model to weigh the importance of different words in a sentence regardless of their position. This architecture has set a new precedent for NLP and AI research, leading to the development of highly advanced models such as BERT (Bidirectional Encoder Representations from Transformers) [16], GPT (Generative Pre-trained Transformer) [17], and their successors. These models have achieved state-of-the-art results in various NLP tasks, including NER, due to their ability to understand and generate human-like text based on contextual understanding.

The introduction of deep learning yielded more accurate results but at a greater computational cost. These models require large amounts of training data and substantial computational power to train, often taking advantage of the parallel processing capabilities of GPUs and specialised hardware like TPUs [10]. Despite the computational challenges, the performance gains in tasks such as NER, where capturing intricate patterns and dependencies in text is crucial, have been significant.

BioNER

In the biomedical field, NER is indispensable for extracting valuable information from vast amounts of textual data, including scientific literature, clinical records, and patents. Biomedical NER (BioNER) helps in identifying entities such as gene names, protein names, diseases, drugs, and other biological terms. This extraction is vital for tasks such as drug discovery, clinical decision support, and literature-based knowledge discovery, making NER a cornerstone in biomedical informatics [18].

BioNER presents unique challenges due to the complexity and variability of medical terminology, the frequent use of abbreviations and acronyms, and the rapid evolution of

biomedical knowledge. Unlike general domain NER, biomedical NER must handle highly specialised vocabulary and context-specific meanings, which complicates the entity recognition process [19].

Foundation Models

Foundation models like BERT, GPT, and their variants have revolutionised NLP by providing pre-trained language models that can be fine-tuned for specific applications. These models leverage vast amounts of text data to learn rich contextual representations, enabling significant improvements in various NLP tasks, including NER. The ability to fine-tune these pre-trained models for domain-specific tasks has led to state-of-the-art performance in biomedical NER [20]

BERT

BERT and models based on its architecture have excelled in BioNER tasks in recent years due to their ability to leverage bidirectional context [21]. Unlike traditional models that process text either left-to-right or right-to-left, BERT reads entire sequences of words simultaneously in both directions, meaning it considers the context of each word by looking at both its preceding and following words. This bidirectional approach is crucial for understanding the full context of each word, which is particularly important in classification tasks like NER, where the meaning of a word often depends on its surrounding context.

BERT achieves this through two pre-training objectives. The first is Masked Language Modelling (MLM), where a certain percentage of words in the input text are randomly masked, and the model is trained to predict the masked words based on the context provided by the unmasked words on both sides. This encourages the model to learn strong contextual representations.

The second objective is Next Sentence Prediction (NSP). In this task, the model is given pairs of sentences and trained to predict whether the second sentence logically follows the first in the original text or if the pair is randomly chosen. NSP helps the model learn about the relationships between sentences, which is especially useful in tasks that require understanding the order, flow, and coherence of text.

This combination of MLM and NSP allows BERT and similar models to understand complex language patterns and relationships, making them highly effective for NER and other natural language processing tasks.

BERT-based models

Some BERT variations include RoBERTa (A Robustly Optimized BERT Pretraining Approach), which is a reimplementation of BERT that optimizes the pre-training process by removing the Next Sentence Prediction (NSP) objective and focusing solely on Masked Language Modelling (MLM). Developed by Facebook AI in 2019, RoBERTa is pre-trained on a significantly larger corpus compared to BERT, allowing it to capture a broader range of linguistic patterns. This optimization has led RoBERTa to achieve state-of-the-art performance on various NLP benchmarks, making it a powerful tool in the field of natural language processing [22]

Domain-specific BERT models, including BioBERT and BioMedBERT, previously called PubMedBERT, have been pre-trained on extensive biomedical corpora, allowing them to better understand and process the complex and unique terminology found in the biomedical field. Unlike general-purpose models, which are trained on a wide range of text data, these specialised models are tailored to the specific needs of biomedical NLP tasks, such as BioNER. This specialised training enables them to achieve superior performance in identifying and categorising biomedical entities, such as gene names, diseases, and chemical compounds, where general-purpose models often struggle [23].

BioBERT's biomedical training is built upon BERT's general pre-training however, BioMedBERT is not as it is trained only on biomedical data from the start making it highly domain-specific. By capturing the nuances of domain-specific language, BioBERT and BioMedBERT offer a significant advantage in applications requiring a deep understanding of biomedical texts.

Bioformer 8L and 16L are lightweight BERT-based models trained from scratch on biomedical data much like BioMedBERT. 8L and 16L refer to the number of layers each model has. BERT has 110M parameters all contributing to its ability to capture complex patterns in text, Bioformer 8L and 16L are more compact with only 43M parameters which is 40% less than BERT. Although these models are smaller, previous studies have shown they

outperform BERT in biomedical NLP tasks such as BioNER while only being slightly less accurate than bigger biomedical-trained models like BioMedBERT [24].

GPT

GPT's ability to generate coherent and contextually relevant text by predicting the next word in a sequence, given the preceding context, has set it apart from other models. This autoregressive approach allows GPT to generate human-like text that is fluent and contextually appropriate, making it suitable for a wide range of applications, from text completion and translation to more complex tasks like summarisation, question answering, and creative writing. While GPT excels in general-purpose NLP tasks, its architecture can be adapted for more specific applications, including named entity recognition (NER) and other domain-specific tasks [25]. However, unlike models like BERT, GPT was initially designed with a focus on text generation rather than token classification, making it less naturally suited for tasks like NER without further modifications or fine-tuning.

Fine-tuning

Each model is pre-trained on a wide range of data, in the case of BioMedBERT this is strictly biomedical data, and so when preparing the model for a task fine-tuning is required. Fine-tuning a model is the stage in which they are trained on a smaller more task-specific dataset, in this study all models will be fine-tuned on a human-annotated biomedical dataset tailored for NER.

During fine-tuning each model will be trained under pre-defined parameters, hyperparameters, such as batch size and learning rate. Batch size refers to the number of sentences processed in each training iteration. For example, a batch size of 32 means that in each iteration, the model is updated based on 32 sentences rather than the entire dataset at once. Smaller batch sizes allow for more frequent updates to the model’s weights, potentially leading to more accurate learning. However, this comes at the cost of increased training time and computational resources. The learning rate controls the magnitude of the weight updates during training. A higher learning rate results in larger updates, which can speed up the training process but also increases the risk of overshooting the optimal weights, leading to instability and inaccuracies in the model. Conversely, a lower learning rate results in smaller,

more precise updates but requires more iterations and time to converge on the optimal solution.

In summary, the evolution of NER from its inception at the MUC-6 conference to the present day has been characterised by significant technological advancements. From the early use of statistical methods and CRFs to the current state-of-the-art deep learning models and Transformer architectures, NER has continually improved in accuracy and efficiency, solidifying its role as a fundamental component of modern NLP applications.

