Noise in Biomedical NLP: Investigating Performance of Pre-trained Models on Biomedical Named Entity Recognition (BioNER) Task

A study submitted in partial fulfilment
of the requirements for the degree of
MSc Data Science

at

THE UNIVERSITY OF SHEFFIELD

by

Sheng-Jui Chang

Word Count: 9716

August 2024

Acknowledgement

I would like to thank Professor Denis Newman-Griffis for his patience, support, and academic guidance during the process of my dissertation.

Table of Contents

List of Figures

List of Tables

0. Structured Abstract

The development of large language models (LLMs) has rapidly advanced the evaluation and benchmarking of these models in essential Natural Language Processing (NLP) tasks. However, human-introduced noise is ubiquitous and can lead to a decline in model performance. Therefore, the dissertation focuses on the biomedical field and examines how noise may impact model performance in biomedical named entity recognition (BioNER).

Objective: The primary objective is to understand the impact of noise on model performance in biomedical named entity recognition tasks and to evaluate the robustness of both baseline and domain-specific models.

Aims: The research aims to explore the recent advances in large language models (LLMs) and biomedical LLMs, focusing on various deep learning techniques.

Conclusions: When examining the impact of common human-introduced linguistic noise on biomedical NLP tasks, the research concludes from the experiments that subtle noise does not significantly harm or downgrade model performance of BERT, ELMo, and BioBERT on bioNER tasks, except for BioGPT. While BioBERT has the lowest balanced performance among the four models, its BioNER results remain consistent with the original BC5CDR dataset. In contrast, although BioGPT achieves the highest precision, recall, and F1-score, it is sensitive to spacing, grammatical, and typographical errors.

Keywords: *Linguistic noise, data noise, LLMs, biomedical LLMs, NLP, biomedical NLP, biomedical named entity recognition, biomedical text mining, and deep learning*.

1. Introduction

The advancement of large language models (LLMs) has accelerated evaluating and benchmarking of these models in critical Natural Language Processing (NLP) tasks (Kejriwal et al., 2024). NLP combines AI and Linguistics, aiming to enable computers to understand statements and words written in human languages (Khurana et al., 2023). NLP applications have expanded across various tasks and fields, including machine translation, information extraction, summarisation, classification, question answering, and healthcare (Khurana et al. 2023).

However, recent research has underscored that tasks requiring significant human judgement may manifest noticeable levels of noise (Kahneman et al., 2021). In addition to noise introduced by humans, Andrade (2023) elucidates the concept of statistical noise. Overall, noise is perceived as undesirable as it diminishes the performance of NLP models when applied (Ebadi et al., 2024). Apart from noise attributed to biases (Kahneman et al., 2021) and statistical fluctuations (Andrade, 2023), the study will focus on another type of noise in NLP: linguistic noise within non-standard textual content (Sharou et al., 2021).

1.1. Evaluating Model Performance on Biomedical Named Entity Recognition (BioNER) Task

The study focuses on the Named Entity Recognition (NER) task in the biomedical field. Biomedical NER (BioNER) is critically important for structuring unstructured texts in biomedical literature for further analysis and has been widely studied in both general (Lample et al., 2016; Zhao et al., 2021) and biomedical NLP (Sahu et al., 2016; Zhao et al., 2021). The dissertation prioritises and discusses BioNER task as it is sensitive to noise (Tong et al., 2022). To be more specific, this dissertation will examine four types of linguistic noise that occur in biomedical language: spacing errors, homophone errors, grammatical errors, and typographical errors (Lee et al., 2020). Investigating these types of linguistic noise and their impacts on biomedical literature and NLP tasks is important because they represent human-introduced errors that are common in the biomedical field. However, there has been little research on their impact on bioNER model performance. The BioCreative V Chemical-Disease Relation (BC5CDR) (Li et al., 2016) dataset is used to run the experiments and analyse the impact of noise on the NER models.

1.2. Deep Learning Techniques

Deep learning is a subset of machine learning that involves neural networks with many layers to understand complex patterns in data. The deep learning techniques which the dissertation will be using fall under the two categories: baseline models and their variants. The baseline models cover BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), BioBERT (Lee et al., 2020), and BioGPT (Luo et al., 2022). The four models are selected for the biomedical task because they are the most common ones in the field and have been extensively studied. The extensive research conducted on these models provide valuable insights into their performance, strengths, and limitations within the biomedical NLP domain. These findings not only inform the specific application of these models in biomedical tasks, but also contribute to the broader context of large language models (LLMs). By understanding how these models perform in the biomedical NLP field, researchers can refine and adapt LLMs for other complex, domain-specific applications, ultimately advancing the capabilities of LLMs within various disciplines.

1.3. Research Questions

Recent studies, however, have shown concerns regarding the disparity between the performance of NLP systems on benchmark datasets and their capacity to operate effectively in real-world situations (Ribeiro et al., 2020; Ethayarajh et al., 2020). After clearly defining the noise, the research will be looking at and knowing the deep learning techniques which will be used for the NLP tasks. The research project hopes to understand the role of noise and how it affects our baseline models and their variants. The study begins by enumerating the two research questions (RQs).

1. How does noise in the biomedical named entity recognition (BioNER) task affect our estimates of the pre-trained models on their output patterns and performance?

2. How robust are the existing domain-specific pre-trained models compared to the baseline models on BioNER task against noise?

For RQ1, I keep the original datasets and introduce noise into the dataset to create both clean and noisy versions. I then implement bioNER on both the original and noisy datasets. Finally, I evaluate our model and task performance on the clean and noisy test datasets. The question focuses on understanding how introducing noise into the dataset affects the performance of models in a biomedical context. For RQ2, I discuss the specific context of biomedical NLP tasks and evaluate the resilience of both baseline and domain-specific models in handling noise.

1.4. Objectives and Aims

The primary objective of the dissertation is to understand the impact of noise on model performance in biomedical named entity recognition tasks and to evaluate the robustness of both baseline and domain-specific models. This dissertation aims to explore the recent advances in large language models (LLMs) and biomedical LLMs, focusing on various deep learning techniques.

2. Literature Review

2.1. Search Strategy for Identifying Literature

The searches are conducted in July 2024 in Google Scholar, arXiv, PubMed, and National Institutes of Health (NIH). I first define the core concepts and topics in the dissertation: noise in NLP, pre-trained models, and biomedical NLP. The study then associates the core concepts and topics with broader related terms including data noise, LLMs, biomedical LLMs, biomedical text mining, and deep learning. The biomedical literature the dissertation searches from and uses majorly comes from National Institutes of Health (NIH) and PubMed. This is largely due to their publicly available database, search engine for biomedical literature and its use of professional language (Zhao et al., 2021). Google Scholar and arXiv are also used to help search a wider range of literature across disciplines. Literature and papers found on the databases and sites are considered high-quality and professional sources. The key words used include *Linguistic noise, data noise, LLMs, biomedical LLMs, NLP, biomedical NLP, biomedical named entity recognition, biomedical text mining, and deep learning*.

2.2. Differentiating Various Types of Noise

Recent studies have pointed out that tasks involving significant human judgement can have significant levels of noise (Kahneman et al., 2021). According to the book ‘Noise: A Flaw in Human Judgement’ published by Kahneman et al. (2021), they collect examples and illustrate how noise is introduced to real-world human decision-making scenarios. The book covers examples derived from scenarios such as a group of fellow experts in setting the price of insurance premiums and in deciding the lengths of custodial sentences to illustrate the concepts of ‘Judgement, Noise, and Bias’ (Kahneman et al., 2021; Sleeman & Gilhooly, 2021). Through discussing and reporting the various levels of disagreement between experts, Kahneman et al. (2021) then conclude that noise and biases are omnipresent and hence recommend conducting a ‘noise audit’ in any tasks where there is significant dependence on human judgement. Since Machine Learning systems are heavily dependent on annotated instances which are often completed by humans or tools developed by experts, experts see the book as of significant importance to understanding the role and effect of noise in the field of AI and Machine Learning (Sleeman & Gilhooly, 2021).

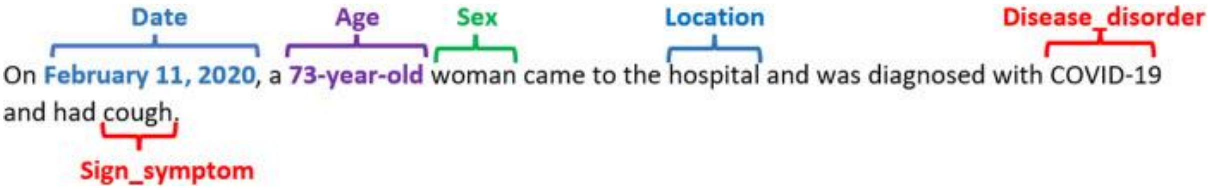
On the other hand, although statistical noise has a long history and often lies within our daily applications (Stuck et al., 1974) and datasets (Fried et al., 2002), Andrade (2023) first discusses the concept of signal and statistical noise in research and argues that signals, the outcome of interest in a study, are distorted by statistical noise. This statistical noise originates from external variables that are measured at different levels and the impact of these variables on subject-to-subject signal variation is quantified by the standard deviation (Andrade, 2023). In other words, there are various types of noise in both our data and real-life situations, and I therefore consider it is important to distinguish these types of noise from the noise which I will be focusing on in the dissertation: harmful linguistic noise which leads to drop in model performance.

2.3. Benchmark Task and Challenges

The vast amount of literature published in PubMed and over the internet are in unstructured format (Missen et al., 2020). With the advancements in technologies and NLP, they provide substantial opportunities to extract valuable insights from vast amounts of unstructured data. In the biomedical domain, for example, identifying entities from electronic health records and clinical trials such unstructured text can be challenging and time-consuming. Now, with such extraction advancements, they help researchers retrieve valuable information from literature and generate reports faster. The benchmark task which I will be focusing on is biomedical Named Entity Recognition (BioNER), a technique to identify biomedical-specific entities from unstructured text and assign them to their corresponding categories. I will then carry out NER on both the annotated BC5CDR dataset and the same dataset with different types of linguistic noise introduced.

2.3.1. Named Entity Recognition (NER) and Biomedical Named Entity Recognition (BioNER)

In 1997, the phrase ‘Named Entity Recognition’ was coined by Chinchor and Robinson (1997) in the Sixth Message Understanding Conference (MUC-6) to streamline information extraction tasks. Nowadays, NER is considered as one of the most common and important BioNLP tasks to accurately identify biomedical-specific entities from unstructured texts (Lample et al., 2016; Zhao et al., 2022; Naseem et al., 2022). It is a technique to identify and separate named entities and group them under predefined categories. These categories can cover, but are not limited to, individual names, locations, organisations, quantities, percentages, and medical notes. Figure 1 shows an example of NER results on a medical record. In contrast, biomedical NER (BioNER) is about identifying biomedical terms of interest from text, such as diseases, drug names, as well as some generic terms including ‘hospital’ and ‘alcohol’. BioNER aims to identify different biomedical entities and predict their entity types. BioNER is a crucial initial step in biomedical literature mining tasks, and its performance significantly impacts subsequent tasks, including biomedical relation extraction (Wei et al., 2016) and drug-drug interaction tasks (Kolchinsky et al., 2015; Liu et al., 2016; Zhao et al., 2021). A widely used tagging method is the IOB format (Settles, 2004), which indicates whether each token is at the Beginning of an entity, Inside, or Outside an entity. The IOB annotation scheme allows for parsing of the entities that share neighbouring word boundaries. Hence, the NER task can be conceptualised as a sequence tagging problem, where each word in the text sequence *X* is assigned to one of three classes. An example of the IOB format can be found in Figure 2, where they represent the token at the Beginning of an entity, Inside, or Outside an entity. In the example, ‘congenital myotonic dystrophy’ is the entity detected and ‘disease’ is its corresponding entity type.



On February 11, 2020, a 73-year-old woman came to the hospital and was diagnosed with COVID-19 and had cough.

Figure 1: NER Results on a Medical Record (Raza et al., 2022)

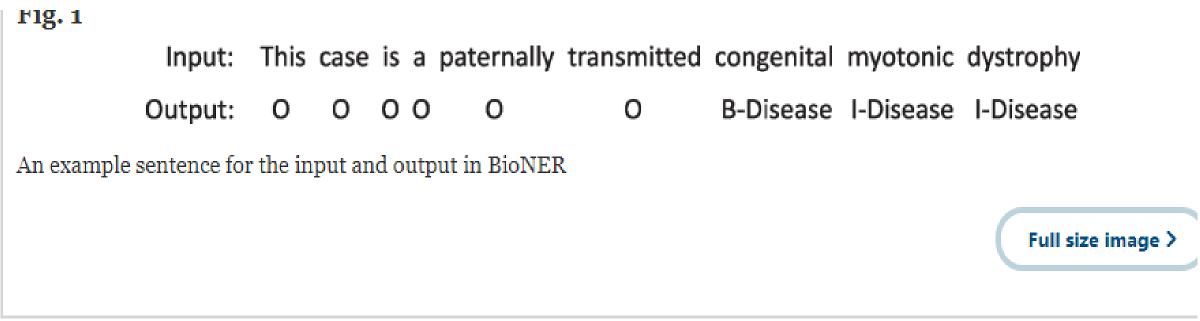


Fig. 1

Input: This case is a paternally transmitted congenital myotonic dystrophy

Output: O O O O O O B-Disease I-Disease I-Disease

An example sentence for the input and output in BioNER

Figure 2: Input and output of BioNER (Zhang & Chen, 2022)

NER Techniques

NER techniques refer to extracting information from unstructured data. They fall mostly into the three categories: 1) rule-based approaches; 2) machine learning approaches; and 3) deep learning approaches. In the past decade, many techniques utilising traditional machine learning approaches have been suggested for biomedical literature mining tasks (Zhao et al., 2021). However, more recently, deep learning models have emerged as a top choice for biomedical text mining and shown satisfactory performance among these tasks (Zhao et al., 2021).

Rule-based approaches involve human experts establishing a set of rules to identify entities in the text based on their grammatical and structural features. They are accurate and are established on language, syntactical, and domain-specific expertise. Examples cover term-matching with existing concept databases such as Unified Medical Language System (UMLS) (Aronson, 2001; Rocktaschel et al., 2012), or pattern matching based on part-of-speech and sentence structure (Eftimov et al., 2017). Since they are not highly generalised, the rules are restricted to a certain domain only (Jehangir et al., 2023). In addition, rule-based NER approaches are often more complex in the domain of biomedical literature due to the characteristics of the field as novel concepts and inventions are published day-by-day (Huang & Lu, 2015; Ramachandran & Arutchelvan, 2021).

