The Open Biomedical Engineering Journal
DOI: 10.2174/0118741207289680240510045617, 2024, 18, e18741207289680
ISSN: 1874-1207
1

RESEARCH ARTICLE **OPEN ACCESS**

Exploring Biomedical Named Entity Recognition *via* SciSpaCy and BioBERT Models

Aman Jolly[1], Vikas Pandey[2], Indrasen Singh[3,*] and Neha Sharma[4]

[1]*KIET Group of Institutions, UP, India*
[2]*Department EE, School of Engineering, BBD University, Lucknow, UP, India*
[3]*School of Electronics Engineering, VIT Vellore, Tamil Nadu, India*
[4]*Delhi Technological University, Delhi, India*

Abstract:

Introduction: Biological Named Entity Recognition (BioNER) is a crucial preprocessing step for Bio-AI analysis.

Methods: Our paper explores the field of Biomedical Named Entity Recognition (BioNER) by closely analysing two advanced models, SciSpaCy and BioBERT. We have made two distinct contributions: Initially, we thoroughly train these models using a wide range of biological datasets, allowing for a methodical assessment of their performance in many areas. We offer detailed evaluations using important parameters like F1 scores and processing speed to provide precise insights into the effectiveness of BioNER activities.

Results: Furthermore, our study provides significant recommendations for choosing tools that are customised to meet unique BioNER needs, thereby enhancing the efficiency of Named Entity Recognition in the field of biomedical research. Our work focuses on tackling the complex challenges involved in BioNER and enhancing our understanding of model performance.

Conclusion: The goal of this research is to drive progress in this important field and enable more effective use of advanced data analysis tools for extracting valuable insights from biomedical literature.

Keywords: Biomedical, BioNER, SciSpaCy, BioBERT, Natural language processing, Named entity recognition.

*Address correspondence to this author at the School of Electronics Engineering, VIT Vellore, Tamil Nadu, India;
E-mail: erindrasen@gmail.com

Cite as: Jolly A, Pandey V, Singh I, Sharma N. Exploring Biomedical Named Entity Recognition *via* SciSpaCy and BioBERT Models. Open Biomed Eng J, 2024; 18: e18741207289680. http://dx.doi.org/10.2174/0118741207289680240510045617

Received: November 14, 2023
Revised: March 13, 2024
Accepted: April 08, 2024
Published: June 05, 2024

1. INTRODUCTION

Named Entity Recognition (NER) [1] is a crucial part of Natural Language Processing (NLP) that involves identifying and categorising named items in written text. Named entities encompass several categories such as personal names, organisational names, and geographical places, among others. The primary objective is to do this by transforming unstructured textual content into a machine-readable format. Named Entity Recognition (NER) is a crucial part of Information Extraction (IE) that involves the detection and classification of named items in unstructured text data. Named entities refer to a broad variety of entities, such as persons, businesses, geographical locations, precise dates, numerical values, and other pertinent categories. Named Entity Recognition (NER) is a prominent area of research that has attracted substantial academic attention and has several practical applications in the field of natural language processing. The use of NER approaches varies across different domains, giving rise to distinct difficulties and possibilities

within this subject. Deep learning techniques for NER have demonstrated encouraging outcomes across diverse areas such as healthcare, legal, and chemical fields. There have been proposals to utilise active learning methodologies, specifically Label Transfer Learning (LTP), as a means to mitigate the expenses associated with data annotation in the field of NER. In addition, there are shared projects and benchmarks, such as the ROCLING-2022 shared task focused on Chinese healthcare NER, which furnish datasets and assessment criteria to facilitate the comparison of diverse NER methodologies. Named Entity Recognition is a specific component within the broader field of Information Extraction that falls within the domain of NLP. The primary objective of this system is to discern and categorise designated items, such as individual names, organisational names, geographical places, dates, and other specific nouns, inside the text that lacks a predetermined framework [2-4]. Named Entity Recognition is an essential preprocessing task in several downstream NLP applications, including but not limited to question answering, conversation systems, and knowledge graph generation [3].

The use of Named Entity Recognition algorithms exhibits variability across many domains, owing to the distinctive characteristics inherent to each specific area [4]. Nevertheless, the procedure often encompasses other essential NLP stages, such as tokenization, part-of-speech tagging, parsing, and model construction [4]. In the field of medicine, for instance, Named Entity Recognition encompasses the task of recognising medical terminology, properties such as negation and severity, and establishing connections between terms and ideas inside domain-specific ontologies [5–9].

In the early stages of development, Named Entity Recognition (NER) systems mostly relied on manually constructed rules, lexicons, and ontologies. However, these systems need extensive dictionaries and meticulous manual feature creation, which constrained their ability to expand and reach maximum performance [10, 11]. In the era of Machine Learning, there are solutions that include machine learning concepts as a potential alternative [12, 13]. However, even these solutions required extensive feature engineering work, limiting their flexibility and effectiveness. The scenario saw significant shifts due to the emergence of the Deep Learning revolution. Recent advancements have brought about advanced deep learning structures like Bi-directional Long Short-Term Memory Networks (BiLSTM) combined with Conditional Random Fields (CRF) and GRAM-CNN [14]. The developments have far beyond the capability of older approaches, demonstrating improved precision and resilience. The greatest significant advancement in Named Entity Recognition (NER) technology occurred with the development of the BERT architecture (Bi-directional Encoder Representations from Transformers). BERT has introduced a new age by using contextual information, allowing NER systems to attain exceptional levels of accuracy and recall [15-18]. The capacity to understand subtle contextual signals has transformed NER accuracy, representing a significant advancement in biomedical text comprehension and data retrieval.

Named Entity Recognition can provide difficulties, particularly in situations involving low-resource languages or topics that necessitate a time-consuming and knowledge-intensive annotation procedure [4, 5, 19]. In order to tackle these issues, scholars have undertaken investigations into many methodologies, including few-shot learning. This particular methodology endeavours to train machine learning models with a very restricted amount of accessible data [19-21]. Furthermore, the integration of entity definition information and domain-specific resources, such as medical dictionaries and ontologies [8, 13, 14, 22], has been shown to enhance the efficacy of Named Entity Recognition models within certain domains [18, 23-32]. Multiple strategies are employed in order to enhance the precision of NER. Several approaches can be identified, including:

1. The process of data augmentation encompasses the generation of supplementary training instances through the modification or synthesis of pre-existing examples. This approach has the potential to enhance the generalisation capabilities of the Named Entity Recognition paradigm [33-36].

2. The hybrid approach involves integrating rule-based methodologies with machine learning techniques, namely Conditional Random Fields (CRF), to enhance the precision of Named Entity Recognition [34, 37, 38].

3. The inclusion of domain-specific resources, such as medical dictionaries and ontologies, together with entity definition information, has been found to enhance the effectiveness of named entity recognition models in certain domains [39-41].

4. Semi-supervised learning is a technique that involves training a model using a little amount of labelled data together with a large amount of unlabeled data. This technique has the potential to improve the effectiveness of named entity recognition models in situations typified by restricted resources, such as low-resource languages or domains where the annotation process requires substantial time, knowledge, and expertise [33, 42, 43]. One way to address the problem of noisy data and the introduction of unwanted patterns during training is to use an iterative training process and data-generating tools. This entails retraining the model using a subset of the original annotated dataset [33, 44, 45].

5. Fine-tuning and transfer learning are often employed techniques in the field of named entity recognition to enhance the efficiency and/or accuracy of NER models [46-48]. Researchers have achieved enhanced precision in NER models across several domains, such as healthcare, legal, and chemical domains, through the use of these methodologies [44, 49]. The job of Named Entity Recognition poses significant challenges in the context of clinical writing, primarily due to the intricate and diverse nature of medical language. Nevertheless, many methodologies may be employed to enhance the precision of Named Entity Recognition in clinical discourse [47].