Aims

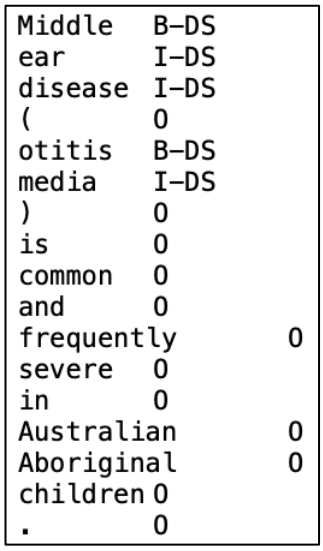
The primary aim of this study is to evaluate the effectiveness of foundation models for biomedical NER. By systematically comparing different models and fine-tuning strategies across various metrics, this research seeks to provide insights into the optimal use of these advanced models in the biomedical domain. Under consistent experimental conditions using standardised datasets to ensure fair comparison, the findings will contribute to the development of more accurate and efficient NER systems, facilitating advancements in biomedical research and applications.

Methodology

Data Collection

Preparing the dataset was a crucial first step in the methodology, the dataset was derived from the Europe PMC annotated full-text corpus, an extensive dataset used for biomedical text-mining [26]. Annotated by experts, the annotations are categorised into three groups: genes/proteins (GP), diseases (DS), and organisms (OG). The dataset was comprised of 300 full-text research articles, with 72,000 mentions of biomedical terms and 114,000 sentences, covering a variety of literature making it great for training and evaluating foundation models on biomedical named entity recognition.

The data is in a tab-separated format where each line has two columns, one for the individual tokens, from the text and the other containing the IOB tags for the associated token. Tokens are the most basic unit of a text that the model processes, they can be words, sub-words or even characters found in other languages. These tags are crucial for identifying the start and continuation of named entities within the text. Each token is labelled as either I, B, or O. B-beginning is the tag given to the first token of an entity, I-inside the tag given to a token within a multi-token entity and O-outside tag is given to anything that is not a part of an entity. An example is Figure 2 below, the tokens “Middle” and “Otitis” are tagged as B as they are the beginning of an entity, while “ear”, “disease”, and “media” are all I for inside as they are part of a multi-token entity. Everything else is tagged as O as they are not an entity. The entities are then further annotated with what type they are DS, GP or OG



```
Middle   B-DS
ear      I-DS
disease  I-DS
(        O
otitis   B-DS
media    I-DS
)        O
is       O
common   O
and      O
frequently       O
severe   O
in       O
Australian       O
Aboriginal       O
children O
.        O
```

***Figure 2**: An extract from the EPMC annotated full-text corpus [20]*

Data Pre-processing and Tokenization

To prepare the data it was loaded into a pandas DataFrame. The two columns were named: 'Word' and 'Tag'. This step ensured that the data was structured correctly for further processing. To form complete sentences and their corresponding tags, the tokens were merged. This was done by splitting the text based on periods and grouping words into sentences. Each word was then associated with its corresponding sentence, enabling the creation of sentence-level annotations necessary for training the model.

Tokenization was performed using the tokenizer from the chosen pre-trained model, ensuring that the tokens and their corresponding labels were aligned correctly. During tokenization, each sentence was split into tokens, and the corresponding labels were mapped to these tokens. Special care was taken to handle cases where labels might be missing or incorrect, especially in BERT models where sub-tokens are common, this ensures that every token has an appropriate label.

To facilitate the training process, a custom dataset class was developed to handle the tokenized sentences and their corresponding labels. This class was responsible for converting the tokens and labels into PyTorch tensors, which are required for model training. The class also managed the batching of data, an essential process for efficient training, particularly when working with large datasets. DataLoader instances were then created to streamline the loading of data during training and evaluation. These DataLoaders handle shuffling and

batching, ensuring that the model is exposed to diverse examples in each epoch, which helps in reducing overfitting and improving generalisation.

Model Selection and Fine-Tuning

When selecting the foundation models to train and evaluate, their track record in NER was taken into consideration. Each model was loaded using the Hugging Face Transformers library, which provides a wide range of pre-trained models for various NLP tasks [27]. The models were then fine-tuned with a sample of the dataset, 80%, for the NER task A custom trainer class was used to handle the training process, focusing on computing the loss for token classification. The training process involved adjusting the model's parameters to minimise the loss, thereby improving its performance on the NER task.

Hyperparameter tuning was performed to find the optimal batch size and learning rate for training the model. Different combinations of batch sizes and learning rates were tried, and the model's performance was evaluated on the validation set for each combination. This process helped identify the best hyperparameters that yielded the highest evaluation metrics.

The training was conducted on the Andrena GPU cluster in Queen Mary's HPC facility, each model was trained on 1 GPU with 12 cores and each with 7.5GB of RAM. This was much faster than previous CPU attempts.

Evaluation metrics

To evaluate the models, they are tested on their ability to correctly label the tokens in the remaining 20% of the dataset, this ensures that the evaluation metrics accurately reflect the ability of the model to label unseen data.

The evaluation metrics were computed using the seqeval library, which is well-suited for sequence labelling tasks like NER. The metrics included precision, recall, and F1-score. These metrics were calculated based on the model's predictions and the true labels from the validation set. Precision measures the proportion of correctly identified entities among all identified entities, recall measures the proportion of correctly identified entities among all

true entities, and F1-score is the harmonic mean of precision and recall, providing a single metric to evaluate the model's performance.

While accuracy is a common metric in other machine learning tasks it was disregarded in this study due to the nature of NER. Most tokens in the dataset are non-entities, labelled ‘O’, and so a model that identifies no named entities and labels everything as ‘O’ can achieve a very high accuracy of ~95%. This is why more emphasis is put on precision, recall and F1 in evaluating the ability of models to perform biomedical NER.

Typically, in BioNER on a standard dataset such as BC5CDR and NCBI, 80% > is considered a good F1 score and anything above 90% is excellent. With some datasets having models that achieved F1 scores of 95% and above [28]. However, what is considered a good F1 score on one dataset might not be the same on another, so it is important to consider the results of previous research with the same dataset, but with the dataset used in this study, there are currently no other studies to compare with.

Challenges and Considerations

Several challenges were encountered during the study, including handling the complexity and variability of biomedical terminology, managing the imbalance in entity distribution, and addressing the computational demands of training deep learning models.

For the GPT models selected, the later models of GPT-3/3.5 Turbo and GPT-4/4.o could not be used as they cannot be fine-tuned for BioNER due to their API-based access where you can only prompt them. This means to test the capabilities of GPT, older models which are more accessible such as GPT-2 are the only option. Even with GPT-2, the architecture of the model is not made for token classification as such requires more modification than the BERT models. One such modification is that the GPT models do not have a token classification head so by adding the ‘AutoModelForTokenClassification’ class to the code this appends a head to the model for NER.