Machine learning approaches automatically learn patterns for detecting named entities, which necessitates substantial amounts of labelled training data. Additionally, they generalise better to unseen data. As a result, supervised machine learning methods like Conditional Random Fields (CRFs) and Support Vector Machines (SVMs) have largely replaced traditional rule-based methods (Cho & Lee, 2019). While more advanced machine learning and recent deep learning models require substantial amounts of training data, they can be expensive and difficult to obtain in the biomedical domain (Dernoncourt et al., 2016). In BioNER, annotated data is often restricted to specific types of entity, such as diseases or chemicals. Hence, the approaches can be limited in that they are only capable of identifying these specific entity types. When working with NER tasks using machine learning approaches, the task can be treated as a multi-class classification problem with named entities as out labels. However, named entities require thorough understanding of the sentence context and the sequence of the word labels inside. Hence, the approaches are limited when it comes to identifying entities in long sentences.

Deep learning approaches leverage large amounts of unstructured data and have gained reputation in working with NLP problems. Gao et al. (2021) discovered that with a base NER model such as Long Short Term Memory Conditional Random Field (BiLSTM-CRF) (Huang et al., 2015) or BERT (Devlin et al., 2018), the combination of transfer learning and semi-supervised learning can lead to high model performance without relying on large amounts of labelled data. In the biomedical domain, domain-independent methods based on deep learning and statistical word embeddings, such as bi-directional long short-term memory network (BiLSTM) with CRF and GRAM-CNN have been shown to outperform specific entity recognition tools (Lample et al., 2016; Zhu et al., 2017). On the other hand, with the advancement of Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), a deep bidirectional pre-trained model that uses self-attention mechanisms from the transformer architecture and trained on a dataset with over 2.5 billion words, it has opened more possibilities and opportunities for various NLP tasks such as named entity recognition (NER) and relation extraction (RE) tasks. Transformers and BiLSTMs will be used as the deep learning approaches in the dissertation.

NER Example:

Problem Definition

Given an input sentence $X = \{x1, x2,\ldots,xL\}$, where xi is the ith word (token), and L represents the length of the sentence. A model M must correctly identify the start and end words of each named entity within X and assign it to a corresponding label $y \in Y$, where Y is a predefined list of all possible entity types.

Overall Process

I will use the example from Perera et al. (2020) to illustrate what a NER task involves (see Figure 3). The first step involves the tagging of biomedical entities. For illustration purposes, I borrow the example from Perera et al. (2020). In the sentence 'BRCA1 gene causes predisposition to breast cancer and ovarian cancer.', the tagged entities are 'BRCA1', 'Breast Cancer', and 'Ovarian Cancer'. Next, relations between the entities of interest are inferred with various techniques. In the example, association indicating verbs techniques was used to help infer relations between entities. With the foundation complete from the NER task, the following steps can be used for later relation detection and analysis tasks. Later, the step involves analysing polarity and strengths of relations. In the example sentence, the sentence demonstrates a negative association between the BRCA1 gene and specified disease such as breast cancer and ovarian cancer, the strength of this relationship can be determined either by the shortest path in the sentence dependency tree or by calculating the word distance. Finally, for relation detection and analysis purposes, the extracted relations can be visualised in a graph to identify both direct and indirect relations and interactions.

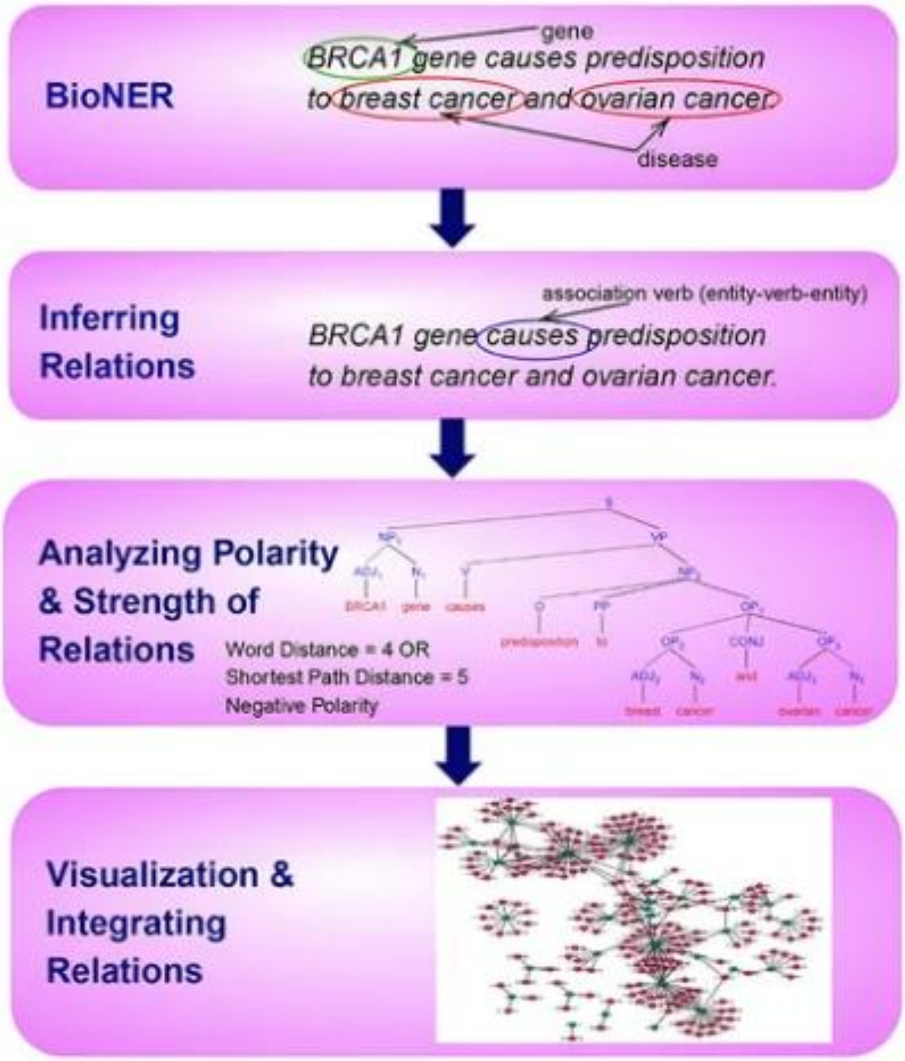

BioNER
gene
BRCA1 gene causes predisposition
to breast cancer and ovarian cancer
disease
Inferring Relations
association verb (entity-verb-entity)
BRCA1 gene causes predisposition
to breast cancer and ovarian cancer.
Analyzing Polarity & Strength of Relations
S
NP
VP
ADJ
N
V
BRCA1
gene
causes
O
PP
OP
predisposition
to
CONJ
and
breast
cancer
ovarian
cancer
Word Distance = 4 OR
Shortest Path Distance = 5
Negative Polarity
Visualization & Integrating Relations

Figure 3: An Overview of the Steps for BioNER and Relation Detection and Analysis (Perera et al., 2020)

Challenges

Common BioNER challenges include data annotations (Jie et al., 2019), complex biomedical texts (Leaman et al., 2015), and handling noisy and misspelt text (Jehangir et al., 2023). Although tackling NER tasks using deep learning techniques can significantly reduce the amounts of labelled data, huge amounts of human-annotated training data are still needed for supervised NER systems. Lengthy and complex sentences are common in biomedical literature. Hence, it is crucial to accurately identify and classify relations inside a lengthy sentence. Leaman et al. (2015) indicated that ambiguous naming convention in biomedical literature is the cause of such ambiguity and complications in biomedical literature. Last, human-introduced linguistic noise such as typos and grammatical mistakes can also be a contributing factor to the complexities and challenges in BioNER (Leaman et al., 2015).

Therefore, focussing on the linguistic errors part, the research decides to investigate how various types of linguistic noise affects model performance.

2.4. Deep Learning Approaches

With the deep learning techniques applied to natural language processing (NLP), major advancements in biomedical text mining models have been indicated (Lee et al., 2019). In the dissertation, bidirectional transformers encoders (BERT, BioBERT), BiLSTM (ELMo), and unidirectional transformer decoder (BioGPT) neural network architectures will be used. The choice of the deep learning techniques is based on the considerable success applied to NLP problems such as NER and RE tasks (Lample et al., 2016; Zhu et al., 2017) and motivated by BERT's (Devlin et al., 2019) state-of-the-art results on various NLP tasks. For instance, given traditional NER systems require a large amount of labelled corpus, Lample et al. (2016) introduced two neural architectures: one based on BiLSTMs combined with CRFs, and another that labels segments using a transition-based approach. These architectures achieved state-of-the-art performance in NER without relying on any language-specific knowledge or external resources. In addition, BERT, which utilises a pre-trained self-attention model based on the Transformer architecture and is trained on over 2.5 billion words, achieves state-of-the-art (SOTA) results in various NLP tasks, including named entity recognition (NER) and relation extraction (RE). Therefore, both transformers and biLSTMs are selected based on their considerable success in working with NER and RE tasks as well as they are effectively contrasted to each other. Also, since both BERT and ELMo are all pre-trained on general knowledge, the study also includes some other domain-specific variants - namely BioBERT and BioGPT - in the experiments and discussions. Given that biomedical text mining tasks require both high accuracy and considerable time investment, the study aims to evaluate the effectiveness of various models in performing biomedical text mining tasks. Given that pre-trained models often achieve satisfactory performance on tasks like NER after fine-tuning on a task-specific dataset, this study aims to explore the role of noise in such tasks and their impact on NER model performance. Since real-life scenarios frequently involve human-induced noise (e.g., typographical and grammatical errors), the research will evaluate how well the models perform under these noisy conditions. Specifically, the research will assess the robustness of BERT, ELMo, BioBERT, and BioGPT to noise and analyse how such noise affects their overall performance.

Transformers

BERT has emerged as a top choice in the NLP community due to its adaptability. BERT (Devlin et al., 2019), which stands for Bidirectional Encoder Representations from Transformers, has emerged as a standout contender. Unlike earlier models which looked at words in left-to-right or right-to-left, using a self-attention that weighs the importance of different words, the transformer-based model is designed to understand the context of words in a sentence by looking at the words that come before and after it. In an architecture of a transformer, each output element is connected to every input element and the weightings between them are dynamically adjusted based on their relationships. Pre-trained on massive datasets, although BERT does not have knowledge in language structure and semantics, its performance correlates with syntactic and semantic patterns. In addition, once pre-trained, BERT can be fine-tuned on another dataset for specific sequence tagging tasks such as Named Entity Recognition (NER) and other text mining downstream tasks. Based on its bidirectional architecture (Devlin et al., 2018), BERT is sensitive to context by considering words and their relations to surrounding words. In addition, at the core of fine-tuning BERT and other transformers lies the principle of transfer learning. Transfer learning, refers to leveraging knowledge from one task to another related task with better performance. This, in turn, reduces the amount of labelled data and facilitates the training process.

Generative Pre-trained Transformers (GPT) are artificial neural networks that are used in NLP tasks (Radford et al., 2018). Alongside BERT, GPT (Radford et al., 2018) also stands out as a major pre-trained language model in the general domain (Luo et al., 2022). Unlike BERT, GPT is a unidirectional architecture and processes text in a left-to-right fashion. The first GPT was introduced by OpenAI in 2018 (Radford et al., 2018) and sets the foundation for more task specific systems such as ChatGPT. Given both machine learning and deep learning require substantial amounts of labelled data for model training, generative pre-training of a language model on unlabeled text followed by discriminative fine-tuning on each individual task enhances transfer efficacy and achieves SOTA results on 9 of 12 benchmarks (Radford et al., 2018). These benchmark tasks, however, do not include NER tasks. The research here will use BioGPT (Luo et al., 2022) - the biomedical variant of GPT to run the experiments and answer the research questions.

On the other hand, since noise is a critical and ubiquitous challenge in biomedical NLP, understanding how SOTA models such as BERT, ELMo, and BioGPT handle noise then becomes essential. Due to the effectiveness of BERT and transformers, they are used in the dissertation as a contrast to another NER top performing architecture: BiLSTM.

BiLSTM

ELMo (Peters et al., 2018) stands for Embeddings from Language Models. It is based on BiLSTMs, a form of recurrent neural network (RNN). BiLSTM is an NLP technique that generates word embeddings in the format of numerical representations of words and captures the context of a word within a sentence. The rationale behind ELMo's use of character-level embeddings is based on the idea that languages often contain morphological clues indicating that a word or sequence of words is a named entity.

2.4.1. BERT

BERT (Devlin et al., 2019) is a contextualised word representation model built upon the architecture of a transformer (Vaswani, 2017). The goal of BERT is to pre-train a deep bidirectional text representation. Unlike other transformer models, BERT is unique in its bidirectional representation of text.

2.4.2. ELMo

Pre-trained word representations (Mikolov et al., 2013; Pennington et al., 2014) are essential components in neural language understanding models. Unlike BERT’s transformer architecture, ELMo (Peters et al., 2018) is a pre-trained bidirectional LSTM (BiLSTM) word embedding approach which covers the whole context and meanings within its embeddings. Unlike traditional static word representations which neglect the context in which they reside, ELMo embeddings consider the comprehensive textual context and dynamically adjust the representation. Like BERT’s bidirectionality, ELMo processes word dependently and enhances language understanding. However, disparities in training objectives between BERT and ELMo can lead to variations in context understanding.

2.4.3. BioBERT

BioBERT (Lee et al., 2020) is a variant of BERT and basically has the same structure as BERT. With the success of pre-training in general NLP tasks (Lee et al., 2019), researchers have been adapting the techniques into the biomedical field. However, applying the general BERT to a specific domain leads to poor performance. Hence, researchers have developed the pre-trained models on biomedical corpora. BioBERT outperforms general BERT on biomedical text generation and mining tasks (Luo et al., 2022).

2.4.4. BioGPT

While BERT and its variants have been extensively studied in the biomedical domain, the possibility of GPT and its variants has not been explored as thoroughly (Luo et al., 2022).

Although Luo et al. (2022) demonstrated that BioGPT has the advantage of generating satisfactory descriptions for biomedical terms and GPT and its variants have achieved significant success in various discriminative biomedical tasks, their lack of generative capabilities limits their application scope (Luo et al., 2022).

3. Methodology and Implementation

3.1. Dataset

BioCreative V Chemical-Disease Relation (BC5CDR) (Li et al., 2016) is composed of 1500 PubMed articles with 4409 annotated chemicals, 5818 diseases and 3116 chemical-disease interactions for biomedical NER and RE tasks (Peng, 2024). The dataset covers annotations of disease and chemical entities, and their interactions (Li et al., 2016; Peng, 2024). The dataset contains the PubMed repository of biomedical literature with a wide range of research articles, reviews, and clinical studies, and is frequently used for biomedical text mining tasks. The study considers biomedical literature has advantages over biomedical texts such as clinical notes based on the three reasons (Zhao et al., 2021): 1) PubMed's publicly available database and search engine; 2) they tend to use professional language and have diverse ways of conveying the same information; 3) there has been proven success of integrating biomedical literature mining tasks into NLP.