Several approaches can be identified, including the integration of domain-specific resources, including entity definition information and specialised resources like medical dictionaries and ontologies, which have been shown to enhance the efficacy of named entity recognition models within the medical domain [18, 50]. The hybrid approach involves integrating rule-based methodologies with machine learning techniques, namely CRF, in order to enhance the precision of Named Entity Recognition [51]. Data augmentation is a methodology that entails the creation of supplementary training instances through the alteration or synthesis of preexisting examples. This approach has the potential to enhance the generalisation capabilities of the Named Entity Recognition paradigm [9]. Semi-supervised learning is a methodology that entails training a model using a small quantity of labelled data with a substantial quantity of unlabelled data. This approach has the potential to enhance the efficacy of Named Entity Recognition models in scenarios characterised by limited resources, such as low-resource languages or domains where the annotation process necessitates substantial effort, skill, and domain knowledge [50, 52]. Fine-tuning and transfer learning techniques can be employed to enhance the efficiency and/or accuracy of Named Entity Recognition models [53, 54]. The utilisation of a multilevel named entity recognition framework can effectively tackle the difficulties associated with clinical NER. This framework enables the construction of models that are capable of handling more intricate NER tasks [53, 55]. Through the use of these methodologies, scholars have successfully enhanced the precision of Named Entity Recognition models within the medical field. As an illustration, a research endeavour pertaining to the identification and extraction of adverse drug events from clinical literature by the use of medicine and associated techniques yielded F1 scores of 93.45% for Named Entity Recognition [50]. A recent investigation pertaining to the identification of medication names and their corresponding properties from discharge summaries demonstrated a notable F1-score of 0.91 in the context of the 2018 National NLP Clinical Challenges (n2c2) shared task [50]. Identifying uncommon medical conditions and their associated clinical presentations, together with the influence of socio-economic determinants of health, can provide challenges within the clinical literature. Moreover, the intricate and diverse nature of medical vocabulary presents a formidable obstacle for Named Entity Recognition when applied to clinical material. The concept of stupor poses a challenge in terms of its specific clinical definition [5]. Nevertheless, scholars have put forth various methodologies to enhance the precision of Named Entity Recognition in clinical text [56-59]. These methodologies encompass the integration of domain-specific resources, the adoption of a hybrid approach, data augmentation techniques, semi-supervised learning methods, fine-tuning and transfer learning strategies, as well as the utilisation of a multilevel NER framework [58, 60–62]. Identifying abbreviations and acronyms in clinical material is difficult for Named Entity Recognition algorithms. Nevertheless, many methodologies may be employed to enhance the precision of Named Entity Recognition in clinical discourse containing abbreviations and acronyms. Several approaches can be identified, including the use of domain-specific resources, such as medical dictionaries and ontologies, that enhance the ability of named entity recognition models to identify abbreviations and acronyms in the clinical text [37]. The hybrid approach involves integrating rule-based methodologies with machine learning techniques, namely CRF, in order to enhance the precision of Named Entity Recognition [51]. Data augmentation is a methodology that entails the creation of supplementary training instan-ces through the alteration or synthesis of preexisting examples. This approach has the potential to enhance the generalisation capabilities of the Named Entity Recog-nition model [55]. Fine-tuning and transfer learning are two techniques that may be employed to enhance the efficiency and/or accuracy of Named Entity Recognition models [53].The utilisation of a multilevel named entity recognition framework presents a viable solution to tackle the obstacles encountered in clinical NER. This framework enables the construction of models that cater to more intricate NER tasks [63].

Through the use of these methodologies, scholars have successfully enhanced the precision of Named Entity Recognition models when applied to clinical material containing abbreviations and acronyms. As an illustration, a research investigation on clinical named entity identification attained an F1-score of 0.91 on the 2018 National NLP Clinical Challenges (n2c2) shared task [27, 50]. When applied to the field of biomedical research, the Biomedical Named Entity Recognition (BioNER) technique is an essential part of the process of biomedical text mining [44, 64-66]. The enormous task of recognising and categorising the numerous things that can be found in biomedical literature is taken on by BioNER. These entities include genes, diseases, chemical names, and medicinal names. The ever-increasing number of biological entities, their plethora of synonyms, the popularity of abbreviations, the use of lengthy entity descriptions, and the combination of letters, symbols, and punctuation within these texts all contribute to this complexity. BioNER, integral to numerous applications in the field of natural language processing, plays a pivotal role in biomedical literature mining [25, 67, 68]. One prominent application involves biomedical relation extraction, which seeks to uncover intricate relationships between diverse biomedical entities, such as diseases, genes, species, and chemicals. The accuracy and quality of downstream relation extraction tasks are directly contingent on the performance of BioNER systems [69]. Furthermore, BioNER significantly contributes to the identification and classification of drugs and their interactions, critical for applications like drug discovery, drug safety monitoring, and personalized medicine [70]. It also facilitates knowledge base completion, enriching biomedical knowledge bases by automatically extracting and categorizing pertinent entities from text. This enriched knowledge is then harnessed for a multitude of purposes, including semantic

search, question answering, and data integration. BioNER's utility extends to biomedical question answering, assisting in understanding and responding to complex biomedical queries by identifying and categorizing relevant entities from both the question and the relevant text [71]. In the realm of biomedical text summarization, BioNER proves invaluable in identifying and extracting key biomedical entities from texts, thereby enabling the generation of concise, informative summaries [71]. Moreover, it enhances the accuracy and relevance of search results in biomedical information retrieval systems, where it identifies and classifies pertinent entities in both the query and the document collection [51, 67, 71, 72]. A simplified representation for identifying a Named entity can be represented as shown in Algorithm 1

***Algorithm 1**: Named Entity Recognition and Classification using BERT Model.*

Input: Tokens: A sequence of tokens representing the input text

Output: Named Entity Labels: A list of predicted entity labels for each token in the input text

```
function Named_Entity_Recognition_And_Classification(tokens):
  Initialize Named_Entity_Labels as an empty list
  for each token t in a sequence of tokens do
    Calculate the probability distribution over possible entity labels:
      P(label | token, context) = fmodel(token, context)
    Apply softmax function to P(label | token, context)
    Assign the label with the highest probability for token t
    Append the predicted entity label to Named_Entity_Labels
  end for
  return Named_Entity_Labels
end function
```

The actual calculation of these probabilities can involve complex machine learning models such as CRFs, BiLSTM-CRFs, or Transformer-based models, which take into account the token, its context, and various features for labelling.

In Algorithm 1:

- P(label | token, context) represents the probability of a specific label given the token and its context.
- fmodel() represents the function that the machine learning model (*e.g.*, CRF, BiLSTM, Transformer) uses to predict this probability.
- The model represents the machine learning model that has been trained to predict entity labels.

SciSpaCy, a Python library and models tailored for practical biomedical and scientific text processing, makes extensive use of spaCy [73]. spaCy itself employs deep learning techniques and, with the release of version 3, has transitioned to using the Transformer architecture as its deep learning model. It offers a customizable architecture that accommodates various deep learning techniques such as LSTMs, CRFs, and transformers to suit specific tasks [74].

BioBERT, on the other hand, is a pre-trained biomedical language representation model based on the BERT architecture [75]. It has been fine-tuned on an extensive corpus of biomedical literature, incorporating articles from PubMed and PMC as shown in Fig. (**1**). The adaptability of this model enables it to perform very well in many biomedical activities, such as biomedical named entity identification, connection extraction, question answering, and drug discovery. Both BioBERT and SciSpaCy are open-source tools, rendering them important resources for NER.

In this work, we present significant contributions to the domain of Biomedical Named Entity Recognition (BioNER), with prior focus on the utilization and examination of two prominent models, SciSpaCy and BioBERT, addressing the intricate challenges inherent in BioNER.

1. Through meticulous training on diverse biological datasets, we systematically evaluate the performance of SciSpaCy and BioBERT across various domains. Our comprehensive assessment employs key metrics such as F1 scores and processing speed, providing a detailed understanding of their efficacy in BioNER tasks.