In conclusion, the methodology employed in this study was designed to systematically prepare, train, and evaluate a foundation model for biomedical NER. By leveraging pre-trained models and fine-tuning them on domain-specific data, this research aimed to contribute to the development of more accurate and efficient NER systems in the biomedical field.

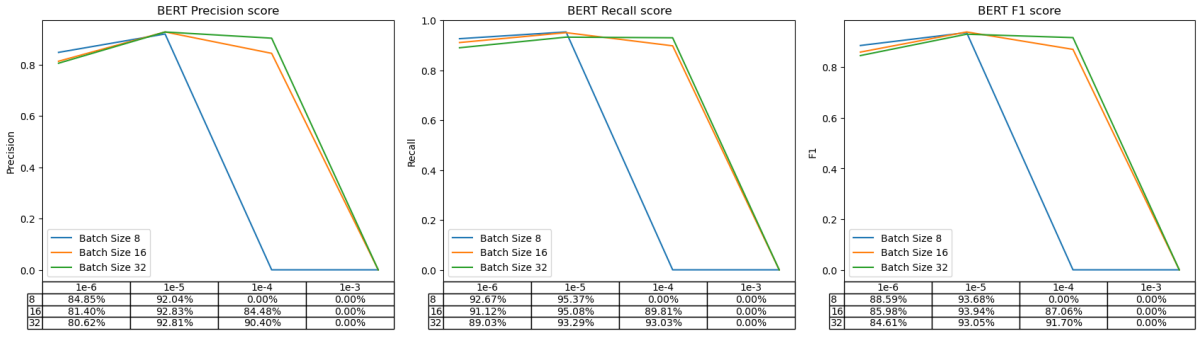
Results and Discussion

BERT Model Performance

In Figs 3a, 3b, and 3c BERT demonstrated strong performance in the BioNER task. The optimal F1-score of 93.94% was achieved when the model was fine-tuned with a batch size of 16 and a learning rate of 1e-5, Fig 3c. This configuration provided a balanced approach, yielding high precision and recall scores, which are critical in ensuring the accurate identification of biomedical entities. Interestingly, the highest recall score of 95.37% was observed with a slightly smaller batch size of 8, also at the same learning rate of 1e-5, Fig 3b. This suggests that while BERT is highly effective in recognising a wide range of entities, slight adjustments in batch size can fine-tune its ability to recall entities with high precision.

However, it is crucial to note that the model exhibited a significant decline in performance at higher learning rates across all batch sizes. Notably, at learning rates of 1e-4 and above, the F1, recall, and precision scores dropped to zero when the model was trained with a batch size of 8. This sharp decline indicates that the model likely overfits the training data rapidly, resulting in poor generalisation of the validation set. The observed overfitting at higher learning rates underscores the importance of selecting an appropriate learning rate, particularly in the context of domain-specific tasks such as BioNER.

The findings suggest that for the BERT model, a moderate learning rate of 1e-5 coupled with smaller batch sizes is optimal for achieving high performance in BioNER tasks.



	1e-6	1e-5	1e-4	1e-3
8	84.85%	92.04%	0.00%	0.00%
16	81.40%	92.83%	84.48%	0.00%
32	80.62%	92.81%	90.40%	0.00%

	1e-6	1e-5	1e-4	1e-3
8	92.67%	95.37%	0.00%	0.00%
16	91.12%	95.08%	89.81%	0.00%
32	89.03%	93.29%	93.03%	0.00%

	1e-6	1e-5	1e-4	1e-3
8	88.59%	93.68%	0.00%	0.00%
16	85.98%	93.94%	87.06%	0.00%
32	84.61%	93.05%	91.70%	0.00%

Figure 3a: BERT Precision Score *- This figure displays the precision scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for BERT with varying batch sizes (8, 16, 32). The precision score indicates the proportion of correctly identified entities among all entities identified by the model.*

 Figure 3b: BERT Recall Score *- This figure presents the recall scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for BERT with varying batch sizes (8, 16, 32). The recall score reflects the model's ability to correctly identify relevant named entities from the dataset.*

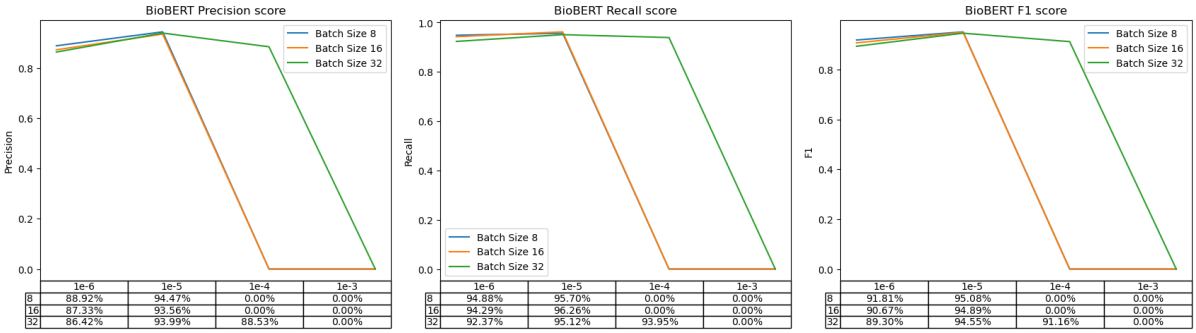
 Figure 3c: BERT F1 Score *- This figure illustrates the F1 scores for BERT across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) with varying batch sizes (8, 16, 32). The F1 score is the harmonic mean of precision and recall, offering a balanced measure of the model's accuracy in the BioNER task.*

BioBERT Model Performance

BioBERT, a variant of BERT that has been pre-trained on large-scale biomedical corpora, exhibited performance trends that were similar to the BERT model but with a notable improvement in the peak F1-score. The highest F1-score of all models at 95.08% was achieved with a batch size of 8 and a learning rate of 1e-5, Fig 4c.

Interestingly, BioBERT showed a similar decline in performance at learning rates of 1e-4 and above, mirroring the behaviour observed with BERT. This drop-off in performance at higher learning rates underscores the model's sensitivity to hyperparameter adjustments, which is consistent with its architecture's reliance on fine-tuning for specific domains.

The results indicate that while BioBERT offers a specialised advantage in biomedical contexts, it shares a common sensitivity with BERT regarding learning rates and batch sizes. Therefore, like BERT, BioBERT necessitates a balanced approach in tuning these parameters to maximise its performance in BioNER tasks. The findings also highlight the importance of leveraging domain-specific pre-training in enhancing the model's capability to recognize and classify biomedical entities effectively.