3.1.1. Ethical Consideration

The BC5CDR dataset is secondary and anonymised. The PubMed literature in the dataset is already publicly accessible and does not contain any personally identifiable information. The use of publicly available publications ensures that ethical standards are met.

3.2. Methodology

The research focussed on linguistic noise that hinders the accuracy of information and performance of our model on the benchmark tasks. The research proposed to adopt a different approach. Unlike previous studies that apply deep learning models directly to various biomedical NLP tasks and discuss the robustness of those models to noise, I introduced different types of noise to our datasets through the 'spaCy' library and 'random' module and discussed how various noise affects the performance and robustness of each model. The choice of spaCy is due to its Pythonic characteristics and balancing speed and ease of use. With the pre-trained models: BERT, BioBERT, and BioGPT provided by the Hugging Face Transformers library (Wolf et al., 2020) as well as ELMo provided by the Tensorflow Hub, I first ran biomedical named entity recognition (bioNER) on the BC5CDR dataset as the base. Later, I introduced the linguistic noise: homophone error, typographical error, grammatical error, and spacing error into the BC5CDR dataset, focussing on discussing one type of noise at a time, analysed model performance on the noisy dataset, and discussed the role of noise in biomedical NLP. In order to achieve final high BioNER model performance, much time was spent on fine-tuning the models.

3.2.1. Input

Preprocessing on the Models

Preprocessing on BERT, BioBERT, and BioGPT to deal with their subword tokenisation characteristics. Subword tokenisation is a technique that breaks words into smaller subword units based on their frequency in a text corpus. The method is particularly useful for languages with morphology, where words can have different meanings. By using subword tokenisation, BERT can better handle out-of-the-vocabulary words, improving the performance of NLP models by reducing the occurrence of out-of-vocabulary words and enhancing the handling of rare words (Devlin et al., 2018). For instance, instead of tokenising a sentence by splitting it based on spaces and ignoring punctuations and symbols, the Hugging Face tokeniser takes both punctuations and symbols into account (see Figure 4 and 5). This then ensures that a model does not have to learn every possible representation of a word and every other punctuation that could attach, leading to more accurate preprocessing. In terms of dealing with BioGPT's subword tokenisation issue, a custom preprocessing method was used, and more details can be found in the following 3.3.2. BioNER Implementation Data Preprocessing section.

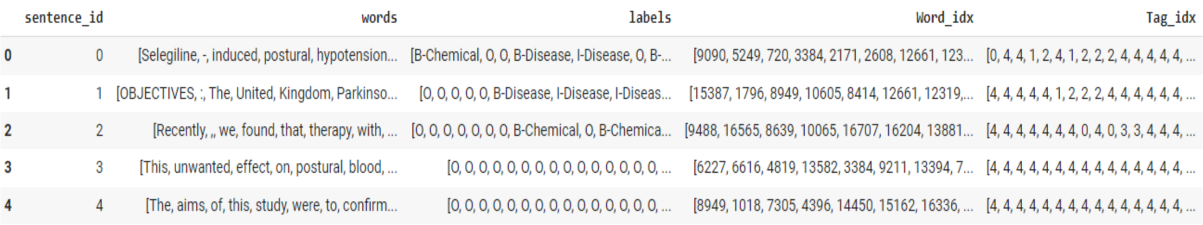
[‘The’, ‘United’, ‘Kingdom’, ‘Parkinson’s’, ‘s’, ‘Disease’, ‘Research’ , ‘GROUP(’, ‘‘UKPDRG)’, ‘trial’...]

Figure 4: Tokenising Text by Splitting It at Spaces

[‘The’, ‘United’, ‘Kingdom’, ‘Parkinson’, “ ‘ “, ‘s’, ‘Disease’, ‘Research’ , ‘GROUP’, ‘(‘, ‘UKPDRG’, ‘)’, ‘trial’...]

Figure 5: Tokenising Text with Tokensier

On the other hand, for ELMo word embeddings, word-to-index and index-to-word mapping were created for conversions for words before the training process and after prediction (see Figure 6).



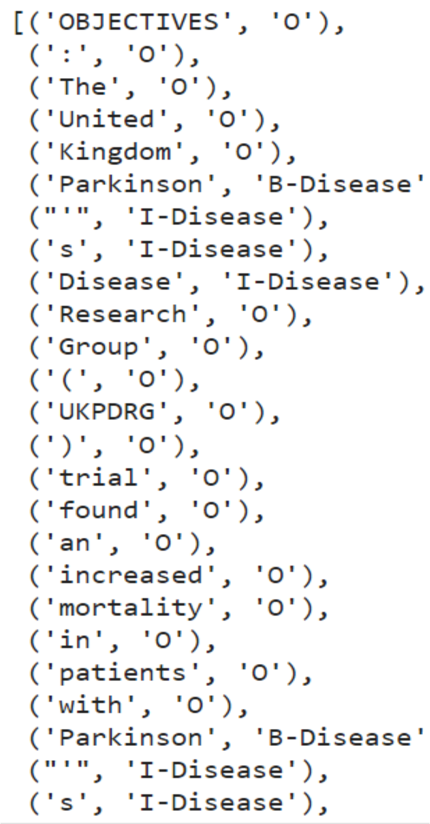
	sentence_id	words	labels	Word_idx	Tag_idx
0	0	[Selegiline, -, induced, postural, hypotension...	[B-Chemical, O, O, B-Disease, I-Disease, O, B-...	[9090, 5249, 720, 3384, 2171, 2608, 12661, 123...	[0, 4, 4, 1, 2, 4, 1, 2, 2, 2, 4, 4, 4, 4, 4, ...
1	1	[OBJECTIVES, :, The, United, Kingdom, Parkinso...	[O, O, O, O, O, B-Disease, I-Disease, I-Diseas...	[15387, 1796, 8949, 10605, 8414, 12661, 12319,...	[4, 4, 4, 4, 4, 1, 2, 2, 2, 4, 4, 4, 4, 4, 4, ...
2	2	[Recently, ,, we, found, that, therapy, with, ...	[O, O, O, O, O, O, O, B-Chemical, O, B-Chemica...	[9488, 16565, 8639, 10065, 16707, 16204, 13881...	[4, 4, 4, 4, 4, 4, 4, 0, 4, 0, 3, 3, 4, 4, 4, ...
3	3	[This, unwanted, effect, on, postural, blood, ...	[O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, ...	[6227, 6616, 4819, 13582, 3384, 9211, 13394, 7...	[4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ...
4	4	[The, aims, of, this, study, were, to, confirm...	[O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, ...	[8949, 1018, 7305, 4396, 14450, 15162, 16336, ...	[4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ...

Figure 6: ELMo’s Word-to-index and Index-to-word Mapping

Input Data: Fed the pre-trained models with the annotated BC5CDR training dataset. The annotated entities in the dataset include ‘B-Disease’, ‘B-Chemical’, ‘I-Disease’, ‘I-Chemical’, and ‘O’.

3.2.2. Output

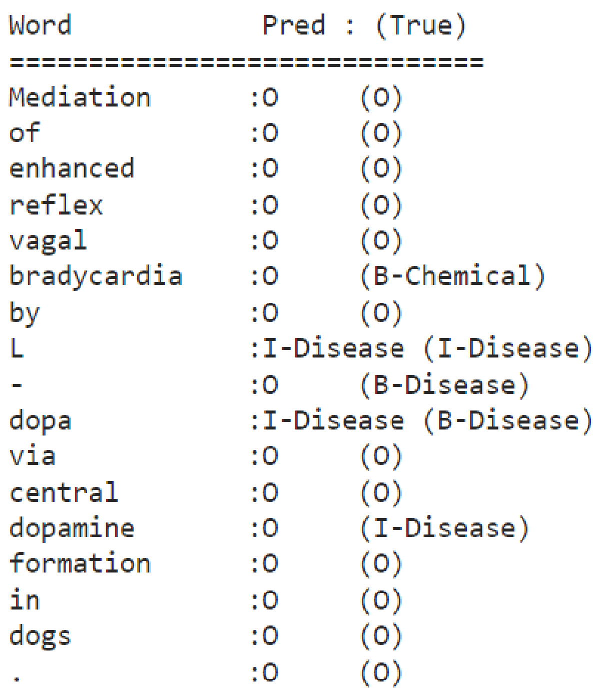
See Figure 7 for an example output of chunked words with their corresponding entity after preprocessing.



```
[('OBJECTIVES', 'O'),
 (':', 'O'),
 ('The', 'O'),
 ('United', 'O'),
 ('Kingdom', 'O'),
 ('Parkinson', 'B-Disease'
 ("'", 'I-Disease'),
 ('s', 'I-Disease'),
 ('Disease', 'I-Disease'),
 ('Research', 'O'),
 ('Group', 'O'),
 ('(', 'O'),
 ('UKPDRG', 'O'),
 (')', 'O'),
 ('trial', 'O'),
 ('found', 'O'),
 ('an', 'O'),
 ('increased', 'O'),
 ('mortality', 'O'),
 ('in', 'O'),
 ('patients', 'O'),
 ('with', 'O'),
 ('Parkinson', 'B-Disease'
 ("'", 'I-Disease'),
 ('s', 'I-Disease'),
```

Figure 7: Example of Chunked Words with Their Corresponding Named Entities

See Figure 8 for an example of model prediction output after training.



```
Word            Pred : (True)
==============================
Mediation      :O      (O)
of             :O      (O)
enhanced       :O      (O)
reflex         :O      (O)
vagal          :O      (O)
bradycardia    :O      (B-Chemical)
by             :O      (O)
L              :I-Disease (I-Disease)
-              :O      (B-Disease)
dopa           :I-Disease (B-Disease)
via            :O      (O)
central        :O      (O)
dopamine       :O      (I-Disease)
formation      :O      (O)
in             :O      (O)
dogs           :O      (O)
.              :O      (O)
```

Figure 8: Example of Model Prediction Output

3.2.3. Types of Linguistic Noise and Error Rate Range

Sharou et al. (2021) highlight the need to first differentiate between ‘harmful noise’ - content that should be removed or normalised because it distorts the NLP systems and 'useful noise’ - content that should be retained to improve the performance of such systems. The research focused on ‘harmful noise’, specifically pertaining to linguistic errors. That is, the research looked at linguistic noise introduced by humans which harms the overall task result. The linguistic noise includes homophone error, typographical error, grammatical error, and spacing error (Lee et al., 2020) (see Table 1).

Error Type	**Cause of Error**	**Example**
Homophone Error	Words that sound identical but are spelled differently	peace/piece
Typographical Error	Striking an incorrect key on keyboard	from /form

Grammatical Error	User did not know exactly what the difference between grammars	among/between
Spacing Error	Wrong blank between words	maybe/may be

Table 1: Types of Errors and Examples (Lee et al., 2020)

After defining the noise to be implemented to the dataset, the error rates for noise implementation were set as follows: 5%, 10%, 20%, 30%.

3.2.4. Structure of the Experiments

Experiments with Original Dataset

Firstly, BERT, BioBERT, and BioGPT were fine-tuned on the BC5CDR training set for the BioNER task. Alongside these fine-tuned models, ELMo was also utilised. However, different to most fine-tuning approaches, a feature-based approach was applied on ELMo for feature extraction. This means instead of manually updating ELMo's parameters during training, its pre-trained embeddings were directly used as features for other models. The necessary training parameters were configured, and the models were evaluated on the test set. Finally, the precision, recall, and F1 scores for each model were recorded to assess their performance.

Experiments with Noisy Dataset

The noise experiments followed the same structure as those conducted with the original dataset. However, in this stage, four types of noise were introduced into the dataset, with one error added at a time. The models were then run on the noisy dataset, and their performance was recorded to compare against the baseline model performance.

3.2.5. Models

Both general and domain-specific transformer models and word embeddings were used in the experiments.

General: BERT and ELMo

Domain-specific models: BioBERT and BioGPT

3.2.6. Evaluation

Both quantitative and qualitative methodologies were applied to evaluate model performance and their output patterns. For quantitative methodology, using the test sets, model performance was then recorded with their individual Precision, Recall, and F1-score (see equation 1 in 3.3.1.5. Model Evaluation). To dig into where the differences are, the research did not want to just look at the performance numbers. Therefore, regarding qualitative methodology, the differences in performance between each model were examined, and the output patterns at various levels of introduced noise were analysed. That is, the research focused on cases where one model performed correctly while another made errors. For qualitative analysis on the output results, the research identified different types of noise, inconsistent corpus annotation, different entity types predicted, overlapping entities or entities within entities, and entity boundaries which refers to incorrect identification of entity start and end positions. In addition, the research looked for patterns in the errors, such as sensitivity to specific types of noise or consistent mistakes.

Assumption

In terms of noise implementation experiment, with the literature supported by Leanman et al. (2015) that human-introduced noise can also be a contributing factor to the challenges in BioNER, the research assumed that the introduction of four types of linguistic noise would lead to a significant performance drop in the models.

3.3. BioNER Implementation

3.3.1. Data Collection

First, the BC5CDR dataset of annotated text was aggregated from Github. The annotated dataset is composed of train, dev, and test sets, and contains examples of text where the named entities are labelled.

3.3.2. Data Preprocessing

Since the data is formatted in the CoNLL IOB type format, I formatted it into a pandas dataframe with the columns: words, pos, chunk, and labels. In the dataframe, I paid attention to 'words', 'labels', and 'sentence_ids' columns, namely the word token, IOB label, and sentence_id to differentiate sentences. Given BERT and BioBERT do 'subword tokenisation,' in which it breaks words into smaller units (e.g. 'hypertension' into 'h', '##yper', '##tens', '##ion'), I hence did some adjustments to the gold labels to deal with this. In order to predict the correct set of entity types, I fine-tuned the two models using the training data so that they predict the correct set of entity types. Regarding ELMo, since it generates word embeddings using character-level convolutions followed by bi-directional LSTM networks, it does not rely on subword tokenisation like BERT. For BERT and BioBERT, I conducted the following: tokenisation using BERT's tokeniser, handling special tokens, and aligning subword tokens. First, text needed to be tokenised before feeding it to models. With BERT's tokensier, it splitted the text into word tokens and mapped them to their corresponding IDs in BERT's own vocabulary. Next, BERT requires special tokens such as '[CLS], '[SEP]' to be introduced for making the start and separation of tokens. The [CLS] token stands for class and is placed at the beginning, while [SEP] token refers to separating sentences for the next sentence prediction task. Next, since a single word after BERT's wordpiece tokenisation spans multiple tokens, I assigned the entity label to the first token of the split word and assigned a label for its later subword tokens.