2. The study offers valuable insights for the selection of tools tailored to specific BioNER requirements. By optimizing Named Entity Recognition in the field of biomedical research, our work aims to contribute to the broader advancement of this critical domain

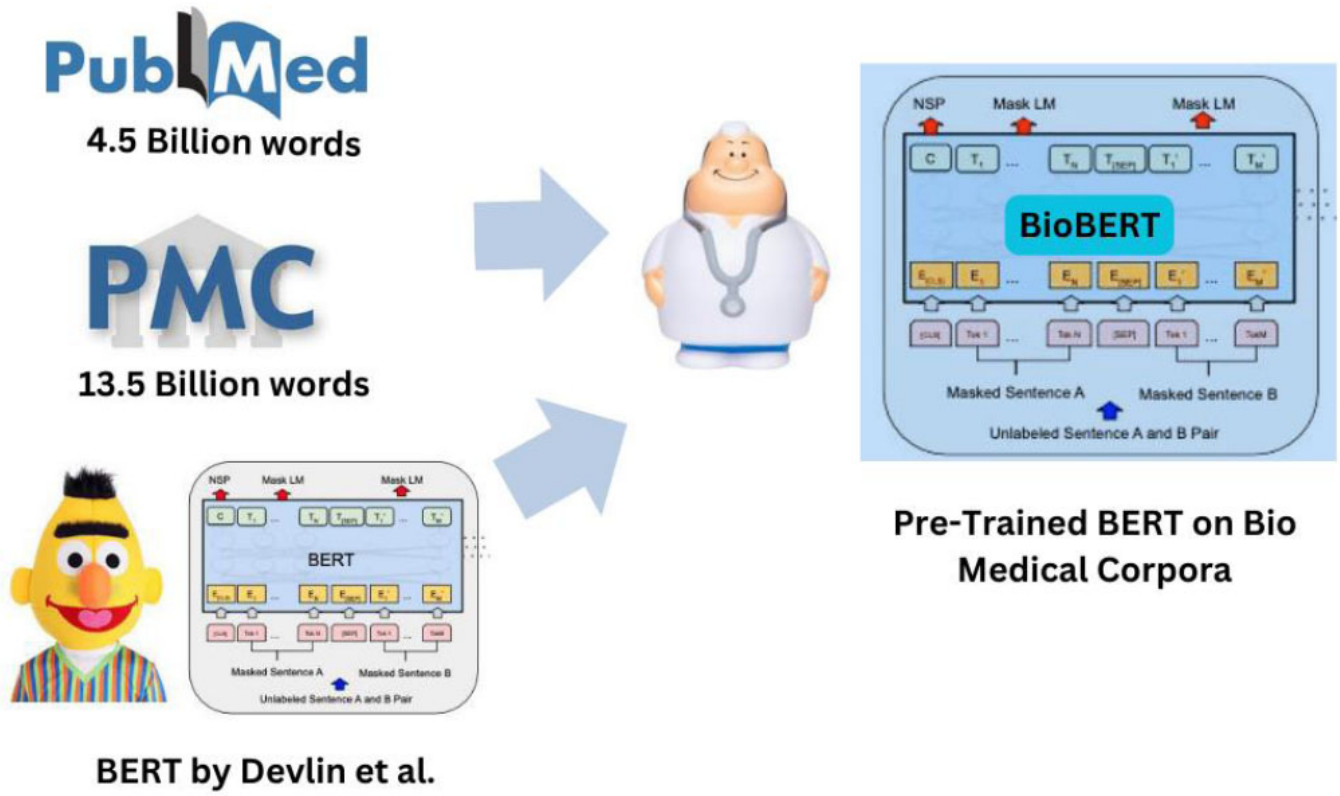

Fig. (1). BioBERT pretraining.

We have performed some literature surveys of the papers that collectively discuss various approaches to biomedical NER. Amith 2017 proposes an ontology-driven method that utilizes information extraction and features of ontologies to identify biomedical software names [76]. Al-Hegami 2017 focuses on the use of machine learning classifiers and a rich feature set to recognize biomedical entities, with the K-Nearest Neighbour classifier showing promising results [77]. Zhang 2013 presents an unsupervised approach to NER, using a noun phrase chunker and distributional semantics for entity extraction and classification [20].

Li 2023 proposes a fusion multi-features embedding method, combining deep contextual word-level features, local char-level features, and part-of-speech features to improve biomedical entity recognition [78]. Kaewphan *et al.*, 2018, published in the "Database Journal of Biological Databases Curation," shows a system for automatically identifying a wide range of biomedical entities within literature, boasting state-of-the-art performance. The study achieved the highest results in named entity recognition, underscoring its capability in biomedical entity recognition and normalization [79].

Priyanka *et al.*, 2021 provided a detailed survey of clinical Named Entity Recognition and Relationship Extraction (RE) techniques, addressing existing NLP models, performance, challenges, and future directions in the context of information extraction from clinical text, offering insights on current research and evaluation metrics [18].

Gong *et al.*, 2009 published in the "International Conference on BioMedical Engineering and Informatics," this study introduces a hybrid approach aimed at recognizing untagged biomedical entities. The approach demonstrates its utility in identifying biomedical entities in the GENIA 3.02 corpus, offering support for biologists in tagging biomedical entities within biomedical literature [80].

Kanimozhi & Manjula in 2018 performed a systematic review that focuses on the identification and classification of named entities within unstructured text documents. The goal is to extract valuable information from such texts by identifying and categorizing the named entities, a crucial step in extracting useful knowledge from unstructured textual data [81].

Overall, these papers highlight different techniques and algorithms for biomedical NER, showcasing the importance of effective natural language processing tools in organizing and extracting information from biomedical literature. Table **1** depicts the core methodology of the above-mentioned research works.

Table 1. NER tabular analysis.

S. No	Title	Authors	Year	Methodology
1.	"Knowledge-Based Approach for Named Entity Recognition in Biomedical Literature: A Use Case in Biomedical Software Identification"	Muhammad Amith, *et al*. [76]	2017	Ontology-driven method to identify familiar and unfamiliar software names.
2.	"A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set"	A. S. Al-Hegami, *et al*. [77]	2017	K-nearest neighbor trained with suitable features for recognizing biomedical named entities.
3.	"Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts"	Shaodian Zhang, *et al*. [20]	2013	Noun phrase chunker followed by a filter based on inverse document frequency for candidate entity extraction.
4.	"Biomedical named entity recognition based on fusion multi-features embedding"	Meijing Li, *et al*. [78]	2023	A proposed multi-feature embedding method with a positive effect on prediction results.
5.	"Wide-scope biomedical named entity recognition and normalization with CRFs, fuzzy matching and character level modelling"	S. Kaewphan, *et al*. [79]	2018	System for automatically identifying a multitude of biomedical entities from the literature with state-of-the-art performance.
6.	"A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts"	Priyanka *et al*. [18],	2021	Surveyed clinical Named Entity Recognition and Relationship Extraction (RE) techniques and addressed existing NLP models, performance, challenges.
7.	"A Hybrid Approach for Biomedical Entity Name Recognition"	Lejun Gong, *et al*. [80]	2009	A hybrid approach for recognizing untagged biomedical entities in GENIA 3.02 corpus.
8.	"A Systematic Review on Biomedical Named Entity Recognition"	U. Kanimozhi, *et al*. [81]	2018	Focus on identifying and classifying named entities for extracting useful information from unstructured text documents.

In addition to this study, there are several domains where Named Entity Recognition may be applied. Potential intersections and possibilities for collaboration or cross-pollination exist across the fields, especially in areas such as Visual Named Entity Recognition, cross-modal learning, semantic comprehension, and multi-modal information fusion.

1. Visual Named Entity Recognition (VNER): is a process that involves detecting and classifying named entities in text-based data. There is an increasing interest in Visual Named Entity Recognition, which refers to the process of recognising and categorising named items in photos or videos. The domain might possibly benefit from the application or adaptation of image processing techniques, particularly those that emphasise feature extraction and semantic interpretation.