	1e-6	1e-5	1e-4	1e-3
8	88.92%	94.47%	0.00%	0.00%
16	87.33%	93.56%	0.00%	0.00%
32	86.42%	93.99%	88.53%	0.00%

	1e-6	1e-5	1e-4	1e-3
8	94.88%	95.70%	0.00%	0.00%
16	94.29%	96.26%	0.00%	0.00%
32	92.37%	95.12%	93.95%	0.00%

	1e-6	1e-5	1e-4	1e-3
8	91.81%	95.08%	0.00%	0.00%
16	90.67%	94.89%	0.00%	0.00%
32	89.30%	94.55%	91.16%	0.00%

 ***Figure 4a**: **BioBERT Precision Score** - This figure displays the precision scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for BioBERT with varying batch sizes (8, 16, 32). The precision score indicates the proportion of correctly identified entities among all entities identified by the model.*

 ***Figure 4b**: **BioBERT Recall Score** - This figure presents the recall scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for BioBERT with varying batch sizes (8, 16, 32). The recall score reflects the model's ability to correctly identify relevant named entities from the dataset.*

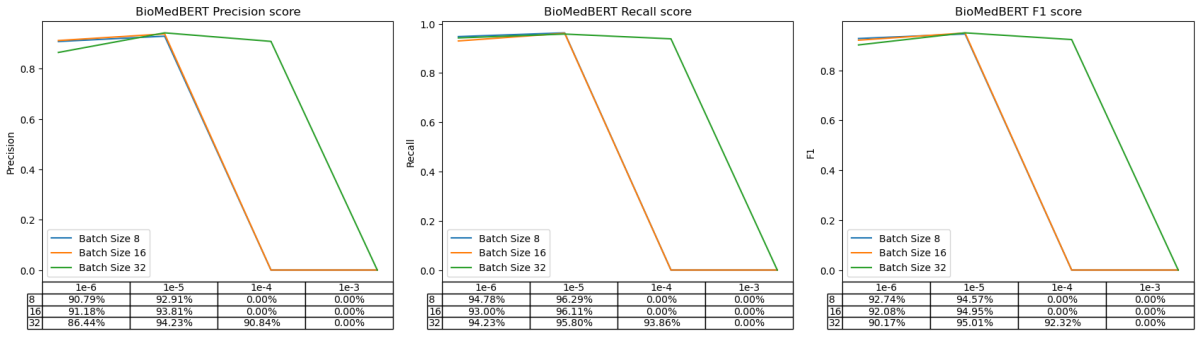
 ***Figure 4c**: **BioBERT F1 Score** - This figure illustrates the F1 scores for BioBERT across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) with varying batch sizes (8, 16, 32). The F1 score is the harmonic mean of precision and recall, offering a balanced measure of the model's accuracy in the BioNER task.*

BioMedBERT Model Performance

The BioMedBERT model, also known as PubMedBERT, also demonstrated excellent performance, particularly when fine-tuned with larger batch sizes and lower learning rates. The model's highest F1-score of 95.01% was achieved with a batch size of 32 and a learning rate of 1e-5, Fig 5c. This configuration also yielded the highest precision score of 94.23%, Fig 5a, suggesting that BioMedBERT is particularly effective in achieving a high level of precision when fine-tuned with larger batch sizes. This might be due to BioMedBERT having not been pre-trained on any general data like BioBERT and BERT, making it more adept at handling the intricacies of biomedical terminology and entity recognition.

However, similar to the other models, BioMedBERT also experienced a sharp decline in performance at higher learning rates (1e-4 and above). At these higher learning rates, the model failed to correctly identify any entities, as reflected in the drop to zero in both recall and precision scores. This sharp decline underscores the need for careful tuning of learning rates, especially in models that are specialised for specific domains like biomedical texts.

The results indicate that for BioMedBERT, a balance between batch size and learning rate is crucial, with the best performance achieved at larger batch sizes and moderate learning rates. The model's strong performance, particularly in precision, highlights its potential as a highly effective tool for BioNER tasks. Its specialisation in biomedical texts likely contributes to its superior performance compared to more general-purpose models like BERT, making it a valuable asset in the field of biomedical informatics.



	1e-6	1e-5	1e-4	1e-3
8	90.79%	92.91%	0.00%	0.00%
16	91.18%	93.81%	0.00%	0.00%
32	86.44%	94.23%	90.84%	0.00%

	1e-6	1e-5	1e-4	1e-3
8	94.78%	96.29%	0.00%	0.00%
16	93.00%	96.11%	0.00%	0.00%
32	94.23%	95.80%	93.86%	0.00%

	1e-6	1e-5	1e-4	1e-3
8	92.74%	94.57%	0.00%	0.00%
16	92.08%	94.95%	0.00%	0.00%
32	90.17%	95.01%	92.32%	0.00%

***Figure 5a**: **BioMedBERT Precision Score** - This figure displays the precision scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for BioMedBERT with varying batch sizes (8, 16, 32). The precision score indicates the proportion of correctly identified entities among all entities identified by the model.*

 ***Figure 5b**: **BioMedBERT Recall Score** - This figure presents the recall scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for BioMedBERT with varying batch sizes (8, 16, 32). The recall score reflects the model's ability to correctly identify relevant named entities from the dataset.*

 ***Figure 5c**: **BioMedBERT F1 Score** - This figure illustrates the F1 scores for BioMedBERT across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) with varying batch sizes (8, 16, 32). The F1 score is the harmonic mean of precision and recall, offering a balanced measure of the model's accuracy in the BioNER task.*

RoBERTa Model Performance

The RoBERTa model exhibited a distinct performance pattern in the BioNER task. The optimal F1-score, recall, and precision were all observed at a batch size of 8 and a learning rate of 1e-5, Figs 6c, 6b and 6a. This indicates that RoBERTa, despite its architectural differences from BERT-based models, such as not using the next sentence prediction objective during pre-training, still benefits from similar fine-tuning strategies when applied to BioNER tasks.

However, RoBERTa's performance dropped sharply to zero at learning rates of 1e-4 and 1e-3, even at batch size 32 where other models, particularly BERT and BioBERT, tend to still maintain some level of performance. This suggests that while RoBERTa is highly effective at lower learning rates, it is particularly susceptible to overfitting at higher learning rates, resulting in a rapid decline in its ability to generalise.

These findings highlight the importance of carefully tuning the learning rate and batch size for RoBERTa, as the model's ability to generalise diminishes rapidly at higher learning rates. Despite this sensitivity, RoBERTa remains a strong contender for BioNER tasks, particularly when fine-tuned with lower learning rates and smaller batch sizes.