A custom preprocessing process was used to deal with BioGPT's subword tokenisation. For tokenisation and label alignment, I manually tokenise each word and align the labels to the first token of each word. Subsequent subword tokens get a special label ('-100') to exclude them from loss calculation. For padding, I tokenised sequence sequences and then padded them to a consistent maximum length. This ensures consistent input size for model training.

In terms of ELMo, I first converted every word with its named entities into a list of tuples [(word, named entity),]. Since ELMo is word embeddings and does character-level tokenisation, it captures the semantics and syntax of words. They are numerical representations of words and words with semantic or syntactic similarity close to each other. Hence, word-to-index and index-to-word mapping were created for conversions for words for later training and prediction use.

3.3.3. Model Fine-tuning

When fine-tuning BERT, BioBERT, and BioGPT, I adapted the pre-trained models for later BioNER tasks. The 'transformers' library from Hugging Face was used. The step includes manually setting up hyperparameters and finding the optimal performance from custom setting. The hyperparameters include the max sequence length, learning rate, train batch size, number of train epochs. Max sequence length refers to the number of words the model will accept from a sentence at once. When exceeded, the extra part will be cut off or processed in chunks. Learning rate determines the step size during the optimisation process. Train batch size is the number of train examples the model processes. The number of train epochs refers to the number of times the model trains. The developers of BERT suggest that 10 is the optimal number of epochs (Devlin et al., 2018). However, an increased number of epochs might expose the model to overfitting. Optimiser and weight decay parameters were included for fine-tuning BERT and BioBERT. In terms of optimiser, AdamW was used, a commonly used optimiser for BERT with weight decay, which helps control the model to stay more general and work well with new unseen data. This is because AdamW optimiser applies weight decay directly to the weights and thus helps with generalisation and prevents overfitting. For more details on fine-tuning and setting up hyperparameters, I referred Liu and Wang (2021)'s work on hyperparameter optimisation for fine-tuning pre-trained language models and Mosbach et al. (2021)'s work focuses specifically on fine-tuning BERT. The key here was to adjust the model's weights using the labelled BC5CDR dataset to achieve satisfactory performance on the NER task. In terms of fine-tuning ELMo, Tensorflow Hub was used for loading ELMo and 'torch' was used for training the PyTorch model. Unlike transformer models which require manually adjusting hyperparameters, fine-tuning ELMo does not require setting hyperparameters and requires a feature-based extraction approach. Hence, at ELMo's fine-tuning stage, the process includes loading the pre-trained ELMo through Tensorflow Hub, feeding the tokenised sentences to obtain contextualised embeddings, and training the model on the BC5CDR dataset. Last, to prevent models from overfitting or underfitting, the number of train epochs was carefully managed.

3.3.4. Model Training

In the training step, the models were trained using the parameters listed in Table 2, 3, and 4. BERT, BioBERT, and BioGPT were fetched from the Hugging Face Transformers library (Wolf et al., 2020). ELMo was fetched through Tensorflow Hub. In the model training stage, the models were trained to adjust to the data. For training evaluation, evaluation data was used to avoid overfitting or underfitting.

Max Sequence Length	512
Learning Rate	2e-5
Train Batch Size	16
Number of Train Epochs	3
Optimiser	AdamW
Weight Decay	0.01

Table 2: Training Parameters of BERT

Max Sequence Length	512
Learning Rate	2e-5
Train Batch Size	16
Number of Train Epochs	3
Optimiser	AdamW
Weight Decay	0.01

Table 3: Training Parameters of BioBERT

Max Sequence Length	128
Learning Rate	1e-3
Train Batch Size	16
Number of Train Epochs	5

Table 4: Training Parameters of BioGPT

3.3.5. Model Evaluation

With the trained NER models, they were evaluated on the test set for their performance. Evaluation metrics including Precision, Recall, and F1-score were used to assess how well the models identify and classify named entities (see equation 1). Precision measures how accurate the model is and indicates the proportion of predicted entities that are accurately labelled. It is the ratio between the correctly identified positives (i.e. true positives) and all identified positives. Recall evaluates the model's effectiveness in identifying actual positives and is represented as the ratio of correctly predicted true positives to the total number of actual entities labelled. The F1-score is used as a balanced measure of both precision and recall. In sequence labelling tasks, NER models assign a label to each token in a sentence (e.g., 'B-Disease' for the beginning of a disease entity, 'I-Chemical' for tokens inside a chemical entity, or 'O' for outside any named entity). In order to count a success vs. false positive or false negative (see equation 1), True Positive (TP), False Positive (FP), and False Negative (FN) are used to determine if the model correctly identifies a named entity. True Positive (TP) refers to the model identifying the entity correctly, both entity type and its exact boundaries or span.

False Positive (FP) occurs when the model identifies an entity mistakenly. This happens either the entity boundaries are incorrect or the model incorrectly identifies an entity type. False Negatives (FN) happens when the model fails to identify an entity that it is supposed to.

$$p = \frac{TP}{TP + FP}, \; r = \frac{TP}{TP + FN}, \; f = \frac{2 * p * r}{p + r}$$

Equation 1: Equations of precision, recall, and f1-score

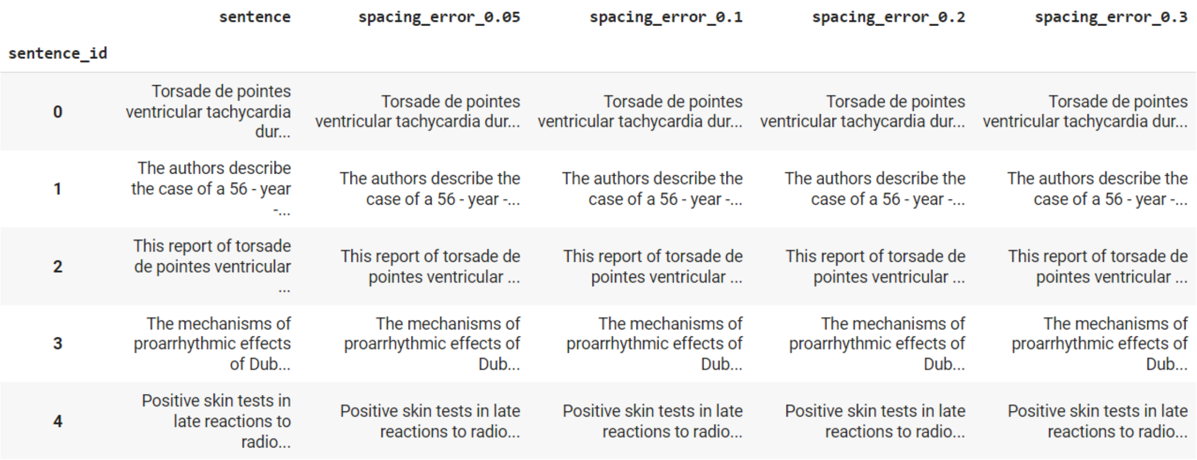
3.3.6. Inference

In this step, models for prediction and inference on unseen text were used. By providing the model with new input, it applied the knowledge it had learned during training to make informed inferences. The step is important because if a model can generalise well, it means it has truly learned the underlying pattern of the data. The step was included to ensure the ability of models to generalise and perform accurately outside of the training environment.

3.4. Noise Implementation

Followed the same process in 3.2.1.1. Data Collection, 3.2.1.2. Data Preprocessing, 3.2.1.3. Model Fine-Tuning, and 3.2.1.4. Model Training, I here introduced the four types of noise: homophone error, typographical error, grammatical error, and spacing error, from 5% to 30% noise level, to the BC5CDR test set. The results were saved into four csv files for the BioNER tasks later. The following presents examples of the BC5CDR dataset with different types of linguistic errors applied.

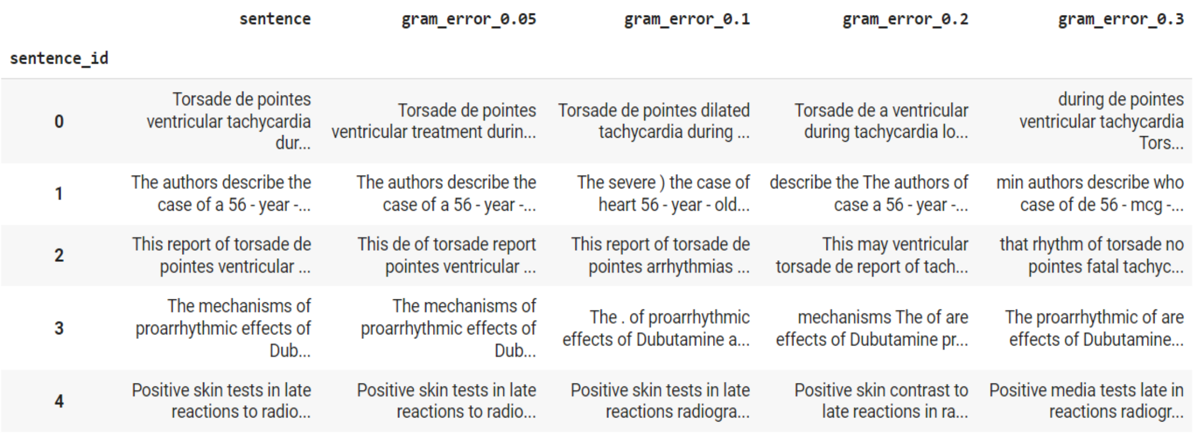
3.4.1. BC5CDR Dataset Applied with Varying Rates of Spacing Errors



sentence_id	sentence	spacing_error_0.05	spacing_error_0.1	spacing_error_0.2	spacing_error_0.3
0	Torsade de pointes ventricular tachycardia dur...	Torsade de pointes ventricular tachycardia dur...	Torsade de pointes ventricular tachycardia dur...	Torsade de pointes ventricular tachycardia dur...	Torsade de pointes ventricular tachycardia dur...
1	The authors describe the case of a 56 - year -...	The authors describe the case of a 56 - year -...	The authors describe the case of a 56 - year -...	The authors describe the case of a 56 - year -...	The authors describe the case of a 56 - year -...
2	This report of torsade de pointes ventricular ...	This report of torsade de pointes ventricular ...	This report of torsade de pointes ventricular ...	This report of torsade de pointes ventricular ...	This report of torsade de pointes ventricular ...
3	The mechanisms of proarrhythmic effects of Dub...	The mechanisms of proarrhythmic effects of Dub...	The mechanisms of proarrhythmic effects of Dub...	The mechanisms of proarrhythmic effects of Dub...	The mechanisms of proarrhythmic effects of Dub...
4	Positive skin tests in late reactions to radio...	Positive skin tests in late reactions to radio...	Positive skin tests in late reactions to radio...	Positive skin tests in late reactions to radio...	Positive skin tests in late reactions to radio...

Figure 9: BC5CDR Dataset with Varying Rates of Spacing Errors

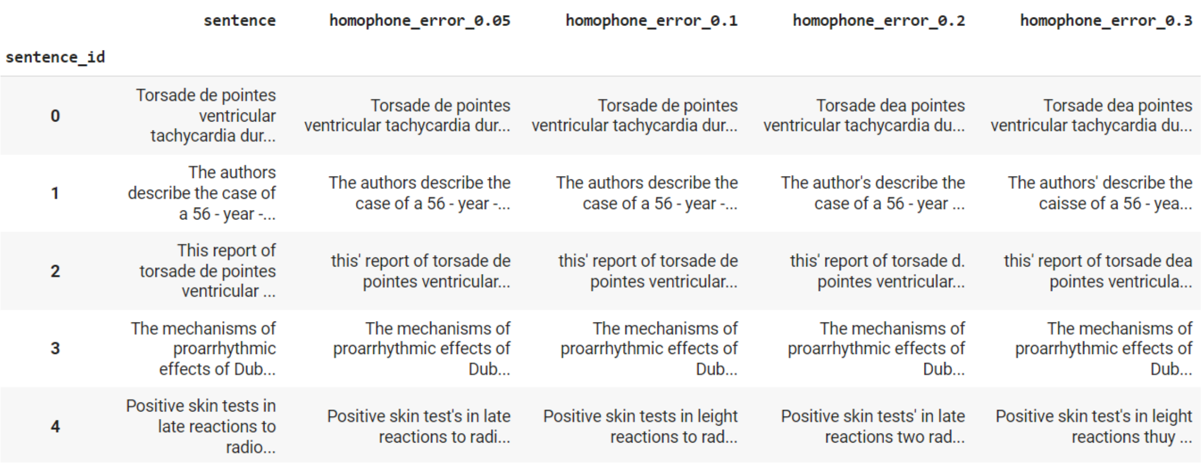
3.4.2. BC5CDR Dataset Applied with Varying Rates of Grammatical Errors



sentence_id	sentence	gram_error_0.05	gram_error_0.1	gram_error_0.2	gram_error_0.3
0	Torsade de pointes ventricular tachycardia dur...	Torsade de pointes ventricular treatment durin...	Torsade de pointes dilated tachycardia during ...	Torsade de a ventricular during tachycardia lo...	during de pointes ventricular tachycardia Tors...
1	The authors describe the case of a 56 - year -...	The authors describe the case of a 56 - year -...	The severe) the case of heart 56 - year - old...	describe the The authors of case a 56 - year -...	min authors describe who case of de 56 - mcg -...
2	This report of torsade de pointes ventricular ...	This de of torsade report pointes ventricular ...	This report of torsade de pointes arrhythmias ...	This may ventricular torsade de report of tach...	that rhythm of torsade no pointes fatal tachyc...
3	The mechanisms of proarrhythmic effects of Dub...	The mechanisms of proarrhythmic effects of Dub...	The . of proarrhythmic effects of Dubutamine a...	mechanisms The of are effects of Dubutamine pr...	The proarrhythmic of are effects of Dubutamine...
4	Positive skin tests in late reactions to radio...	Positive skin tests in late reactions to radio...	Positive skin tests in late reactions radiogra...	Positive skin contrast to late reactions in ra...	Positive media tests late in reactions radiogr...