2. Cross-Modal Learning: Cross-modal learning involves leveraging information from different modalities (*e.g.*, text and images) to improve performance on various tasks. Techniques developed for image processing, especially those involving attention mechanisms and feature extraction, could potentially be integrated with NER systems to improve entity recognition performance, especially in cases where textual context alone may not be sufficient.

3. Semantic Understanding: Both image processing and NER involve understanding the semantic content of data, albeit in different modalities. Techniques developed for image inpainting, super-resolution, or feature extraction may enhance the semantic understanding of textual data, which could indirectly benefit NER systems by providing richer contextual information.

4. Multi-Modal Information Fusion: Another potential linkage is through multi-modal information fusion techniques. Research in image processing often involves integrating information from different sources or modalities to improve the quality of image reconstruction or understanding. Similarly, integrating information from images (*e.g.*, scenes, objects) with textual data (*e.g.*, context, descriptions) could enhance NER systems' performance, especially in scenarios where textual and visual information are both available.

Some interesting works where NER can be combined are done by Yuantao Chen *et al*. in image processing techniques such as image inpainting and super-resolution. In a series of innovative contributions by Yuantao Chen, Runlong Xia, Kai Yang, and Ke Zou, their research endeavors have significantly advanced the field of image processing and restoration. In June 2023, they introduced "DARGS Image inpainting algorithm *via* deep attention residuals group and semantics," proposing a method that combines Semantic Priors Network, Deep Attention Residual Group, and Full-scale Skip Connection to effectively restore missing regions in images. Building on this foundation [82], in October 2023, they presented "GCAM: lightweight image inpainting *via* group convolution and attention mechanism," offering a lightweight approach that utilizes group convolution and rotating attention mechanism to enhance image inpainting while optimizing resource usage [83]. In January 2024, they unveiled "MFMAM: Image inpainting *via* multi-scale feature module with attention module," introducing a network utilizing multi-scale feature module and improved attention mechanisms for enhanced image inpainting, particularly focusing on texture and semantic detail preservation [84]. Their work continued with "DNNAM: Image inpainting algorithm *via* deep neural networks and attention mechanism," published in March 2024, where they introduced a method incorporating partial multi-scale channel attention mechanism and deep neural networks to improve image inpainting accuracy and quality [85]. Lastly, in July 2024, their paper titled "MICU: Image super-resolution *via* multi-level information compensation

and U-net" presented a novel approach for image super-resolution using multi-level information compensation and U-net architecture, significantly improving image reconstruction quality compared to existing methods [86]. Collectively, these papers underscore the author group's commitment to advancing image processing techniques with innovative methodologies and improved performance metrics. These works are very close in the possibilities for collaborating NER in areas such as Visual Named Entity Recognition, cross-modal learning, semantic comprehension, and multi-modal information fusion.

2. METHODOLOGY

The work aimed to rigorously evaluate the performance of BioBERT and SCISpaCy in biomedical Named Entity Recognition, considering the data's origin, preprocessing, training, and evaluation processes. The choice of these models and the training process was guided by the aim of achieving accurate and generalizable recognition of biomedical entities. These considerations are explained as follows:

2.1. Data Collection

The dataset utilized in this study was curated from diverse sources of biomedical Named Entity Recognition datasets. These datasets, including JNLPBA [15], BC4_CHEMD [87], BIONLP13CG [88], EXPERT 206 [70], and BC5CDR [89], were selected to encompass a wide range of biomedical entity types and domains, offering a comprehensive evaluation of BioNER models as shown in Table 2.

Entity Types and Datasets: The evaluation encompassed multiple datasets, each representing distinct biomedical entity types. The datasets included.

- JNLPBA: Targeted entities included DNA, CELL_TYPE, CELL_LINE, RNA, and PROTEIN.
- BC4_CHEMD: Focused on the CHEMICAL entity type.
- BIONLP13CG: Encompassed a diverse set of entity types, such as AMINO_ACID, ANATOMICAL_SYSTEM, CANCER, and more.
- EXPERT 206: Included GENE, DISEASE, and VARIANT as recognized entities.
- BC5CDR: Considered both CHEMICAL and DISEASE entities.

2.2. Data Preprocessing

2.2.1. Data Conversion for BioBERT

To prepare the data for training BioBERT, the datasets were converted into a BIO (Beginning-Inside-Outside) format. This involved annotating each token in the dataset with one of three labels: "B-" for the beginning of an entity, "I-" for internal parts of an entity, and "O" for tokens outside of any entity. This format enables BioBERT to understand and recognize the boundaries of biomedical entities within the text.

2.2.2. Data Conversion for SCISpaCy

For training SCISpaCy, the datasets were transformed into a format suitable for spaCy's NER model. The conversion process required structuring the data as a list of sentences, each followed by a list of (start, end, label) triples, where the triples define the start and end positions of the entity within the sentence and its associated label. For example, ("Tokyo Tower is 333m tall.", [(0, 11, "BUILDING")]) represents a sentence, the start and end positions of the entity "BUILDING," and its label.

2.3. Data Splitting

To facilitate model evaluation, the dataset was divided into training and testing subsets. The data was partitioned using an 80-20 split, allocating 80% of the data for training and 20% for testing. This ensures that the models are trained on a substantial portion of the data while preserving an independent dataset for assessing their generalization performance.

2.4. Hardware and Software Environment

Training of the model was carried out on carefully managed hardware consisting of Google Collab T4 GPUs. The environment had the essential programming languages, libraries, and frameworks for the construction and execution of the models, as well as particular GPU specifications for rapid training.

Table 2. Dataset analysis.

Datasets	Entity Type	No. of Entities
BIONLP13CG	Tissues	587
BIONLP13CG	Organisms	2,093
BIONLP13CG	Gene/Protein	7,908
EXPERT 206	Disease	7,390
EXPERT 206	Gene/Protein	7,393
BC5CDR	Disease	12,204
BC5CDR	Chemical	12,204
BC4_CHEMD	Chemical	84,249
JNLPBA	Cell Line	4,315

Table 3. NER results of SciSpacy and BioBERT.

Model	Entity Types	SCISpaCy F1 Score	BioBERT F1 Score
JNLPBA	"DNA, CELL_TYPE, CELL_LINE, RNA, PROTEIN"	73.1	73.67
BC4_CHEMD	"CHEMICAL"	85.1	86.07
BIONLP13CG	"AMINO_ACID, ANATOMICAL_SYSTEM, CANCER, CELL, CELLULAR_COMPONENT, DEVELOPING_ANATOMICAL_STRUCTURE, GENE_OR_GENE_PRODUCT, IMMATERIAL_ANATOMICAL_ENTITY, MULTI-TISSUE_STRUCTURE, ORGAN, ORGANISM, ORGANISM_SUBDIVISION, ORGANISM_SUBSTANCE, PATHOLOGICAL_FORMATION, SIMPLE_CHEMICAL, TISSUE"	78.13	86.07
EXPERT 206	"GENE, DISEASE, VARIANT"	91.08	91.23
BC5CDR	"CHEMICAL, DISEASE"	85.53	87.83

3. RESULTS AND DISCUSSION

3.1. Model Performance

3.1.1. JNLPBA Dataset

For the JNLPBA dataset, SCISpaCy and BioBERT both displayed competitive performance, with F1 scores of 73.1 and 73.67, respectively. These ratings were uniform across all entity types in this sample.

In the BC4 CHEMD dataset, which concentrates on CHEMICAL items, both models demonstrated excellent performance. BioBERT surpassed SCISpaCy with an F1 score of 86.07, while SCISpaCy received an F1 score of 85.1.

3.1.2. BIONLP13CG Dataset

The BIONLP13CG dataset featured a large variety of entity types, posing a challenge to the models' ability to recognise a wide range of biomedical items. BioBERT's F1 score of 86.07 was superior to SCISpaCy's score of 78.13.

The EXPERT 206 dataset, which consists of GENE, DISEASE, and VARIANT entities, produced high F1 scores for both models. SCISpaCy obtained an F1 value of 91.08, but BioBERT fared somewhat better with an F1 score of 91.23.