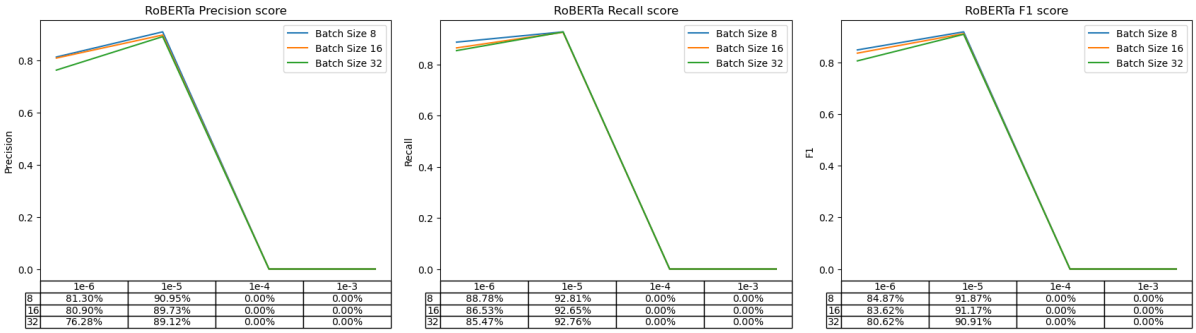


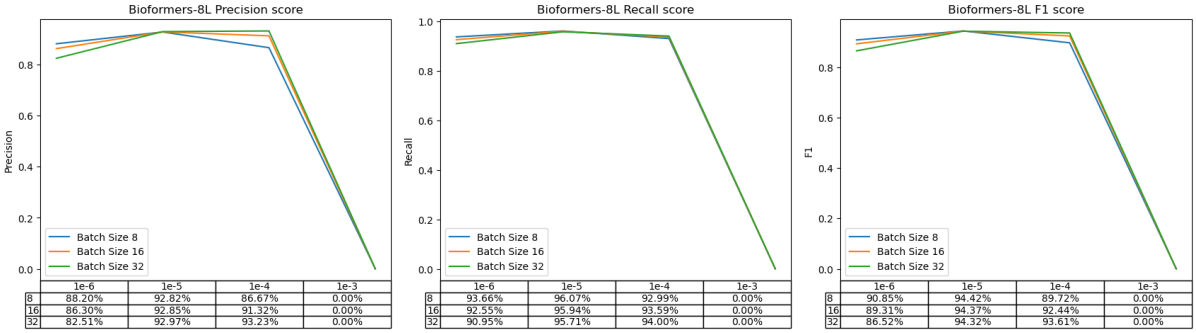
 ***Figure 6a**: **RoBERTa Precision Score** - This figure displays the precision scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for RoBERTa with varying batch sizes (8, 16, 32). The precision score indicates the proportion of correctly identified entities among all entities identified by the model.*

 ***Figure 6b**: **RoBERTa Recall Score** - This figure presents the recall scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for RoBERTa with varying batch sizes (8, 16, 32). The recall score reflects the model's ability to correctly identify relevant named entities from the dataset.*

 ***Figure 6c**: **RoBERTa F1 Score** - This figure illustrates the F1 scores for RoBERTa across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) with varying batch sizes (8, 16, 32). The F1 score is the harmonic mean of precision and recall, offering a balanced measure of the model's accuracy in the BioNER task.*

Bioformer 8L and 16L Performances

The Bioformer 8L and 16L models generated great F1 scores of 94.42% and 94.70% respectively, Figs 7c and 8c. Being pre-trained from scratch on biomedical data explains why the models are comparable to BioBERT and BioMedBERT despite being 60% more compact than the latter. The models are also the only ones able to achieve a good F1 score on all batch sizes under the learning rate 1e-4 especially batch size 8 where no other BERT-based model maintained any performance. However, much like the other BERT-based, they could not correctly identify any labels at the learning rate 1e-3. The two models are almost identical in every parameter, with any differences being marginal.



	1e-6	1e-5	1e-4	1e-3
8	88.20%	92.82%	86.67%	0.00%
16	86.30%	92.85%	91.32%	0.00%
32	82.51%	92.97%	93.23%	0.00%

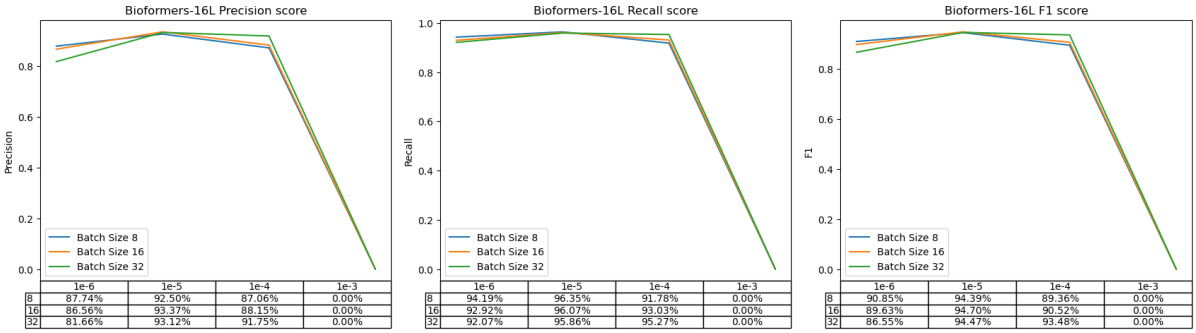
	1e-6	1e-5	1e-4	1e-3
8	93.66%	96.07%	92.99%	0.00%
16	92.55%	95.94%	93.59%	0.00%
32	90.95%	95.71%	94.00%	0.00%

	1e-6	1e-5	1e-4	1e-3
8	90.85%	94.42%	89.72%	0.00%
16	89.31%	94.37%	92.44%	0.00%
32	86.52%	94.32%	93.61%	0.00%

 ***Figure 7a**: **Bioformer-8L Precision Score** - This figure displays the precision scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for Bioformer-8L with varying batch sizes (8, 16, 32). The precision score indicates the proportion of correctly identified entities among all entities identified by the model.*

 ***Figure 7b**: **Bioformer-8L Recall Score** - This figure presents the recall scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for Bioformer-8L with varying batch sizes (8, 16, 32). The recall score reflects the model's ability to correctly identify relevant named entities from the dataset.*

 ***Figure 7c**: **Bioformer-8L F1 Score** - This figure illustrates the F1 scores for Bioformer-8L across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) with varying batch sizes (8, 16, 32). The F1 score is the harmonic mean of precision and recall, offering a balanced measure of the model's accuracy in the BioNER task.*



	1e-6	1e-5	1e-4	1e-3
8	87.74%	92.50%	87.06%	0.00%
16	86.56%	93.37%	88.15%	0.00%
32	81.66%	93.12%	91.75%	0.00%