Figure 10: BC5CDR Dataset with Varying Rates of Grammatical Errors

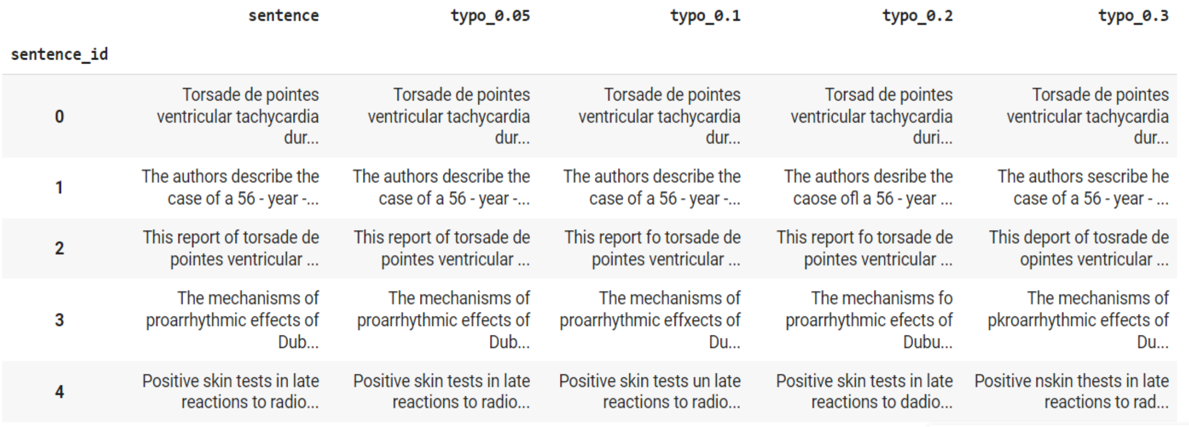
3.4.3. BC5CDR Dataset Applied with Varying Rates of Homophone Errors



sentence_id	sentence	homophone_error_0.05	homophone_error_0.1	homophone_error_0.2	homophone_error_0.3
0	Torsade de pointes ventricular tachycardia dur...	Torsade de pointes ventricular tachycardia dur...	Torsade de pointes ventricular tachycardia dur...	Torsade dea pointes ventricular tachycardia du...	Torsade dea pointes ventricular tachycardia du...
1	The authors describe the case of a 56 - year -...	The authors describe the case of a 56 - year -...	The authors describe the case of a 56 - year -...	The author's describe the case of a 56 - year ...	The authors' describe the caisse of a 56 - yea...
2	This report of torsade de pointes ventricular ...	this' report of torsade de pointes ventricular...	this' report of torsade de pointes ventricular...	this' report of torsade d. pointes ventricular...	this' report of torsade dea pointes ventricula...
3	The mechanisms of proarrhythmic effects of Dub...	The mechanisms of proarrhythmic effects of Dub...	The mechanisms of proarrhythmic effects of Dub...	The mechanisms of proarrhythmic effects of Dub...	The mechanisms of proarrhythmic effects of Dub...
4	Positive skin tests in late reactions to radio...	Positive skin test's in late reactions to radi...	Positive skin tests in leight reactions to rad...	Positive skin tests' in late reactions two rad...	Positive skin test's in leight reactions thuy ...

Figure 11: BC5CDR Dataset with Varying Rates of Homophone Errors

3.4.4. BC5CDR Dataset Applied with Varying Rates of Typographical Errors



	sentence	typo_0.05	typo_0.1	typo_0.2	typo_0.3
sentence_id					
0	Torsade de pointes ventricular tachycardia dur...	Torsade de pointes ventricular tachycardia dur...	Torsade de pointes ventricular tachycardia dur...	Torsad de pointes ventricular tachycardia duri...	Torsade de pointes ventricular tachycardia dur...
1	The authors describe the case of a 56 - year -...	The authors describe the case of a 56 - year -...	The authors describe the case of a 56 - year -...	The authors desribe the caose ofl a 56 - year ...	The authors sescribe he case of a 56 - year - ...
2	This report of torsade de pointes ventricular ...	This report of torsade de pointes ventricular ...	This report fo torsade de pointes ventricular ...	This report fo torsade de pointes ventricular ...	This deport of tosrade de opintes ventricular ...
3	The mechanisms of proarrhythmic effects of Dub...	The mechanisms of proarrhythmic effects of Dub...	The mechanisms of proarrhythmic effxects of Du...	The mechanisms fo proarrhythmic efects of Dubu...	The mechanisms of pkroarrhythmic effects of Du...
4	Positive skin tests in late reactions to radio...	Positive skin tests in late reactions to radio...	Positive skin tests un late reactions to radio...	Positive skin tests in late reactions to dadio...	Positive nskin thests in late reactions to rad...

Figure 12: BC5CDR Dataset with Varying Rates of Typographical Errors

Evaluation

Likewise, all precision, recall, and f-score were used to evaluate NER model performance on the noisy dataset. The key purpose of the evaluation metrics here was to evaluate how these NER models were robust to noise.

4. Results and Discussions

4.1. Results

4.1.1. Baseline NER Performance on BC5CDR Dataset

The model with the single highest overall balanced accuracy and results on the three metrics was BioGPT (see Table 5 and Figure 13). Despite the results, it is challenging to determine which model has the best NER model performance among BERT, BioBERT, and ELMo, as the precision, recall, and f-score of the models were all very similar. Hence, the research concludes that the four fine-tuned models all demonstrated comparable and strong BioNER performance on the BC5CDR dataset, with BioGPT showing a slightly better balanced performance than the others.

The result that BioGPT slightly outperformed the others could be attributed to the fact that it was already pre-trained on substantial amounts of biomedical text data (Luo et al., 2022) and thus had the ability to better identify different biomedical named entities compared to others. Although BioBERT was also pre-trained on a large amount of biomedical text data, it did not perform as well as BioGPT, even showing the lowest NER performance among the four models. This could be due to BioGPT being more effectively fine-tuned on the BC5CDR dataset, which allowed it to achieve higher NER performance.

Both ELMo and BERT achieved high balanced accuracy metrics due to their massive pre-training on large corpus (Peters et al., 2018; Devlin et al., 2018). Although published bioNER experiments using ELMo are less common than those using BERT or BioBERT, previous analyses on the BC5CDR dataset have demonstrated BERT's strong performance for bioNER tasks within this domain (Peng et al., 2019). One major departure that my study has from previous results is that BERT outperforms BioBERT. That is, my study finding is not the same as Lee et al. (2019) and Luo et al. (2022). However, the experiment result is consistent with Peng et al. (2019)'s findings that their fine-tuned Base BERT outperforms BioBERT. Looking at the results, the high precision and recall indicated that all the four models handled the named entities effectively.

Model	**P**	**R**	**F**
BERT	0.85	0.92	0.89
ELMo	0.89	0.91	0.88
BioBERT	0.83	0.91	0.87
BioGPT	0.94	0.94	0.94

Table 5: Performance of NER Models on BC5CDR Dataset, Evaluated Using IOB Tag Matching

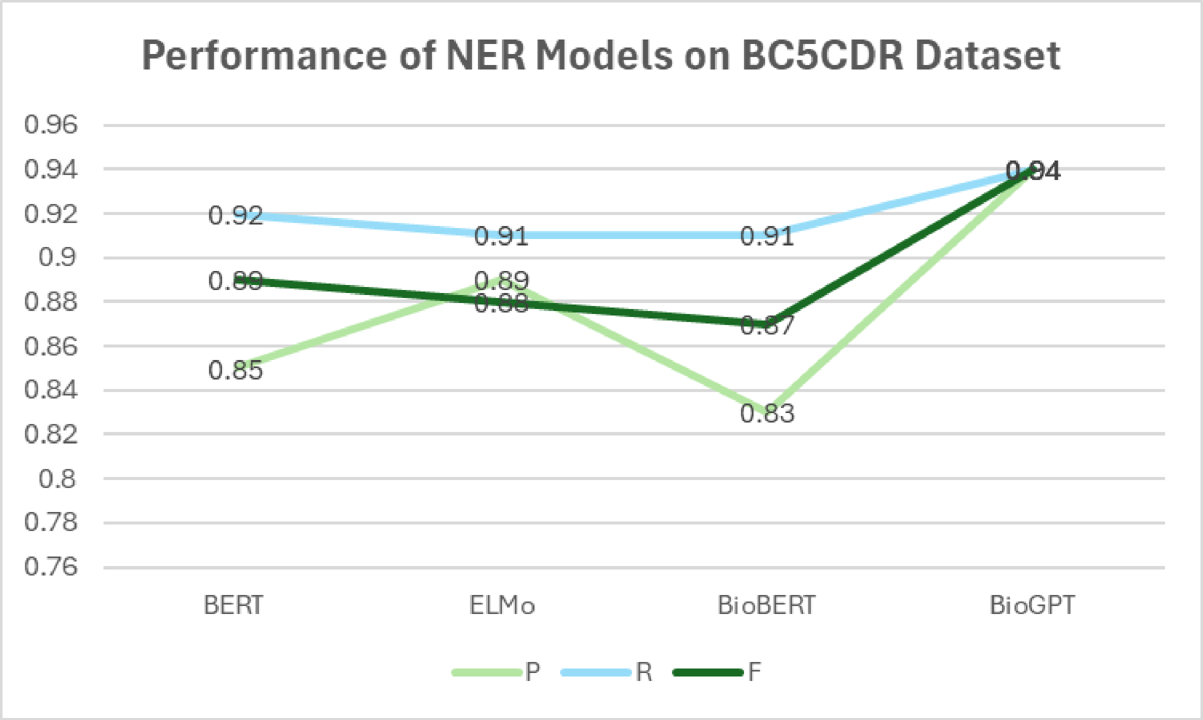


Figure 13: Line Plot Presentation of Table 5

4.1.2. NER Model Performance on Noisy BC5CDR Dataset

While BioBERT's model performance remained unchanged across all the four types of noise introduced into the BC5CDR dataset, BERT experienced a slight f-score drop when 10%, 20%, and 30% of grammatical errors, as well as 5%, 10%, 20%, and 30% of typographical errors, were introduced into the dataset. To summarise the findings on the noisy dataset: BERT and BioBERT’s performance remained unchanged when various levels of spacing errors were introduced (see Table 6 and Figure 14). Similarly, BERT, BioBERT, and ELMo maintained consistent performance with different levels of homophone errors (see Table 8 and Figure 16). However, while most models experienced a slight performance drop with grammatical and typographical errors, BioBERT’s performance remained unchanged (see Table 7 and Table 9).

With the introduction of 10%, 20%, and 30% grammatical errors, BERT's f-score slightly decreased from 89% to 88%. However, BERT’s precision surprisingly rose from 85% to 88% when 20% and 30% of grammatical errors were introduced (see Table 7 and Figure 15). In terms of ELMo, it experienced a slight precision and recall drop when both 20% and 30% level of typographical errors were introduced to the dataset (see Table 9 and Figure 17). Likewise, BERT’s f-score also dropped slightly when various levels of typographical errors were introduced into the dataset (see Table 9 and Figure 17). Interestingly, when it comes to

BioGPT, although it demonstrated the highest bioNER model performance among the four models on the clean BC5CDR dataset, its precision, recall, and F-score decreased more than those of the other models when exposed to spacing, grammatical, and typographical errors. For instance, while other models remained steady when exposed to different levels of spacing errors, BioGPT's precision significantly dropped from 94% to 82% and its f-score dropped from 94% to 83% when 30% spacing errors were introduced to the dataset (see Table 6 and Figure 14).

Put simply, the spacing, homophone, and typo errors did not significantly disrupt or degrade the performance of the models, except for BioGPT. However, it is crucial to identify the essence of the BC5CDR dataset and factors affecting the performance. There are several factors that could potentially attribute these departures from previous results to. These factors include the size of the dataset, data preprocessing, model architecture, and hyperparameters setting, and noise implementation, as they are all contributing factors to the achieved results. This highlights the challenge of defining the best-performing model without thorough and comprehensive experimentation. Several human factors come into play, such as how the dataset was processed and hyperparameters were set, as well as the fact that transformers and word embeddings are built on different model architectures. More in-depth error analysis and contributing factors are discussed in the next 4.2. Discussions section.

Spacing Error

Error Rate	**5%**			**10%**			**20%**			**30%**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
BERT	0.85	0.92	0.89	0.85	0.92	0.89	0.85	0.92	0.89	0.85	0.92	0.89
ELMo	0.88	0.91	0.88	0.88	0.91	0.88	0.88	0.91	0.87	0.87	0.91	0.87
BioBERT	0.83	0.91	0.87	0.83	0.91	0.87	0.83	0.91	0.87	0.83	0.91	0.87
BioGPT	0.86	0.87	0.87	0.86	0.87	0.86	0.84	0.86	0.85	0.82	0.85	0.83

Table 6: Performance of NER models on BC5CDR Dataset with Spacing Errors, Evaluated Using IOB Tag Matching

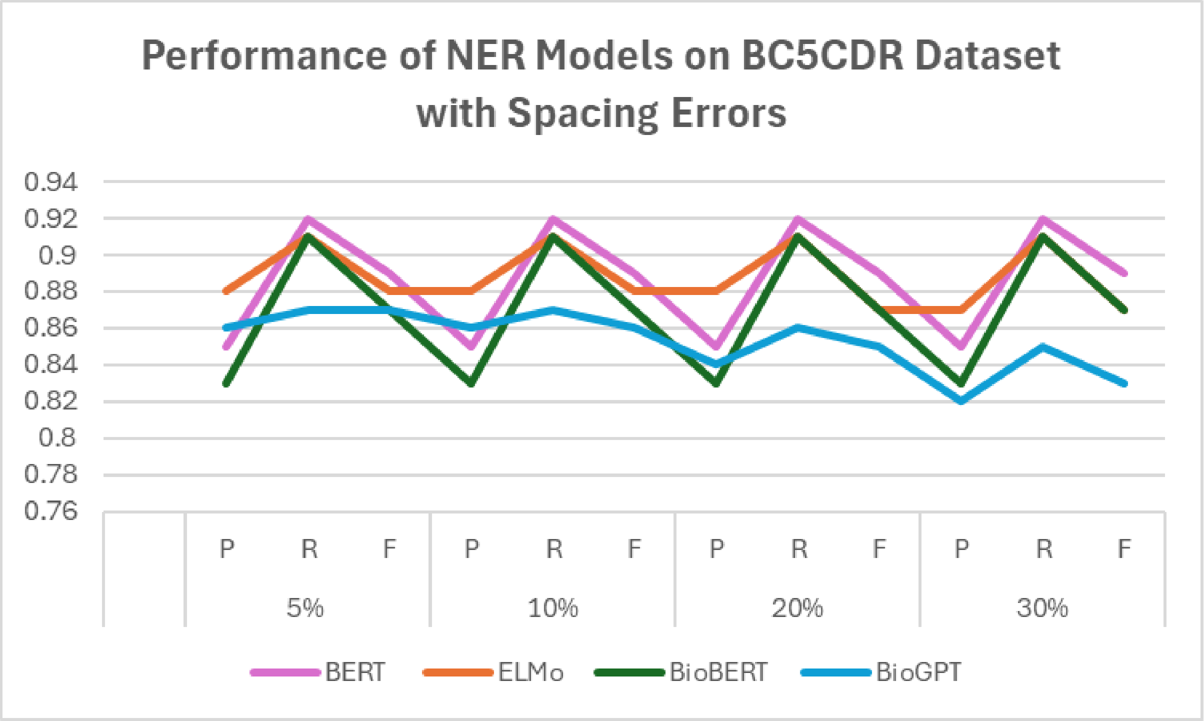


Figure 14: Line Plot Presentation of Table 6

Grammatical Error

Error Rate	**5%**			**10%**			**20%**			**30%**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
BERT	0.85	0.92	0.89	0.85	0.92	0.88	0.88	0.92	0.88	0.88	0.92	0.88
ELMo	0.88	0.91	0.88	0.87	0.91	0.87	0.87	0.91	0.87	0.87	0.91	0.87
BioBERT	0.83	0.91	0.87	0.83	0.91	0.87	0.83	0.91	0.87	0.83	0.91	0.87
BioGPT	0.86	0.87	0.87	0.86	0.87	0.87	0.86	0.88	0.87	0.86	0.88	0.87