3.1.3. BC5CDR Dataset

For the BC5CDR dataset, which involves CHEMICAL and DISEASE entity recognition, both models displayed strong performance. SCISpaCy achieved an F1 score of 85.53, while BioBERT exhibited superior performance with an F1 score of 87.83. These results are showcased in Table **3**.

3.2. Inference and Performance Trade-offs

The evaluation of inference times revealed that SCISpaCy processed an average sentence in approximately 90 milliseconds, while BioBERT required an average of 278 milliseconds per sentence as shown in Fig. (**2**). It is important to note that the relatively longer inference time of BioBERT can be attributed to various internal tasks, including tokenizer loading and token classification.

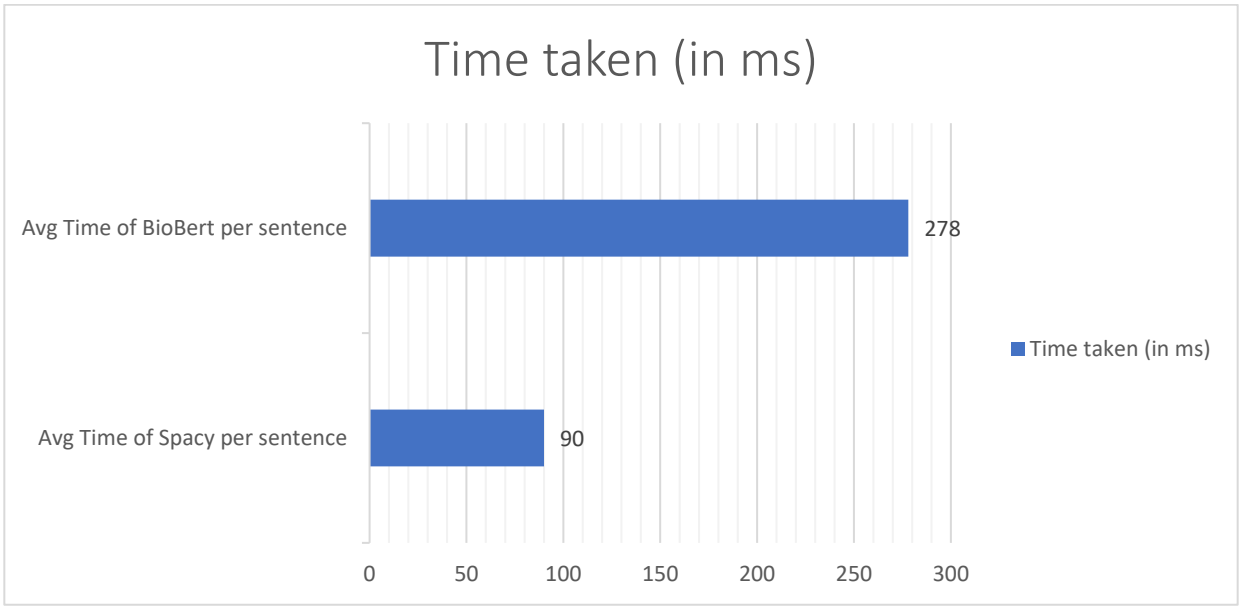

Fig. (2). Time complexity analysis of NER models.

However, BioBERT compensates for this by efficiently processing multiple sentences in parallel. This capability enhances throughput, making BioBERT a favorable choice in scenarios where parallelism is crucial.

4. DISCUSSION

Based on the results showcased in Table **3**. Some inferences have been deduced. These inferences can help to guide the selection of the appropriate model for specific biomedical NER tasks based on the dataset's entity types and performance requirements.

4.1. Entity Types and Specialization

The F1 scores vary across different entity types and datasets. For instance, in the JNLPBA dataset, DNA, CELL_TYPE, CELL_LINE, RNA, and PROTEIN have relatively similar F1 scores, indicating that the model's performance is consistent across these biomedical entity types. On the other hand, in the BIONLP13CG dataset, there is a wide range of entity types with varying F1 scores, suggesting that the model's performance is more specialized for some entity types.

4.2. BioBERT's Consistency

BioBERT consistently achieves higher F1 scores compared to SCISpaCy across all datasets and entity types. This indicates BioBERT's robustness and effectiveness in biomedical named entity recognition tasks.

4.3. High Performance for EXPERT 206

The EXPERT 206 dataset demonstrates the highest F1 scores for both models. This suggests that the entities in this dataset, such as Genes, Diseases, and variants, are well-recognized by both SCISpaCy and BioBERT.

4.4. Strong BioBERT Performance in BC5CDR

In the BC5CDR dataset, which focuses on Chemicals, BioBERT significantly outperforms SCISpaCy with an F1 score of 87.83, indicating its strength in recognizing chemical entities.

4.5. Balanced Performance in BC4_CHEMD

The BC4_CHEMD dataset, consisting of DISEASE and CHEMICAL entity types, shows similar and high F1 scores for both models. This suggests that both SCISpaCy and BioBERT are proficient in recognizing these types of entities.

4.6. Biomedical Specialization

The choice of model may depend on the specific biomedical entity types in the dataset. While BioBERT generally outperforms SCISpaCy, practitioners should consider the nature of the entities they are dealing with when selecting the most suitable model for their task.

4.7. Overall Robustness

Both SCISpaCy and BioBERT demonstrate strong performance in biomedical named entity recognition. However, BioBERT's consistently higher F1 scores and adaptability to various entity types make it a valuable choice for a wide range of biomedical applications.

4.8. Time Complexity on Same Hardware

On the same hardware configuration with a T4 GPU (utilizing Google Colab's default GPU settings), we observed that spaCy demonstrated a faster average inference time per sentence, approximately 50 milliseconds, compared to BioBERT's average speed of 178 milliseconds per sentence. It is worth noting that the apparent time difference in favor of spaCy may be attributed to its streamlined processing for individual sentences. However, it is important to consider that BioBERT exhibits a significant advantage when processing multiple sentences in parallel. This capability compensates for the additional time incurred by BioBERT in various internal tasks such as tokenizer loading and token classification. Consequently, BioBERT's suitability for parallel sentence processing makes it a strong contender in scenarios where high throughput and parallelism are critical.

CONCLUSION

SCISpaCy and BioBERT, along with other BioNER models, have demonstrated their capabilities, benefits, and relevant considerations. This study evaluated diverse datasets, and performance metrics, offering valuable insights to scholars and practitioners in the field of biological natural language processing. The study produced intriguing findings. The selection of a model such as SCISpaCy or BioBERT significantly impacts the detection of many biological events. BioBERT consistently achieved higher F1 scores than SCISpaCy across several datasets, entity types, and domains. The durability and adaptability of BioBERT in the biomedical field make it a highly promising choice for BioNER.The datasets exhibited notable variations in F1 scores across different item categories. The F1 ratings for both models in the EXPERT 206 dataset of Genes, Diseases, and Variants indicate that SCISpaCy and BioBERT effectively recognise and categorise these entity types. The BIONLP13CG dataset exhibited a diverse range of entity kinds, each with distinct F1 scores. The performance of the model may be tailored to specific entities. Parallel processing is an additional advantage of BioBERT. Although BioBERT may have longer phrase inference times, it has the ability to process several sentences simultaneously, which makes it an excellent choice for high-throughput applications. It is essential to assess the different categories within the dataset and determine the optimal trade-off between efficiency and precision when recognising biological entities. The selection of the model should be guided by the intrinsic aspects of the work. Our analysis demonstrates the significance of BioNER in the field of biomedical research and the capabilities of current models. Before making a decision, it is important to thoroughly comprehend the specific criteria and objectives of each biological Named Entity Recognition task, and then select either SCISpaCy, BioBERT, or a combination of both accordingly. The advancement of NLP models in the field of biomedicine provides professionals with a diverse

array of tools to effectively extract pertinent information from texts related to biology.