	1e-6	1e-5	1e-4	1e-3
8	94.19%	96.35%	91.78%	0.00%
16	92.92%	96.07%	93.03%	0.00%
32	92.07%	95.86%	95.27%	0.00%

	1e-6	1e-5	1e-4	1e-3
8	90.85%	94.39%	89.36%	0.00%
16	89.63%	94.70%	90.52%	0.00%
32	86.55%	94.47%	93.48%	0.00%

 ***Figure 8a**: **Bioformer-16L Precision Score** - This figure displays the precision scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for Bioformer-16L with varying batch sizes (8, 16, 32). The precision score indicates the proportion of correctly identified entities among all entities identified by the model.*

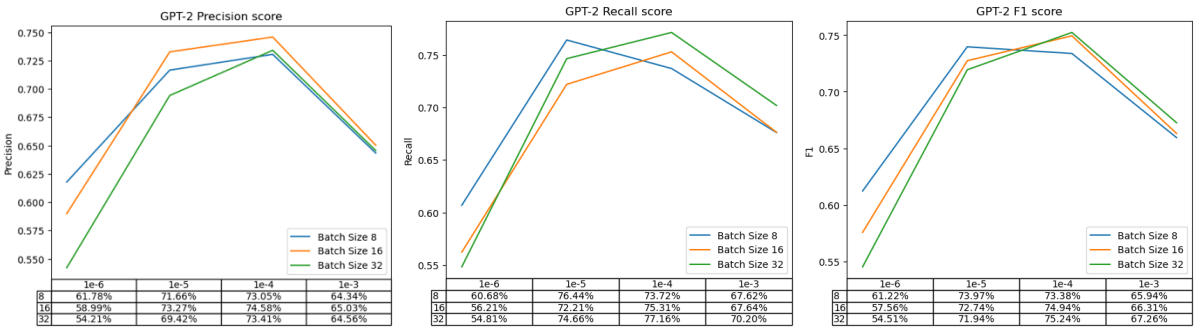
 ***Figure 8b**: **Bioformer-16L Recall Score** - This figure presents the recall scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for Bioformer-16L with varying batch sizes (8, 16, 32). The recall score reflects the model's ability to correctly identify relevant named entities from the dataset.*

 ***Figure 8c**: **Bioformer-16L F1 Score** - This figure illustrates the F1 scores for Bioformer-16L across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) with varying batch sizes (8, 16, 32). The F1 score is the harmonic mean of precision and recall, offering a balanced measure of the model's accuracy in the BioNER task.*

GPT-2 Model Performance

GPT-2, which is traditionally designed for text generation tasks, exhibited radically different behaviour in the BioNER task compared to the BERT-based models. The highest F1-score of 75.24% was obtained at a batch size of 32 and a learning rate of 1e-4, Fig 9c. The precision and recall metrics also peaked around this configuration, Figs 9a and 9b, suggesting that GPT-2, though less commonly used for NER tasks, can still be fine-tuned effectively for such purposes. However, it is important to note that GPT-2's overall performance was lower than that of the BERT-based models.

This lower performance could be attributed to the architectural differences inherent in GPT-2, which is designed primarily for autoregressive text generation rather than token classification tasks. The relatively lower metrics across the board indicate that while GPT-2 can be adapted for BioNER, it does not achieve the same level of performance as models specifically designed or fine-tuned for token classification tasks like BERT, BioBERT, and BioMedBERT. This underscores the importance of selecting the appropriate model architecture for the task at hand, particularly in domain-specific applications like biomedical NER.



	1e-6	1e-5	1e-4	1e-3
8	61.78%	71.66%	73.05%	64.34%
16	58.99%	73.27%	74.58%	65.03%
32	54.21%	69.42%	73.41%	64.56%

	1e-6	1e-5	1e-4	1e-3
8	60.68%	76.44%	73.72%	67.62%
16	56.21%	72.21%	75.31%	67.64%
32	54.81%	74.66%	77.16%	70.20%

	1e-6	1e-5	1e-4	1e-3
8	61.22%	73.97%	73.38%	65.94%
16	57.56%	72.74%	74.94%	66.31%
32	54.51%	71.94%	75.24%	67.26%

 ***Figure 9a**: **GPT-2 Precision Score** - This figure displays the precision scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for GPT-2 with varying batch sizes (8, 16, 32). The precision score indicates the proportion of correctly identified entities among all entities identified by the model.*

 ***Figure 9b**: **GPT-2 Recall Score** - This figure presents the recall scores across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) for GPT-2 with varying batch sizes (8, 16, 32). The recall score reflects the model's ability to correctly identify relevant named entities from the dataset.*

 ***Figure 9c**: **GPT-2 F1 Score** - This figure illustrates the F1 scores for GPT-2 across different learning rates (1e-6, 1e-5, 1e-4, 1e-3) with varying batch sizes (8, 16, 32). The F1 score is the harmonic mean of precision and recall, offering a balanced measure of the model's accuracy in the BioNER task.*

Optimal parameters

Learning rate 1e-5 was the preferred learning rate across all BERT-based models, outperforming the 3 other learning rates in every metric across every batch size. The optimal batch size was much more model-dependent as each model performed differently. Some, such as BioBERT and RoBERTa performed better at the smaller batch size 8, while BioMedBERT showed better results at batch size 32. BERT was in the middle with its best F1 score being under batch size 16, Fig 3c. GPT-2 had different optimal parameters as it showed better results at learning rate 1e-4 for batch sizes 16 and 32 but for batch size 8 the learning rate 1e-5 was where it peaked. This finding highlights the importance of fine-tuning learning rates to match the specific requirements of BioNER, particularly for models that have been pre-trained on domain-specific corpora.

BioBERT vs BioMedBERT

From Fig 10 we can see BioMedBERT and BioBERT both demonstrated exceptional performance, each holding 3 out of the 7 highest F1 scores across all models and parameter combinations tested in this study. This consistency underscores their robustness and adaptability in handling BioNER tasks.