Table 7: Performance of NER Models on BC5CDR Dataset with Grammatical Errors, Evaluated Using IOB Tag Matching

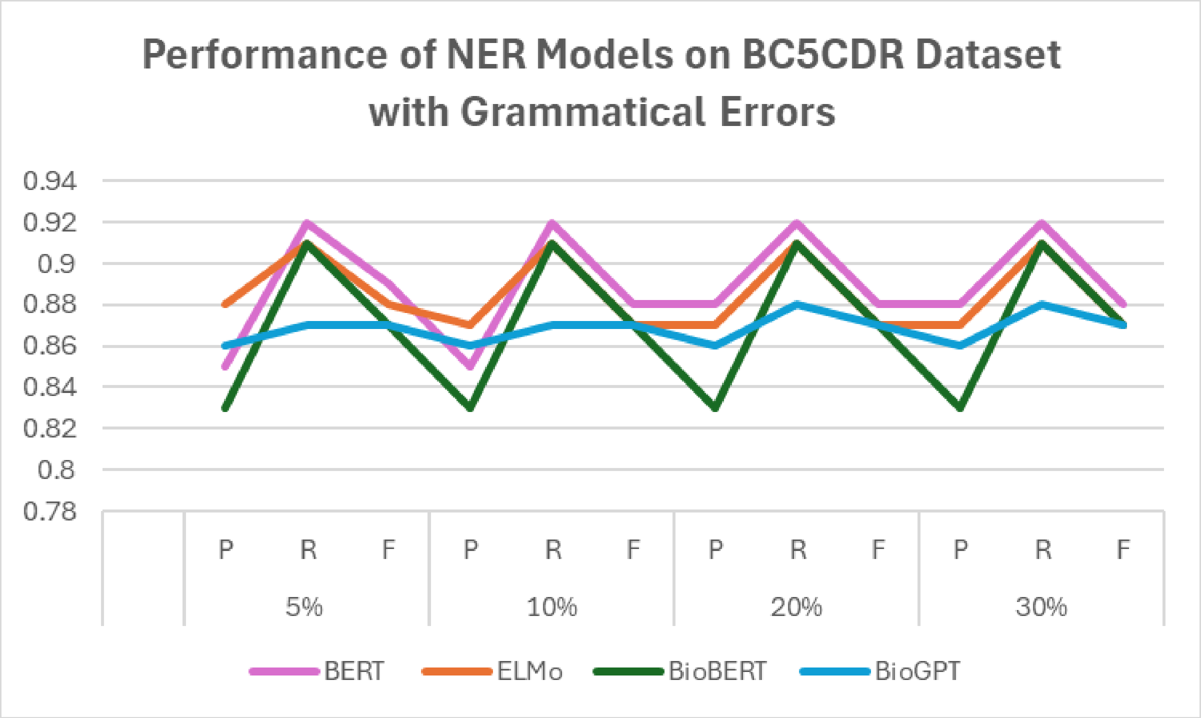


Figure 15: Line Plot Presentation of Table 7

Homophone Error

Error Rate	**5%**			**10%**			**20%**			**30%**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
BERT	0.85	0.92	0.89	0.85	0.92	0.89	0.85	0.92	0.89	0.85	0.92	0.89
ELMo	0.89	0.91	0.88	0.89	0.91	0.88	0.89	0.91	0.88	0.89	0.91	0.88
BioBERT	0.83	0.91	0.87	0.83	0.91	0.87	0.83	0.91	0.87	0.83	0.91	0.87
BioGPT	0.93	0.94	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93

Table 8: Performance of NER Models on BC5CDR Dataset with Homophone Errors, Evaluated Using IOB Tag Matching

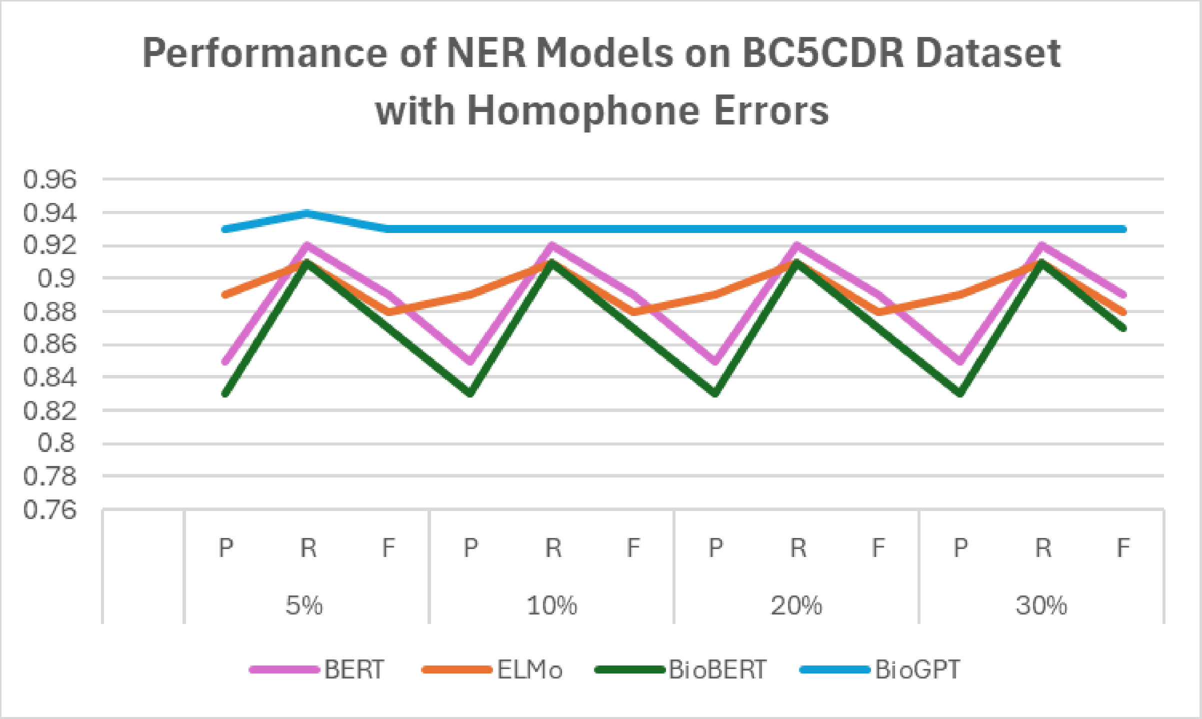


Figure 16: Line Plot Presentation of Table 8

Typographical Error

Error Rate	**5%**			**10%**			**20%**			**30%**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
BERT	0.85	0.92	0.88	0.85	0.92	0.88	0.85	0.92	0.88	0.85	0.92	0.88
ELMo	0.88	0.91	0.88	0.88	0.91	0.88	0.87	0.91	0.87	0.87	0.91	0.87
BioBERT	0.83	0.91	0.87	0.83	0.91	0.87	0.83	0.91	0.87	0.83	0.91	0.87
BioGPT	0.86	0.87	0.86	0.86	0.86	0.86	0.85	0.86	0.86	0.85	0.86	0.85

Table 9: Performance of NER Models on BC5CDR Dataset with Typographical Errors, Evaluated Using IOB Tag Matching

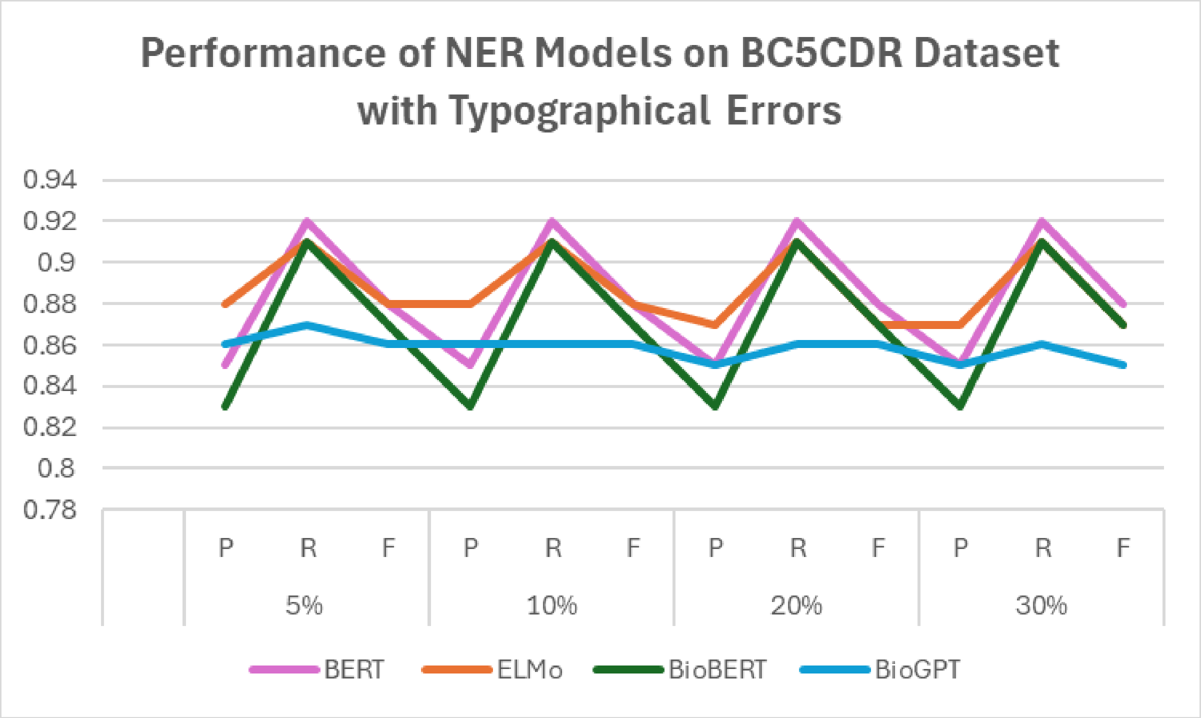


Figure 17: Line Plot Presentation of Table 9

4.2. Discussions

4.2.1. Results Analysis

Interpretation of the Results

In investigating RQ1, the research quantifies the level of noise by using noise metrics like error rate and compares the noise levels in different noisy data to assess its distribution. Based on the experimental results, the research concludes that the four types of linguistic noise do not cause a significant performance drop in the models, except for BioGPT on the dataset with spacing, homophone, and typo errors. That is, with less than 30% of the defined linguistic errors introduced to the dataset, all BERT, ELMo, BioBERT, and BioGPT can still perform BioNER well on the dataset. The defined linguistic errors, then cannot be categorised as 'harmful noise' (Sharou et al., 2021) to BERT, ELMo, and BioBERT, as they do not significantly decrease the performance of the three. In terms of the identified prediction errors, the study summarises the error cases on the test dataset as follows: different entity types predicted, inconsistent annotation within the dataset, and instances where the model incorrectly identifies non-entities as entities or fails to recognise entities.

Surprisingly, BERT's precision slightly increases as grammatical errors increase. Regarding the scenario where there is an increase in precision while recall and f-score remain the same is not a typical scenario. Instead, this very likely indicates that there exists a bias or a flaw in the dataset with grammatical errors. Since the models were evaluated for overfitting concerns, the research suggests that this may be due to bias in the noisy data. In other words, the increase in grammatical errors might have introduced a bias that favours BERT's ability to correctly identify positive instances, potentially due to an overrepresentation of certain error types. For instance, if the rise in grammatical errors is largely attributed to specific mistakes or grammatical rules (e.g., pronoun usage), the model may become more adept at handling these errors, resulting in improved precision despite other influencing factors. Additionally, the introduced errors could create correlations between certain mistakes, making it easier for BERT to recognise and address them. On the other hand, conversely, BERT’s f-score decreased slightly, while precision and recall remain unchanged when typographical error rates of 5%, 10% 20%, and 30% are introduced into the dataset. The scenario in which recall and precision remained constant while the f-score dropped slightly due to typographical errors in the dataset is expected and is a common occurrence in NLP tasks. Based on these results, the research suggests that this slight decrease in f-score might be attributed to BERT's architecture or training process, which may not be fully optimised to handle typographical errors.

When examining NER model output patterns across different noisy datasets, the results from the noisy datasets do not significantly differ from those produced on the clean BC5CDR dataset. For example, in the case of spacing errors, this might be because such errors were filtered out during tokenisation, as the tokeniser primarily splits words based on spaces. As for homophone errors, since they resemble typographical errors after the introduction of noise and do random punctuation additions, even with varying levels of homophone errors, the performance of all NER models remained unchanged. These findings are not only supported by numerical figure results but also verified by manually reviewing the model outputs from the noise implementation experiments.

In investigating RQ2, the experiments show that noise causes subtle or nearly zero effects on both general and domain-specific models. However, there is a clear contrast between the two biomedical domain-specific models, BioBERT and BioGPT. Put simply, while BioBERT shows the lowest balanced performance among the four models, its BioNER performance remains consistent with that on the original BC5CDR dataset. In contrast, although BioGPT achieves the highest precision, recall, and F1-score, it is sensitive to spacing, grammatical, and typographical errors. Earlier, the research question was set under the assumption that domain-specific pre-trained models such as BioBERT and BioGPT would outperform BERT and ELMo. From the experiments, nevertheless, apart from BioGPT, the results show almost no differences between models’ baseline performance on the clean dataset and the noisy dataset. Based on the experimental results, the research concludes that BioBERT is more resistant to linguistic noise compared to general-domain pre-trained models like BERT and ELMo. However, with the inclusion of BioGPT, it is observed that BERT, ELMo, and BioBERT all demonstrate better robustness to noise than BioGPT.

With such results, the study dives deep into exploring the possible factors involved in affecting model performance.