LIMITATIONS, FUTURE WORK AND RESEARCH DIRECTIONS

To improve the thoroughness of our assessments, it is crucial to further investigate domain-specific algorithms for Named Entity Recognition in the context of future studies. In order to achieve this objective, we suggest a deliberate enlargement of our study scope to encompass notable algorithms such as Bert-PKD, CollaboNet, TinyBert, BERN2, and other state-of-the-art techniques. Integrating these algorithms into our future endeavours will enhance the strength and comprehensiveness of our evaluation system. The limitations in the field of Biomedical Named entity recognition include data availability, expert annotation requirements, and the vast space of biomedical concepts. Our objective is to analyse a wider range of NER algorithms in a systematic manner, in order to reveal subtle variations in performance and find methods that may have distinct benefits in particular biological scenarios. This strategic growth is in line with our dedication to developing the area of Biomedical Named Entity Recognition and guaranteeing that our research offers practical insights for researchers, practitioners, and professionals in the biomedical sector. Integrating new algorithms will enhance the comprehensiveness of our studies and promote a more nuanced comprehension of the advantages and constraints of different techniques. To summarise, more study on domain-specific algorithms for NER is crucial. There are potential intersections and opportunities for collaboration or cross-pollination between the fields, particularly in areas such as Visual Named Entity Recognition, cross-modal learning, semantic understanding, and multi-modal information fusion. We are committed to expanding the reach and significance of our findings in future studies.

LIST OF ABBREVIATIONS

BioNER	=	Biological Named Entity Recognition
NER	=	Named Entity Recognition
IE	=	Information Extraction
NLP	=	Natural Language Processing
CRF	=	Conditional Random Fields

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIAL

All the data and supporting information is provided within the article.

FUNDING

None.

CONFLICT OF INTEREST

The authors declared no conflict of interest financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, and Z. Lu, "BioRED: A rich biomedical relation extraction dataset", *Brief. Bioinform.,* vol. 2022, no. 5, pp. 1-12, 2022.

[2] Q. Xi, Y. Ren, S. Yao, G. Wu, G. Miao, and Z. Zhang, "Chinese named entity recognition", *Applicat. Challeng.,* vol. 12647, pp. 51-81, 2021. [http://dx.doi.org/10.1007/978-3-030-71590-8_4]

[3] A. Ahmed, A. Abbasi, and C. Eickhoff, "Benchmarking modern named entity recognition techniques for free-text health record deidentification", *AMIA Jt Summits Transl Sci Proc,* vol. 2021, pp. 102-111, 2021.

[4] R. Hema, and A. Devi, "Chemical named entity recognition using deep learning techniques", *Deep Natural Language Processing and AI Applications for Industry* pp.59-73, 2021. [http://dx.doi.org/10.4018/978-1-7998-7728-8.ch004]

[5] K. Dawar, A.J. Samuel, and R. Alvarado, "Comparing topic modeling and named entity recognition techniques for the semantic indexing of a landscape architecture textbook", *2019 Syst. Inf. Eng. Des. Symp.,* 2019pp. 1-6

[6] S. Tedeschi, and R. Navigli, "MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and Disambiguation)", *Findings of the Association for Computational Linguistics: NAACL* Seattle, United States pp.801-812, 2022.

[7] A. Groza, "Detecting fake news for the new coronavirus by reasoning on the Covid-19 ontology", *arXiv:2004.12330,* 2020.

[8] H. Liu, Z. Sun, and F. Ning, "Named entity recognition method for cnc machine tool design knowledge text", *2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)* 11-12 December 2022, Dalian, China, pp.1452-1455, 2022.

[9] S.R. Kundeti, J. Vijayananda, S. Mujjiga, and M. Kalyan, "Clinical named entity recognition: Challenges and opportunities", *2016 IEEE International Conference on Big Data (Big Data)* 05-08 December 2016, Washington, DC, USA, pp.1937-1945, 2016.

[10] M.C. Durango, T.E.A. Silva, and O.A. Duque, "Named entity recognition in electronic health records: A methodological review", *Healthc. Inform. Res.,* vol. 29, no. 4, pp. 286-300, 2023. [http://dx.doi.org/10.4258/HIR.2023.29.4.286] [PMID: 37964451]

[11] N. Perera, M. Dehmer, and F. Emmert-Streib, "Named entity recognition and relation detection for biomedical information extraction", *Front Cell Dev Biol,* vol. 8, p. 673, 2020.

[12] M. Ehrmann, A. Hamdi, E.L. Pontes, M. Romanello, and A. Doucet, "Named entity recognition and classification in historical documents: A survey", *ACM Comput. Surv.,* vol. 56, no. 2, 2023. [http://dx.doi.org/10.1145/3604931]

[13] H. Cho, and H. Lee, "Biomedical named entity recognition using deep neural networks with contextual information", *BMC Bioinformat.,* vol. 20, no. 1, p. 735, 2019. [http://dx.doi.org/10.1186/S12859-019-3321-4/FIGURES/1] [PMID: 31881938]

[14] V. Kocaman, and D. Talby, "Accurate clinical and biomedical named entity recognition at scale", *Softw. Impacts,* vol. 13, p. 100373, 2022. [http://dx.doi.org/10.1016/J.SIMPA.2022.100373]

[15] M-S. Huang, P-T. Lai, R.T-H. Tsai, and W-L. Hsu, "Revised JNLPBA corpus: A revised version of biomedical NER corpus for relation extraction task", *Brief. Bioinform.,* vol. 21, no. 6, pp. 2219-2238, 2020. [http://dx.doi.org/10.1093/bib/bbaa054] [PMID: 32602538]

[16] K. Wang, "NERO: A biomedical named-entity (recognition) ontology with a large, annotated corpus reveals meaningful associations through text embedding", *NPJ Syst Biol Appl,* vol. 7, p. 38, 2021.

[17] A. Goyal, V. Gupta, and M. Kumar, "Recent named entity recognition and classification techniques: A systematic review",

Comput. Sci. Rev., vol. 29, pp. 21-43, 2018.
[http://dx.doi.org/10.1016/J.COSREV.2018.06.001]

[18] P. Bose, S. Srinivasan, W.C. Sleeman, J. Palta, R. Kapoor, and P. Ghosh, "A survey on recent named entity recognition and relationship extraction techniques on clinical texts", *Appl. Sci.,* vol. 11, no. 18, p. 8319, 2021.
[http://dx.doi.org/10.3390/app11188319]

[19] V. Moscato, M. Postiglione, and G. Sperlí, "Few-shot named entity recognition: Definition, taxonomy and research directions", *ACM Trans. Intell. Syst. Technol.,* vol. 14, no. 5, pp. 1-46, 2023.
[http://dx.doi.org/10.1145/3609483]

[20] S. Zhang, and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts", *J. Biomed. Inform.,* vol. 46, no. 6, pp. 1088-1098, 2013.
[PMID: 23954592]

[21] Y. Tian, W. Shen, Y. Song, F. Xia, M. He, and K. Li, "Improving biomedical named entity recognition with syntactic information", *BMC Bioinformat.,* vol. 21, no. 1, p. 539, 2020.
[http://dx.doi.org/10.1186/s12859-020-03834-6] [PMID: 33238875]

[22] J. Kocerka, M. Krześlak, and A. Gałuszka, "Ontology extraction from software requirements using named-entity recognition", *Adv. Sci. Technol. Res. J.,* vol. 16, no. 3, pp. 207-212, 2022.
[http://dx.doi.org/10.12913/22998624/149941]

[23] G.L. Ciampaglia, P. Shiralkar, L.M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks", *PLoS One,* vol. 10, no. 6, p. e0128193, 2015.
[http://dx.doi.org/10.1371/JOURNAL.PONE.0128193] [PMID: 26083336]

[24] R. Meyes, M. Lu, C.W. de Puiseau, and T. Meisen, "Ablation studies in artificial neural networks", *arXiv:1901.08644,* 2019.

[25] Y. Xiong, "Improving deep learning method for biomedical named entity recognition by using entity definition information", *BMC Bioinformat.,* vol. 22, no. S1, p. 600, 2021.