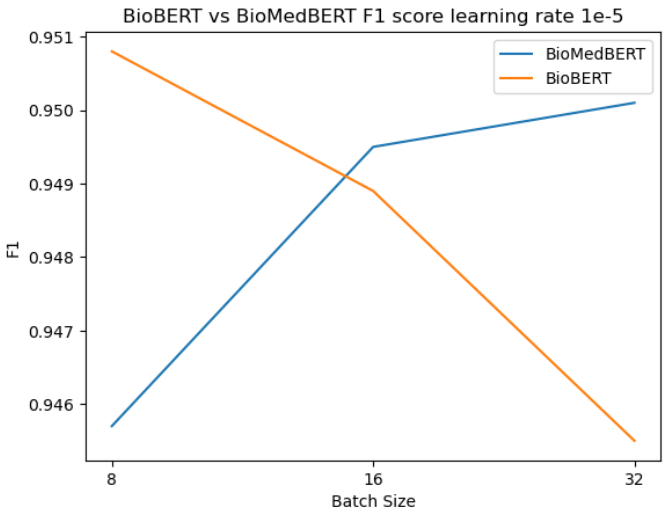


***Figure 10**: Comparison of F1 Scores between BioBERT and BioMedBERT across different batch sizes (8, 16, 32) with a learning rate of 1e-5. BioBERT shows slightly higher F1 scores at batch sizes 8 and 16, while BioMedBERT surpasses BioBERT at batch size 32.*

BioBERT, which is pre-trained on general data and biomedical literature, achieved the highest F1 score of 95.08% at a batch size of 8. This suggests that BioBERT may benefit from smaller batch sizes, which could allow the model to better capture the intricate details, and nuanced patterns present in biomedical texts. Smaller batch sizes often result in more frequent updates to the model weights, potentially leading to better generalisation, particularly in complex and specialised domains like biomedicine.

On the other hand, BioMedBERT, pre-trained strictly on a vast corpus of biomedical text, achieved its peak F1 score of 95.01% at a much larger batch size of 32. This indicates that BioMedBERT may be more capable of handling larger amounts of data in each training iteration, possibly due to its architectural optimisations or the diversity of the biomedical texts it was exposed to during pre-training. The ability to maintain high performance across a range of batch sizes makes BioMedBERT particularly versatile and effective for large-scale biomedical text mining tasks.

In summary, while BioBERT's highest F1 score was slightly higher at 95.08% compared to BioMedBERT's 95.01%, the latter's ability to perform consistently well across larger batch sizes makes it an equally compelling option, especially in scenarios where computational efficiency is a priority. This comparative analysis between BioMedBERT and BioBERT not only highlights their strengths but also provides insights into the nuances of fine-tuning pre-trained models for specialised applications like biomedical named entity recognition. The difference is marginal and will require more studies to conclude a better model.

BERT vs RoBERTa

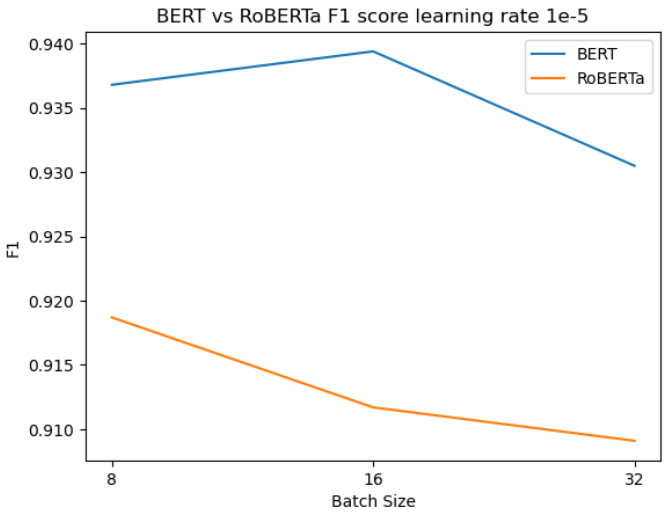


Figure 11: *Comparison of the F1 scores of BERT and RoBERTa across different batch sizes (8, 16, 32) with a learning rate 1e-5*

As the two non-domain-specific BERT models, BERT and RoBERTa's difference in performance was interesting, they had peak F1 scores of 93.94% and 91.87% respectively which is very good given they were not pre-trained on biomedical data. Looking at Figure 11, BERT still outperforms RoBERTa across the board on every parameter, even though RoBERTa is a reimplementation of BERT and is trained on more data. One reason for this could be the tokenization process of each model, BERT uses WordPiece, with each word WordPiece looks for it in the vocabulary and if it cannot find the word it will break the word down into sub-words that are in its vocabulary. On the other hand, RoBERTa uses Byte-Pair encoding, a more aggressive tokenization method, BPE starts with individual characters and then begins merging the most common combinations. BPE tokenization can cause the loss of rarer words which is more common in domain-specific tasks such as BioNER.

Computational and Time cost

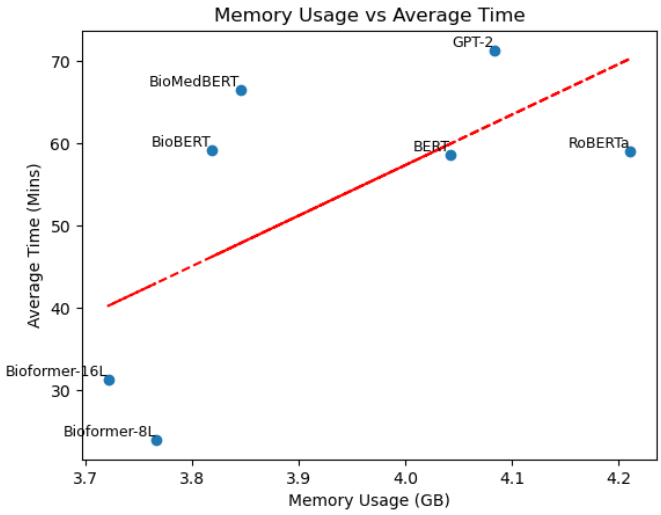


Figure 12: Memory Usage vs. Average Time for Different Models

Scatter plot showing the relationship between memory usage (in GB) and average processing time (in minutes) for various models. The average time was calculated as the total time it took for all parameter combinations to train divided by how many combinations there were (12). The red dashed line represents the line of best fit, indicating the general trend between memory usage and processing time. Models that fall below the line are more efficient, as they use less time for the same amount of memory usage compared to the trend, while those above

the line are less efficient. The models include Bioformer-8L, Bioformer-16L, BioBERT, BioMedBERT, BERT, RoBERTa, and GPT-2.

In Fig 12 we can see that Bioformer-8L and 16L were the most efficient for time and memory usage, which is expected, while BERT, BioBERT and RoBERTa all took roughly the same time to train per parameter combination however, BioBERT used less memory than the others. BioMedBERT and GPT-2 had the longest run times with GPT-2 reaching 70+ minutes per run and having the 2nd most memory usage. BioMedBERT did not use as much memory, but this does not explain why it has such a long run time.