Factors Affecting Model Performance

Dataset Quality

Although BC5CDR is a reputable biomedical dataset, the utilised dataset contains only ‘chemical’ and ‘disease’ entity types. In theory, since the experiment is about comparing the noise robustness of the selected models and word embeddings, it would be more relevant if label types could be included as many as possible. Since the study aims to understand change in performance from noise, what matters is the relative difference in performance under different noise conditions, rather than the absolute performance overall. Hence, including more labels gives a better picture of real-world variability which is very useful and not commonly done for biomedical NLP. As mentioned, with only two entity types included, there exists a limitation on the BioNER task and there might be a possibility that BERT and BioBERT are coincidentally good at these two entity types: chemical and disease. In terms of diversifying and balancing the existing two entity types, conducting manual annotations and applying other supervised learning methods can both be the solutions to the issue. However, due to technical and time constraints, the study utilised the annotated BC5CDR dataset which has been provided publicly. On the other hand, the imbalance in entity types could also be influencing the results. Among the entity types—'B-Chemical', 'B-Disease', 'I-Chemical', 'I-Disease', and 'O'—'O' entities dominate the dataset. As a result, even randomly assigning 'O' to a word would likely result in a correct match.

Data Preprocessing

In theory, effective data preprocessing can help improve model performance. In this case, when doing BioNER with transformers such as BERT and their variants, subword tokenisation becomes such a crucial issue to handle. For transformers, during the experimentation phase, if the subword tokenisation issue was not handled correctly, it would very likely screw up the whole model performance results. While employing wordpiece subword tokenisation can help BERT and BioBERT handle subword tokenisation problems respectively to handle out-of-vocabulary words, this can add complexity to the models, making the models difficult to align

NER labels with tokens. Also, given some named entities span multiple tokens, making it challenging to accurately extract entities. When it comes to the impacts of subword tokenisation on NER, not only it can lead to inaccurate label boundaries, but also BERT models trained on general corpus, could exhibit biases towards certain common entities but fail to identify the ones which are less common. In our case, despite BERT, ELMo, and BioBERT achieving high F-scores on both noisy and clean datasets, there is a possibility—besides the data imbalance issue mentioned earlier, where even assigning 'O' to random words might result in a high match—that the high performance of both models could be due to the commonality of the words and entities they encountered.

Hyperparameters Tuning

In terms of fine-tuning, the research also observes that all the models are highly sensitive to hyperparameters such as batch size and learning rate, and overfitting is a common challenge. For instance, when fine-tuning BioGPT, I observed that performance starts to decline when the number of epochs is above 10, hence suggesting an optimal setting of around 5 epochs. When fine-tuning BERT, a batch size of 16 was used for training. This is not consistent with the batch size of 32 suggested by Devlin et al. (2018). The experiments also tested batch sizes of 32 and 64, finding that 16 yielded high performance. Additionally, when fine-tuning BERT, the optimal performance is generally achieved after three or four epochs rather than 10 epochs suggested by Devlin et al. (2018), leading to a final setting of 3 epochs in the experiments for BERT and BioBERT. This finding does not align with the recommendations of Devlin et al. (2018). In addition, although the utilised models are pre-trained models already, they are still computationally intensive models, requiring massive computational resources for effective fine-tuning and training.

Model Architectures

Different model architectures and characteristics can result in different model performance results. From the experiments, BERT and BioBERT exhibit strong performance on the BioNER task. Supported by Devlin et al. (2018)'s literature on bidirectional transformers, which highlights features such as bidirectionality, attention mechanisms, and self-attention, the research concludes that the bidirectional transformer architecture used by BERT and BioBERT is a key factor contributing to their strong performance on the BioNER task. Not only can bidirectional transformers capture context from both preceding and following tokens, but they also utilise attention and self-attention mechanisms to capture contextual information. When it comes to handling noise, the attention mechanism enables the model to focus on relevant information, enhancing its robustness against the introduced noise in the data. Unlike

BERT, BioGPT uses a bidirectional text processing mechanism. BioGPT employs a unidirectional approach but still achieves the highest overall performance. This exceptional performance can be attributed to BioGPT's massive pre-training on biomedical text and its specialised fine-tuning process, which enhance its ability to accurately identify and classify biomedical entities. In terms of ELMo, since it is commonly used as a feature extractor rather than like transformers that are fine-tuned directly for BioNER, from the experiments, the research observes with ELMo's contextual embeddings, this feature-based approach leverages ELMo's ability to do BioNER tasks well. In addition, also with its bidirectional text processing approaches, it helps ELMo to capture the entire context and make informed entity prediction.

4.2.2. Comparison of Model Performance to Previous Results

Previous studies have demonstrated that transformers outperform LSTMs on various NLP tasks and that transformers consistently achieve better performance than BiLSTM architectures (Lai & Lu, 2020). Based on the experimental results, the research cannot agree with the above finding as BioBERT's BioNER model performance does not outperform ELMo's. Also, the models all demonstrate similar and high performance on the dataset. On the other hand, Peng et al. (2019) indicate that BERT-Base outperforms state-of-the-art (SOTA) models in most biomedical NLP tasks, as measured by the f-score. In their study, when using the BC5CDR dataset focused on disease entities, the fine-tuned BERT-Base model (f-score: 86.6) outperformed BioBERT (f-score: 85.9). This trend is also reflected in the BC5CDR-chemical dataset, where their fine-tuned BERT-Base achieved an f-score of 93.5, compared to BioBERT's 93.0 (Peng et al., 2019). Their finding that BERT outperforms BioBERT on BioNER tasks is consistent with my experiment findings. However, these findings differ from those of Lee et al. (2019) and Luo et al. (2022), who conclude that BioBERT outperforms BERT. As discussed in Section 4.1 and earlier, the key findings of this research are that BERT outperforms BioBERT and that BioGPT surpasses ELMo, BERT, and BioBERT in performance. Although Wang et al. (2023) and Xie et al. (2023) have pointed out that GPT models demonstrate poor performance on NER tasks compared to conventional supervised models, the research observes that BioGPT achieves the highest performance among BERT, BioBERT, and ELMo. Also, despite NER not being included in the NLP benchmark tasks for GPT models that have achieved SOTA results as reported by Radford et al. (2018), BioGPT demonstrates significant potential in the NER area.

4.2.3. Comparison of Baseline Performance and Performance with Noise

The overall NER model performance of BERT, ELMo, and BioBERT on the noisy dataset is nearly identical to their performance on the original BC5CDR dataset. Even when introducing one type of linguistic noise at a time, the models produce nearly the same results as on the clean dataset. Nevertheless, although demonstrating the highest model performance on the clean dataset, it is observed that BioGPT is sensitive to spacing, grammatical, and typographical errors. Next, it is crucial to note that the error rates introduced to the dataset ranged from 5% to 30%, specifically at intervals of 5%, 10%, 20%, and 30%. The study suggests that either the maximum error rate of 30% is still considered subtle and minor for both the dataset and models, or the likelihood of errors affecting critical words is low. Put simply, with the utilisation of the python ‘random’ library, looking at their output patterns, it is hard to control which words would receive errors. In addition, if the chunked word was already a nonsense, for example, a punctuation or an ‘O’ entity type, adding linguistic errors to them would not affect the NER models overall accuracy. In addition, take homophone noise implementation for example, by manually counting, with the set 30% error rate, I calculate manually and observe that there is only 23% error rate introduced. Apart from the possible randomness issues mentioned, this also indicates that the homophone database might not be able to cover all the homophones for the noise implementation.

4.2.4. Transfer Learning

‘After supervised learning - Transfer Learning will be the next driver of machine learning commercial success.’ - Andrew Ng, founder of DeepLearning.AI.

In theory, traditional machine learning and deep learning are designed to work in isolation. Since knowledge is not retained or accumulated, conventional machine learning has to be rebuilt from scratch once the task changes, whole transfer learning overcomes the isolated learning paradigm and is able to learn new tasks faster that relies on the previous learned tasks. In the field of biomedical NLP, BioNER (Devlin et al., 2018) is one of the most prominent applications of transfer learning and has shown better performance in the NER accuracy. From the experiments, the study observes that all transformers exhibit strong transfer learning characteristics. Additionally, these characteristics contribute to reduced computational training time and enhanced performance in named entity recognition accuracy. On top of transformers demonstrating strong transfer learning ability as mentioned in the earlier literature review section, ELMo’s strong transfer learning ability is also observed. However, from the experiments, I observe that ELMo uses a feature-based approach rather than a fine-tuning approach like BERT. That is, in contrast to BERT, where it allows fine-tuning the model for a specific task, ELMo embeddings can be seen as a feature extractor and are incorporated into other models for later NER tasks. In the scope of contextual transfer learning, BioGPT model emerges as an alternative also as it is a biomedical version of GPT and could serve as a clear contrast to the others. Similar to ELMo, BioGPT model captures contextual information within a sentence, which makes it crucial in the BioNER area. GPT (Radford et al., 2018), which stands for Generative Pre-trained Transformer, like BERT (Devlin et al, 2019) is also based on transformer architecture and pre-trained on a large corpus of text data using a masked language modelling objective. This includes randomly replacing some of the input tokens with a mask token and training the model to predict the original token. After pre-training, GPT can also be fine-tuned on a specific dataset for NER tasks.

4.2.5. Model Inclusion

BioELMo (Jin et al., 2019) was planned to be included as a biomedical domain variant contrats to ELMo. Unfortunately, at the time I finished the experiments with ELMo, I faced challenges in fine-tuning and getting BioELMo for use due to the following reasons: 1) BioELMo is no longer available through TensorFlow Hub or AllenNLP, making it difficult to conduct BioNER tasks with it. In addition, even the official GitHub page for AlllenNLP is deprecated.; 2) There is a lack of documentation and community support for BioELMo which complicates troubleshooting and customisation.; 3) BioELMo has dependency conflicts with the current versions of TensorFlow and PyTorch, requiring significant adjustments to set up.; 4) Fine-tuning from scratch without pre-training resources demands substantial amounts of computation resources, which is not feasible with limited time and the current infrastructure. Given the reasons listed above, I do not consider the exclusion of BioELMo from this research to be a limitation, especially since point 1 already highlighted the inherent challenges associated with working with BioELMo.

From previous studies, GPT models have not emerged as top performers for BioNER tasks; instead, they have often been utilised for pre-annotation approaches, such as zero-shot learning (Moradi et al., 2021). Essentially, GPT models are text generation models (Qin et al., 2023) and are not designed with the same objectives as BERT and ELMo. Although GPT models can be fine-tuned for specific biomedical NLP tasks, like BERT, even the proponents of GPT-NER (Wang et al., 2023) did not detail the fine-tuning process, focusing instead on prompt engineering. The research, however, overcomes the challenges and concludes that BioGPT has strong bioNER capability and its ability to generate texts in essence has potential for future generative or creative tasks.

4.3. Critical Evaluation of Work

4.3.1. Strengths

While human-introduced noise is prevalent in both everyday life and the biomedical field, there is limited research on how noise impacts biomedical literature and NLP tasks. This study employs a novel approach by introducing various types of linguistic noise into the dataset to compare against baseline performance. Specifically, the study first defines different types of noise encountered in daily life and develops tailored linguistic noise for BioNER experiments. Additionally, to assess the models' NER performance and robustness to noise, the study examines both transformers and embeddings, highlighting their contrasts.

4.3.2. Limitations

The research acknowledges the challenges involved in managing biomedical texts, noisy datasets, and the appropriate use of large language models (LLMs). For example, the noise implementation process included defining and developing noise scripts, applying only one type of noise to the dataset at a time with varying error rates, and testing each model's performance individually. Due to these complexities, the noise implementation takes longer than anticipated, which consequently reduces the time available for the actual implementation of BioNER tasks and conducting more comprehensive experiments on the noisy dataset. Taking the same noise implementation case mentioned above, looking back at hindsight, to have a better understanding of the role of noise in biomedical NLP, the study could have experimented three cases: a clean training set with a noisy test set, a noisy training set with a clean test set, and a noisy training set with a noisy test set.

5. Future Work and Conclusion

BioNER is often the fundamental task in biomedical language processing and sets the basis for later downstream tasks such as relation extraction and question answering. For instance, with the name entities extracted, they can then be used for detecting and classifying relations among various biomedical concepts within texts and even assisting with clinical decision making (Demner-Fushman et al., 2009). Common examples of extraction of specific biomedical relationship types include chemical induced disease relationships and drug-drug interactions.

In the dissertation, different types of noise were first identified, with a particular focus on linguistic noise, which was clarified for discussion in relation to the research questions. Next, BioNER was introduced with an illustrative example, followed by a comparison between transformers and biLSTMs, as discussed in Chapter 2. Chapter 3 then detailed the BioNER methodology and its implementation. In Chapter 4, the results of experiments conducted on both clean and noisy datasets were presented, along with an analysis and discussion of the findings. Finally, the study concludes that BioGPT demonstrated the highest balanced BioNER model performance among BERT, BioBERT, and ELMo. However, in terms of robustness to noise, BioBERT shows stronger resistance to noise on the dataset compared to BERT, BioBERT, and ELMo.

Motivated by Kahneman et al. (2021)'s *Noise: A Flaw in Human Judgment*, the study decided to focus on the role noise plays in our real-life scenarios and applications. In the biomedical field, with increasing volume of biomedical literature generated at a high demand for extracting knowledge from it (Zhao et al., 2021; Naseem et al., 2022), clinical notes and biomedical reports have become popular sources for information extraction. Fortunately, with the advancements of LLMs and other AI technologies, they have revolutionised and accelerated the efficiency and speed for retrieving valuable information from unstructured texts. These improvements have been particularly impactful in specialised fields like biomedical NLP, where the accurate extraction of information is critical for research and applications. For RQ1, when examining the impact of common human-introduced linguistic noise on biomedical NLP tasks, the research concludes from the experiments that subtle noise does not significantly harm or downgrade model performance on bioNER tasks, except for BioGPT. That is, even with 30% of the defined linguistic errors introduced to the dataset, BERT, ELMo, and BioBERT, still perform well on BioNER tasks. Therefore, the defined linguistic errors cannot be categorised and generalised as harmful noise for BERT, ELMo, and BioBERT. For RQ2, the experiments reveal that noise has a minor impact on both general and domain-specific models on top of BioGPT. Nevertheless, a noticeable contrast exists between the two biomedical models, BioBERT and BioGPT. While BioBERT has the lowest balanced performance among the four models, its BioNER results on the noisy dataset remain consistent with the original BC5CDR dataset. In contrast, although BioGPT achieves the highest precision, recall, and F1-score, it is sensitive to spacing, grammatical, and typographical errors.

6. References

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *PubMed*, 17–21. https://pubmed.ncbi.nlm.nih.gov/11825149

Andrade, C. (2022). Understanding Statistical Noise in Research: 1. Basic Concepts. Indian Journal of Psychological Medicine, 45(1), 89–90. https://doi.org/10.1177/02537176221139665

Chinchor, N., & Robinson, P. (1997). *MUC-7 named entity task definition*. In *Proceedings of the 7th Conference on Message Understanding* (pp. 1–21).