[26] A. Goyal, M. Kumar, and V. Gupta, "Named entity recognition: Applications, approaches and challenges", *IJARCCE,* vol. 6, no. 2, pp. 259-262, 2017.

[27] L. Chang, R. Zhang, J. Lv, W. Zhou, and Y. Bai, "A review of biomedical named entity recognition", *J. Comput. Methods Sci. Eng.,* vol. 22, no. 3, pp. 893-900, 2022.
[http://dx.doi.org/10.3233/JCM-225952]

[28] P. Kalamkar, A. Agarwal, A. Tiwari, S. Gupta, S. Karn, and V. Raghavan, "Named entity recognition in indian court judgments", *arXiv:2211.03442,* 2022.

[29] B. Yaman, M. Pasin, and M. Freudenberg, "Interlinking SciGraph and DBpedia datasets using link discovery and named entity recognition techniques", *Open Access Ser. Informat. (OASIcs),* vol. 70, pp. 15:1-15:8, 2019.

[30] L.S. Cocca, B. Maier, C. Nawroth, P. Kevitt, and M. Hemmje, "Named entity recognition for the extraction of emerging technological knowledge from medical literature", *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management,* 2022pp. 101-108

[31] M. Liu, Z. Tu, T. Zhang, T. Su, X. Xu, and Z. Wang, "LTP: A new active learning strategy for CRF-based named entity recognition", *Neural Process. Lett.,* vol. 54, no. 3, pp. 2433-2454, 2022.
[http://dx.doi.org/10.1007/s11063-021-10737-x]

[32] H. Luo, and B. Gao, "Development of smart wearable sensors for life healthcare", *Eng. Regen.,* vol. 2, pp. 163-170, 2021.
[http://dx.doi.org/10.1016/J.ENGREG.2021.10.001]

[33] Z. Li, S. Zhang, Y. Song, and J. Park, "Extrinsic factors affecting the accuracy of biomedical NER", *arXiv:2305.18152,* 2023.

[34] N. Garg, "Research Proposal Research Question " Can Named Entities improve", Available from: https://www.semanticscholar.org/paper/Research-Proposal-Research-Question-

[35] J. Barua, and D. Patel, "Named entity classification using search engine's query suggestions", *European Conference on Information Retrieval,* vol. 2017, 2017.

[36] C. Sabty, I. Omar, F. Wasfalla, M. Islam, and S. Abdennadher, "Data augmentation techniques on arabic data for named entity recognition", *Procedia Comput. Sci.,* vol. 189, pp. 292-299, 2021.
[http://dx.doi.org/10.1016/j.procs.2021.05.092]

[37] M.H. Khanam, M.A. Khudhus, and M.S.P. Babu, "Named entity recognition using machine learning techniques for telugu language", *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)* 26-28 August 2016, Beijing, China, 2016.

[38] W. Shishah, "Fake news detection using BERT model with joint learning", *Arab. J. Sci. Eng.,* vol. 46, pp. 9115-9127, 2021.
[http://dx.doi.org/10.1007/s13369-021-05780-8]

[39] J. Kalluru, "Enhancing data accuracy and efficiency: An overview of fuzzy matching techniques", *Int. J. Sci. Res.,* vol. 12, no. 8, pp. 685-690, 2023.

[40] E. Tokarchuk, D. Thulke, W. Wang, C. Dugast, and H. Ney, "Investigation on data adaptation techniques for neural named entity recognition", *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop* pp.1-15, 2021.

[41] G. S. Mahalakshmi, and A.A. L. Adline, "Boosting formal concept analysis based definition extraction via named entity recognition", In: *Smart Innovation, Systems and Technologies,* Springer: New Delhi, 2015.L. Jain, H. Behera, J. Mandal, and D. Mohapatra,

[42] T. Mehmood, I. Serina, A. Lavelli, L. Putelli, and A. Gerevini, "On the use of knowledge transfer techniques for biomedical named entity recognition", *Future Internet,* vol. 15, no. 2, p. 79, 2023.
[http://dx.doi.org/10.3390/fi15020079]

[43] L. Stepanyan, "Automated custom named entity recognition and disambiguation", Available from: https://api.semanticscholar.org/CorpusID:219618671

[44] M. Pajas, A. Radovan, and I.O. Biškupić, "Multilingual named entity recognition solution for optimizing parcel delivery in online commerce: Identifying person and organization names", *2023 46th MIPRO ICT and Electronics Convention (MIPRO)* 22-26 May 2023, Opatija, Croatia, pp.1119-1124, 2023.
[http://dx.doi.org/10.23919/MIPRO57284.2023.10159789]

[45] C. Jilek, M. Schröder, R. Novik, S. Schwarz, H. Maus, and A. Dengel, "Inflection-tolerant ontology-based named entity recognition for real-time applications", *arXiv:1812.02119,* 2019.

[46] X. Jiang, and T. Ouyang, "End-to-end speech to named entity recognition system", Available from: https://www.semanticscholar.org/paper/End-to-End-Speech-to-Named-Entity-Recognition-Jiang-Ouyang/b2c43c47151c339c92b8e1be629c152626afb90a

[47] Q. Wei, Z. Ji, Z. Li, J. Du, J. Wang, J. Xu, Y. Xiang, F. Tiryaki, S. Wu, Y. Zhang, C. Tao, and H. Xu, "A study of deep learning approaches for medication and adverse drug event extraction from clinical text", *J. Am. Med. Inform. Assoc.,* vol. 27, no. 1, pp. 13-21, 2020.
[http://dx.doi.org/10.1093/jamia/ocz063] [PMID: 31135882]

[48] Arakelyan, "Automated custom named entity recognition and disambiguation", Available from: https://www.semanticscholar.org/paper/Automated-Custom-Named-Entity-Recognition-and-Stepanyan/24d73ef1aa9f2fb7c1651be67b4f3e40b55ff31e

[49] B. Powley, and R. Dale, "High accuracy citation extraction and named entity recognition for a heterogeneous corpus of academic papers", Available from: https://www.semanticscholar.org/paper/High-accuracy-citation-extraction-and-named-entity-Powley-Dale/e478b6069a7fe2dae89673553edad449ec6c329b

[50] G. Alfattni, M. Belousov, N. Peek, and G. Nenadic, "Extracting drug names and associated attributes from discharge summaries: Text mining study", *JMIR Med. Inform.,* vol. 9, no. 5, p. e24678, 2021.
[http://dx.doi.org/10.2196/24678] [PMID: 33949962]

[51] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, "Biomedical named entity recognition using BERT in the machine

reading comprehension framework", *J. Biomed. Inform.,* vol. 118, p. 103799, 2021.
[http://dx.doi.org/10.1016/j.jbi.2021.103799] [PMID: 33965638]

[52] T. Zhang, "BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection", *2020 International Joint Conference on Neural Networks (IJCNN)* 19-24 July 2020, Glasgow, UK, 2020.

[53] T.M. Luu, R. Phan, R. Davey, and G. Chetty, "A multilevel NER framework for automatic clinical name entity recognition", *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* 18-21 November 2017, New Orleans, LA, USA, pp.1134-1143, 2017.

[54] U. Qazi, M. Imran, and F. Ofli, "GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information", *arXiv:2005.11177,* 2020.

[55] S. Raza, D.J. Reji, F. Shajan, and S.R. Bashir, "Large-scale application of named entity recognition to biomedicine and epidemiology", *PLOS Digit. Heal.,* vol. 1, no. 12, p. e0000152, 2022.
[http://dx.doi.org/10.1371/journal.pdig.0000152] [PMID: 36812589]

[56] H. Gobbi, and M. De Brot, "Papillary tumors of the breast", In: S. Stolnicu, I. Alvarado-Cabrero, Eds., *Practical Atlas of Breast Pathology,* Springer: Cham, 2018.

[57] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news", *Proceedings of the 27th International Conference on Computational Linguistics* Santa Fe, New Mexico, USA, pp.3391–3401, 2018.