	BERT	BioBERT	BioMedBERT	RoBERTa	GPT-2	Bioformer-8L	Bioformer-16L
Batch Size 8	93.68%	95.08%	94.57%	91.87%	73.97%	94.42%	94.39%
Batch Size 16	93.94%	94.89%	94.95%	91.17%	72.74%	94.37%	94.70%
Batch Size 32	93.05%	94.55%	95.01%	90.91%	71.94%	94.32%	94.47%

Figure 13 F1 Scores for Different Models and Batch Sizes:

This table presents the F1 scores achieved by various models (BERT, BioBERT, BioMedBERT, RoBERTa, GPT-2, Bioformer-8L, and Bioformer-16L) across three batch sizes (8, 16, and 32) in learning rate 1e-5. The F1 score, which is the harmonic mean of precision and recall, is a key metric for evaluating the performance of Named Entity Recognition (NER) models.

To summarise, of the 7 models trained across 12 batch sizes and learning rate combinations, learning rate 1e-5 was by far the best-performing learning rate. BioBERT and BioMedBERT had the highest F1 scores and the only scores to reach higher than 95%, Fig 13, which in other BioNER gold standard datasets would be excellent or even state-of-the-art. Bioformer-8L and Bioformer-16L were the most efficient models while still maintaining excellent F1 scores of 94.42% and 94.70% not that far off of the top 2 models. For the case of Bioformer-8L, the model was 66% faster than BioMedBERT and 60% faster than BioBERT while its F1 score was only 0.59% lower than BioMedBERT and 0.66% lower than BioBERT. The small loss in performance for improved efficiency is a worthwhile exchange on a larger scale. F1 scores of 94%-95% are impressive in BioNER but of course, more studies on the dataset, much like previous more popular datasets, will be required to find out if these are benchmark results in the field [28].

Conclusion

The results of this study highlight the distinct strengths and weaknesses of various foundation models in performing Biomedical Named Entity Recognition (BioNER). The study evaluated BERT-based models (BERT, BioBERT, BioMedBERT, RoBERTa, Bioformer-8L and Bioformer-16L) and GPT-2 on a human-annotated biomedical dataset of 300 full-text articles [26]. Overall, BioBERT and BioMedBERT emerged as the top performers, achieving F1 scores above 95%, which are considered excellent in the context of BioNER. These domain-specific models, pre-trained on extensive biomedical corpora, demonstrated a superior ability to handle the intricate terminology unique to biomedical texts, confirming their efficacy for such tasks.

The findings reveal that while general-purpose models like BERT and RoBERTa can still perform well, they are outperformed by models specifically pre-trained on biomedical data. BERT's performance, though strong, was consistently below that of BioBERT and BioMedBERT, particularly in the F1 scores, underscoring the advantage of domain-specific pre-training. RoBERTa, although a more recent and robust variant of BERT, did not match BERT's performance in this BioNER task, potentially due to its tokenization approach and pre-training strategy.

On the other hand, GPT-2, while being a robust model for text generation, lagged behind in NER performance, reflecting its architectural design that is more suited for generative tasks rather than token classification. The significant drop in performance metrics for GPT-2 across various configurations suggests that it may not be the most suitable choice for NER tasks without substantial fine-tuning or architectural modifications.

Interestingly, Bioformer models (8L and 16L) proved to be the most efficient in terms of computational resources, offering a compelling trade-off between performance and efficiency. Despite their smaller size, these models maintained competitive F1 scores, slightly lower than BioBERT and BioMedBERT but with significantly reduced training times and memory usage.

In conclusion, this study demonstrates that while general-purpose models like BERT and RoBERTa have their merits, domain-specific models such as BioBERT and BioMedBERT are more adept at handling the complexities of biomedical text, making them the better choice for BioNER tasks. The efficiency of the Bioformer models also presents a promising direction for developing more resource-efficient yet effective NER models in the biomedical domain. Future work should explore further fine-tuning strategies and the potential integration of these models in real-world biomedical applications to maximise their utility.

References

1. Text Annotations in the News Industry - DataScienceCentral.com [Internet]. Available from: https://www.datasciencecentral.com/text-annotations-in-the-news-industry/
2. Grishman R, Sundheim B. Message Understanding Conference-6: A Brief History. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996); 1996.
3. Bikei DM, Miller S, Schwartz R, Weischedel R. Nymble: a High-Performance Learning Name-finder. In: Proceedings of the Fifth Conference on Applied Natural Language Processing. Washington, DC: Association for Computational Linguistics; 1997.
4. Lafferty J, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the 18th International Conference on Machine Learning (ICML 2001); 2001.
5. Tjong Kim Sang EF, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003; 2003.
6. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural Language Processing (Almost) from Scratch. J Mach Learn Res. 2011;12:2493-2537.
7. Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Process Mag. 2012;29(6):82-97.
8. LeCun Y, Bengio Y, Hinton G. Deep Learning. Nature. 2015;521:436-444.
9. Rumelhart DE, Hinton GE, Williams RJ. Learning Representations by Back-Propagating Errors. Nature. 1986;323:533-536.
10. Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010); 2010.
11. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems (NeurIPS 2012); 2012.

12. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Comput. 1989;1(4):541-551.
13. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9(8):1735-1780.
14. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014); 2014.
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS 2017); 2017.
16. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019); 2019.
17. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. OpenAI; 2018.
18. Wei CH, Peng Y, Leaman R, et al. Overview of the BioCreative V Chemical Disease Relation (CDR) Task. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop; 2015.
19. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform. 2018;87:12-20. doi: 10.1016/j.jbi.2018.09.008
20. Lee J, Yoon W, Kim S, et al. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. Bioinformatics. 2020;36(4):1234-1240.
21. Cariello MC, Lenci A, Mitkov R. A Comparison between Named Entity Recognition Models in the Biomedical Domain. In: Mitkov R, Sosoni V, Giguère JC, Murgolo E, Deysel E, editors. Proceedings of the Translation and Interpreting Technology Online Conference. Held Online: INCOMA Ltd.; 2021
22. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. 2019.

23. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2019;36(4):1234-1240. doi: 10.1093/bioinformatics/btz682.
24. Fang L, Chen Q, Wei CH, Lu Z, Wang K. Bioformer: an efficient transformer language model for biomedical text mining. Bioinformatics. 2022. DOI: 10.1093/bioinformatics/btac575.
25. Wang S, Sun X, Li X, Ouyang R, Wu F, Zhang T, Li J, Wang G. Gpt-ner: Named entity recognition via large language models. arXiv preprint arXiv:2304.10428. 2023
26. Yang X, Saha S, Venkatesan A, Tirunagari S, Vartak V, McEntyre J. Europe PMC annotated full-text corpus for gene/proteins, diseases and organisms. Sci Data. 2023;10(1):722. Published 2023 Oct 19. doi:10.1038/s41597-023-02617-x
27. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*. 2019;abs/1910.03771.
28. Kocaman V, Talby D. Accurate Clinical and Biomedical Named Entity Recognition at Scale. Software Impacts. 2022 Aug 1;13:100373.