Cho, H., & Lee, H. (2019). Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-3321-4

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, *42*(5), 760–772. https://doi.org/10.1016/j.jbi.2009.08.007

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2016, June 10). *De-identification of Patient Notes with Recurrent Neural Networks*. arXiv.org. https://arxiv.org/abs/1606.03475

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). 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, Volume 1 (Long and Short Papers) (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Ebadi, N., Morgan, K., Tan, A., Linares, B., Osborn, S., Majors, E., Davis, J., & Rios, A. (2024). Extracting Biomedical Entities from Noisy Audio Transcripts.

Ethayarajh, K., & Jurafsky, D. (2020). Utility is in the Eye of the User: A Critique of NLP Leaderboards. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4846–4853). Online: Association for Computational Linguistics.

Eftimov, T., Seljak, B. K., & Korošec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS ONE*, *12*(6), e0179488. https://doi.org/10.1371/journal.pone.0179488

Fried, H. O., Lovell, C. A. K., Schmidt, S. S., et al. (2002). Accounting for Environmental Effects and Statistical Noise in Data Envelopment Analysis. Journal of Productivity Analysis, 17, 157–174. https://doi.org/10.1023/A:1013548723393

Gao, S., Kotevska, O., Sorokine, A., & Christian, J. B. (2021). A pre-training and self-training approach for biomedical named entity recognition. *PloS one*, *16*(2), e0246310. https://doi.org/10.1371/journal.pone.0246310

Gardner, M., Grus, J., Neumann, M., Tafjord, Ø., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., & Zettlemoyer, L. S. (2018). *AllenNLP: A deep semantic natural language processing platform*. arXiv. https://doi.org/10.48550/arXiv.1803.07640

Huang, C., & Lu, Z. (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in Bioinformatics, 17(1), 132–144. https://doi.org/10.1093/bib/bbv024

Huang, Z., Xu, W., & Yu, K. (2015, August 9). *Bidirectional LSTM-CRF models for sequence tagging*. arXiv.org. https://arxiv.org/abs/1508.01991

Jie, Z., Xie, P., Lu, W., Ding, R., & Li, L. (2019). Better modeling of incomplete annotations for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 729–734).

Khurana, D., Koli, A., Khatter, K., et al. (2023). Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications, 82(14), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Kejriwal, M., Santos, H., Shen, K., et al. (2024). A noise audit of human-labeled benchmarks for machine commonsense reasoning. Scientific Reports, 14, 8609. https://doi.org/10.1038/s41598-024-58937-4

Kolchinsky, A., Lourenço, A., Wu, H., Li, L., & Rocha, L. M. (2015). Extraction of Pharmacokinetic Evidence of Drug–Drug Interactions from the Literature. *PLoS ONE*, *10*(5), e0122199. https://doi.org/10.1371/journal.pone.0122199

Jin, Q., Dhingra, B., Cohen, W., & Lu, X. (2019). Probing Biomedical Embeddings from Language Models. In Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP (pp. 82–89). Minneapolis, USA: Association for Computational Linguistics.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise. HarperCollins UK.

Kumar, S. (2017, May 10). *A survey of deep learning methods for relation extraction*. arXiv.org. https://arxiv.org/abs/1705.03645

Lafferty, J.D., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *International Conference on Machine Learning*.

Lai, P., & Lu, Z. (2020). BERT-GT: cross-sentence n-ary relation extraction with BERT and Graph Transformer. *Bioinformatics*, *36*(24), 5678–5685. https://doi.org/10.1093/bioinformatics/btaa1087

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 260–270). San Diego, California: Association for Computational Linguistics.

Leaman, R., Khare, R., & Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, *57*, 28–37. https://doi.org/10.1016/j.jbi.2015.07.010

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Lee, J., Kim, M., & Kwon, H. (2020). Deep Learning-Based Context-Sensitive Spelling typing error correction. IEEE Access, 8, 152565–152578. https://doi.org/10.1109/access.2020.3014779

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., & Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database, 2016, baw068. https://doi.org/10.1093/database/baw068

Liu, S., Tang, B., Chen, Q., & Wang, X. (2016). Drug-Drug interaction extraction via convolutional neural networks. *Computational and Mathematical Methods in Medicine*, *2016*, 1–8. https://doi.org/10.1155/2016/6918381

Liu, X., & Wang, C. (2021). *An Empirical study on Hyperparameter Optimization for Fine-Tuning Pre-trained Language Models*. arXiv.org. https://arxiv.org/abs/2106.09204

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, *23*(6). https://doi.org/10.1093/bib/bbac409

Mikolov, T., Chelba, C., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1312.3005

Missen, M. M. S., Naeem, A., Asmat, H., Salamat, N., Akhtar, N., Coustaty, M., & Prasath, V. B. S. (2020). Improving seller–customer communication process using word embeddings. *Journal of Ambient Intelligence and Humanized Computing*, *12*(2), 2257–2272. https://doi.org/10.1007/s12652-020-02323-1

Moradi, M., Blagec, K., Haberl, F., & Samwald, M. (2021, September 6). *GPT-3 Models are Poor Few-Shot Learners in the Biomedical Domain*. arXiv.org. https://arxiv.org/abs/2109.02555

Mosbach, M., Andriushchenko, M., & Klakow, D. (2020). *On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines*. arXiv.org. https://arxiv.org/abs/2006.04884

Naseem, U., Dunn, A. G., Khushi, M., & Kim, J. (2022). Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. BMC Bioinformatics, 23(1). https://doi.org/10.1186/s12859-022-04688-w

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics.

Peng, N., Poon, H., Quirk, C., Toutanova, K., & Yih, W. (2017). Cross-Sentence N-ary Relation Extraction with Graph LSTMs. Transactions of the Association for Computational Linguistics, 5, 101–115. https://doi.org/10.1162/tacl_a_00049

Peng, Y., Yan, S., & Lu, Z. (2019). *Transfer Learning in Biomedical Natural Language Processing: An evaluation of BERT and ELMO on ten benchmarking datasets*. arXiv.org. https://arxiv.org/abs/1906.05474

Peng, C. (2024, April 2). *NCBI; BC5CDR; i2b2 2010; HPRD50; AIMed; MedNLI*. IEEE DataPort. https://ieee-dataport.org/documents/ncbi-bc5cdr-i2b2-2010-hprd50-aimed-mednli

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics.

Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology*, *8*. https://doi.org/10.3389/fcell.2020.00673

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023, February 8). *Is ChatGPT a General-Purpose natural language processing task solver?* arXiv.org. https://arxiv.org/abs/2302.06476

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. language_understanding_paper.pdf (openai.com)

Ramachandran, R., & Arutchelvan, K. (2021). Named entity recognition on bio-medical literature documents using hybrid based approach. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-021-03078-z

Raza, S., Reji, D. J., Shajan, F., & Bashir, S. R. (2022). Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, *1*(12), e0000152. https://doi.org/10.1371/journal.pdig.0000152

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4902–4912). Online: Association for Computational Linguistics.

Rocktäschel, T., Weidlich, M., & Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, *28*(12), 1633–1640. https://doi.org/10.1093/bioinformatics/bts183

Sahu, S., & Anand, A. (2016). Recurrent neural network models for disease name recognition using domain invariant features. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2216–2225). Berlin, Germany: Association for Computational Linguistics.

Sharou, K. A., Li, Z., & Specia, L. (2021, September 1). Towards a better understanding of noise in natural language processing. ACL Anthology. https://aclanthology.org/2021.ranlp-1.7

Settles, B. (2004). *Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets*. ACL Anthology. https://aclanthology.org/W04-1221

Sleeman, D. H., & Gilhooly, K. (2023). Groups of experts often differ in their decisions: What are the implications for AI and machine learning? A commentary on Noise: A Flaw in Human Judgment, by Kahneman, Sibony, and Sunstein (2021). In the AI Magazine/AI Magazine, 44(4), 555–567. https://doi.org/10.1002/aaai.12135

Stuck, B. W., & Kleiner, B. (1974). A Statistical analysis of telephone noise. the Bell System Technical Journal, 53(7), 1263–1320. https://doi.org/10.1002/j.1538-7305.1974.tb02791.x

TensorFlow Hub (n.d.). Tensorflow. Retrieved 20th August, 2024 from GitHub - tensorflow/hub: A library for transfer learning by reusing parts of TensorFlow models.

Tong, Y., Zhuang, F., Zhang, H., Fang, C., Zhao, Y., Wang, D., Zhu, H., & Ni, B. (2022). Improving biomedical named entity recognition by dynamic caching inter-sentence information. Bioinformatics, 38(16), 3976–3983. https://doi.org/10.1093/bioinformatics/btac422

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. CoRR.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023, April 20). *GPT-NER: Named Entity Recognition via large Language models*. arXiv.org. https://arxiv.org/abs/2304.10428

Wei, C., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., Wiegers, T. C., & Lu, Z. (2016). Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database, 2016. https://doi.org/10.1093/database/baw032

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., & others. (2020). *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). Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-demos.6

Xie, T., Li, Q., Zhang, J., Zhang, Y., Liu, Z., & Wang, H. (2023). *Empirical Study of Zero-Shot NER with ChatGPT*. arXiv.org. https://arxiv.org/abs/2310.10035

Zhang, Z., & Chen, A. L. P. (2022). Biomedical named entity recognition with the combined feature attention and fully-shared multi-task learning. *BMC Bioinformatics*, *23*(1). https://doi.org/10.1186/s12859-022-04994-3

Zhao, S., Su, C., Lu, Z., & Wang, F. (2021). Recent advances in biomedical literature mining. Briefings in Bioinformatics, 22(3). https://doi.org/10.1093/bib/bbaa057

Zhu, Q., Li, X., Conesa, A., & Pereira, C. (2017). GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics*, *34*(9), 1547–1554. https://doi.org/10.1093/bioinformatics/btx815

7. Appendix

7.1. Ethics Approval Letter

Downloaded: 31/08/2024
Approved: 19/06/2024 Sheng-Jui Chang
Registration number: 230140637
Information School [a.k.a iSchool]
Programme: MSc Data Science

Dear Sheng-Jui
PROJECT TITLE: Noise in Biomedical NLP: Investigating Performance of Pre-trained Models on Biomedical NLP Tasks
APPLICATION: Reference Number 063028 This letter confirms that you have signed a University Research Ethics Committee-approved self-declaration to confirm that your research will involve only existing research, clinical or other data that has been robustly anonymised.

You have judged it to be unlikely that this project would cause offence to those who originally provided the data, should they become aware of it. As such, on behalf of the University Research Ethics Committee, I can confirm that your project can go ahead on the basis of this self-declaration. If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since full ethical review may be required.
Yours sincerely

Nia Hunt
Departmental Ethics Administrato

7.2. Research Diary and Reflection

Meeting 1 (May) Date:

Work completed:

2,50 word Dissertation proposal submission

Meeting contents:

Discussed about the methodology part for the dissertation project: Look at biomedical literature (e.g. autogenerated data like PubMed, WikiMed, Silver standard resource etc.), introduce noise to the literature, and see how different models perform on each biomedical tasks, namely Named-Entity Recognition (NER), Relation Extraction (RE), Document Classification (DC) (and Question Answering (QA)).

BERT and ELMo ar fine, but also consider including GPT

Planned goals:

To look at MedMentions for NER and SemRep for RE

Meeting 2 (June) Date:

Work completed:

Literature review section

Amendments to the Ethics Application form

Meeting contents:

Denis suggested the student incorporate GPT to the dissertation project and can help set up the student's HPC account. Also. Denis recommended using specific datasets as the student's suggested dataset (e.g. NCBI-Disease) is narrow ·

Planned goals:

Examining and using the recommended dataset

The student to receive proposal feedback from the supervisor

Meeting 3 (July) Date:

Work completed:

Research and review on deep learning techniques and how to use the deep learning models to conduct biomedical NER

Meeting contents:

Q n A session regarding the dissertation project with the supervisor

The supervisor provided details and suggestions on the linguistic noise (e.g. spacing error, grammatical errors etc.) implementation and tools to be used

Planned goals:

The student to focus on completing the code for implementing different types and levels of noise to the dataset

Meeting 4 (August) Date:

Work completed:

Experiments on the large language models

Meeting contents:

We discussed the NER implementation issues and the supervisor provided some possible solutions to BERT's subtokenisation issue. The suggested solutions include taking a couple of test examples or a single file and looking at the output. The supervisor also explained that the goal of a master's thesis is to engage in the research and its difference to a phd thesis. In terms of the noise implementation, the superior suggested three cases: baseline performance, clean train noisy test, noisy train noisy test, and noisy train clean test.

Planned goals:

To submit the draft to the supervisor by 20th August

Reflection on Research

Supervisor Feedback 1:

Reflection on feedback:

The supervisor provided very constructive and detailed feedback on the dissertation proposal. The feedback for each section (e.g. introduction, literature review, methodology) helped guide the direction for future work. For instance, the supervisor pointed out that the student needed to specify the deep learning models chosen and their relations to noise. In addition, the supervisor kindly reminded the student that the methodology part needed to be more specific and detailed.

Supervisor Feedback 2:

Reflection on feedback:

With the helpful feedback and a couple research papers provided by the supervisor, the student was then able to find the relevant biomedical literature regarding noise, biomedical NLP, deep learning, and NER. The supervisor also provided some useful information about the biomedical NLP tasks with clear contrast. These biomedical NLP tasks include: named entity recognition, relation extraction, and document classification.

Supervisor Feedback 3:

Reflection on feedback:

First, the supervisor helped review the linguistic noise python script for later noise implementation. With such help, the student was then able to ensure the noise was accurate. Also, the supervisor suggested incorporating GPT to the discussion. Although incorporating GPT to the experiment would lead to the student experiencing and solving some technical complexities, looking back at hindsight, it was much worth the effort.

Supervisor Feedback 4:

Reflection on feedback:

As always, the supervisor was very helpful in providing constructive feedback to the student during the entire dissertation project. The supervisor communicated well and provided much help to the student during the entire duration. The student received clear information and useful research papers on both technical and biomedical NLP domains. Also, the supervisor kindly helped request the student’s HPC access. Unfortunately, even with much time and effort spent on learning HPC, it turned out to be too much of a technical challenge for the student. For example, even with a successful job script submitted to the stanage cluster for later programme execution, the student struggled to find a way to retrieve the outputs from the cluster.

Supervisor Feedback 5:

Reflection on feedback:

As always, the supervisor was very helpful and provided possible solutions such as looking at a single file first and its corresponding output to deal with BERT’s subword tokenisation issue. The differences between a phd thesis and a master’s thesis was made clear to the student also. Finally, in terms of the noise experiments, including the cases truly helped spice up the discussion section in the dissertation.