[58] M. Linzer, E.H. Yang, N.A. Estes III, P. Wang, V.R. Vorperian, and W.N. Kapoor, "Diagnosing syncope. Part 1: Value of history, physical examination, and electrocardiography", *Ann. Intern. Med.,* vol. 126, no. 12, pp. 989-996, 1997.
[http://dx.doi.org/10.7326/0003-4819-126-12-199706150-00012] [PMID: 9182479]

[59] N.S. Pagad, and N. Pradeep, "Clinical named entity recognition methods: An overview", In: *Advances in Intelligent Systems and Computing,* vol. 1388. Springer: Singapore, 2022, pp. 151-165.
[http://dx.doi.org/10.1007/978-981-16-2597-8_13]

[60] S. Raza, and B. Schwartz, "Detecting biomedical named entities in COVID-19 texts", Available from: https://www.semanticscholar.org/paper/Detecting-Biomedical-Named-Entities-in-COVID-19-Raza-Schwartz/d38a9d72a347c404f69097be57bec6967b2e5bb1

[61] S.I. Bedmar, C.D. Perdonas, and G.S. Aspizua, "Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts", *BMC Bioinformat.,* vol. 23, no. 1, p. 263, 2021.

[62] H-J. Dai, C-H. Wei, H-Y. Kao, R-L. Liu, R.T-H. Tsai, and Z. Lu, "Text mining for translational bioinformatics", *BioMed Res. Int.,* vol. 2015, p. 368264, 2015.
[http://dx.doi.org/10.1155/2015/368264] [PMID: 26380272]

[63] L-H. Lee, C-Y. Chen, L-C. Yu, and Y-H. Tseng, "Overview of the ROCLING 2022 shared task for chinese healthcare named entity recognition", Available from: https://www.semanticscholar.org/paper/Overview-of-the-ROCLING-2022-Shared-Task-for-Named-Lee-Chen/0f3b2ac0d7aa2c5d4c50b5bd066b94abbf53d0f1

[64] B. Tang, D. Jiang, Q. Chen, X. Wang, J. Yan, and Y. Shen, "De-identification of clinical text via Bi-LSTM-CRF with neural language models", *AMIA Annu Symp Proc.,* vol. 2019, pp. 857-863, 2019.

[65] C. Xarhoulacos, A. Anagnostopoulou, G. Stergiopoulos, and D. Gritzalis, "Misinformation vs. situational awareness : The art of deception and the need for cross-domain detection", *Sensors,* vol. 21, no. 16, p. 5496, 2021.

[66] A. Radford, "Better language models and their implications", Available from: https://openai.com/blog/better-language-models/

[67] H. Kim, and J. Kang, "How do your biomedical named entity recognition models generalize to novel entities?", *IEEE Access,* vol. 10, pp. 31513-31523, 2022.
[http://dx.doi.org/10.1109/ACCESS.2022.3157854] [PMID: 35582496]

[68] M. Alzantot, Y. Sharma, A. Elgohary, B-J. Ho, M.B. Srivastava, and K-W. Chang, "Generating natural language adversarial examples", Available from: https://github.com/nesl/nlp_

[69] Z. Zhang, and A.L.P. Chen, "Biomedical named entity recognition with the combined feature attention and fully-shared multi-task learning", *BMC Bioinformatics,* vol. 23, no. 1, p. 458, 2022.
[http://dx.doi.org/10.1186/s12859-022-04994-3] [PMID: 36329384]

[70] P-H. Li, T.F. Chen, J.Y. Yu, S.H. Shih, C.H. Su, Y.H. Lin, H.K. Tsai, H.F. Juan, C.Y. Chen, and J.H. Huang, "pubmedKB: An interactive web server for exploring biomedical entity relations in the biomedical literature", *Nucleic Acids Res.,* vol. 50, no. W1, pp. W616-W622, 2022.
[http://dx.doi.org/10.1093/nar/gkac310] [PMID: 35536289]

[71] S. Jansen, "Who's who and what's what: Advances in biomedical named entity recognition (BioNER)", Available from: https://towardsdatascience.com/whos-who-and-what-s-what-advances-in-biomedical-named-entity-recognition-bioner-c42a3f63334c

[72] L. Luo, C-H. Wei, P-T. Lai, R. Leaman, Q. Chen, and Z. Lu, "AIONER: All-in-one scheme-based biomedical named entity recognition using deep learning", *Bioinformatics,* vol. 39, no. 5, p. btad310, 2023.
[http://dx.doi.org/10.1093/bioinformatics/btad310] [PMID: 37171899]

[73] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and robust models for biomedical natural language processing", *arXiv:1902.07669,* 2019.

[74] "Industrial-strength natural language processing", Available from: https://spacy.io/

[75] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics,* vol. 36, no. 4, pp. 1234-1240, 2020.
[http://dx.doi.org/10.1093/bioinformatics/btz682] [PMID: 31501885]

[76] M. Amith, Y. Zhang, H. Xu, and C. Tao, "Knowledge-based approach for named entity recognition in biomedical literature: A use case in biomedical software identification", In: S. Benferhat, K. Tabia, M. Ali, Eds., *Advances in Artificial Intelligence: From Theory to Practice,* vol. Vol. 10351. Springer International Publishing: Cham, 2017, pp. 386-395.

[77] A.S. Al-Hegami, A.M.F. Othman, and F.T. Bagash, "A biomedical named entity recognition using machine learning classifiers and rich feature set", *Int. J. Comput. Sci. Netw. Secur.,* vol. 17, no. 1, p. 170, 2017.

[78] M. Li, H. Yang, and Y. Liu, "Biomedical named entity recognition based on fusion multi-features embedding", *Technol Health Care.,* vol. 31, no. S1, pp. 111-121, 2023.

[79] S. Kaewphan, K. Hakala, N. Miekka, T. Salakoski, and F. Ginter, "Wide-scope biomedical named entity recognition and normalization with CRFs, fuzzy matching and character level modeling", *Database,* vol. 2018, pp. 1-10, 2018.
[PMID: 30239666]

[80] L-J. Gong, Y. Yuan, Y-B. Wei, and X. Sun, "A hybrid approach for biomedical entity name recognition", *2009 2nd International Conference on Biomedical Engineering and Informatics* 17-19 October 2009, Tianjin, China, 2009.

[81] U. Kanimozhi, and D. Manjula, "A systematic review on biomedical named entity recognition", In: *Data Science Analytics and Applications,* Springer, 2018, pp. 19-37.

[82] Y. Chen, R. Xia, K. Yang, and K. Zou, "DARGS: Image inpainting algorithm via deep attention residuals group and semantics", *J. King Saud Univ. Comput. Inf. Sci.,* vol. 35, no. 6, p. 101567, 2023.
[http://dx.doi.org/10.1016/J.JKSUCI.2023.101567]

[83] Y. Chen, R. Xia, K. Yang, and K. Zou, "GCAM: Lightweight image inpainting via group convolution and attention mechanism", *Int. J. Mach. Learn. Cybern.,* no. Oct, pp. 1-11, 2023.
[http://dx.doi.org/10.1007/S13042-023-01999-Z/METRICS] [PMID:

37360881]
[84] Y. Chen, R. Xia, K. Yang, and K. Zou, "MFMAM: Image inpainting via multi-scale feature module with attention module", *Comput. Vis. Image Underst.,* vol. 238, p. 103883, 2024. [http://dx.doi.org/10.1016/J.CVIU.2023.103883]
[85] Y. Chen, R. Xia, K. Yang, and K. Zou, "DNNAM: Image inpainting algorithm via deep neural networks and attention mechanism", *Appl. Soft Comput.,* vol. 154, p. 111392, 2024. [http://dx.doi.org/10.1016/J.ASOC.2024.111392]
[86] Y. Chen, R. Xia, K. Yang, and K. Zou, "MICU: Image super-resolution via multi-level information compensation and U-net", *Expert Syst. Appl.,* vol. 245, p. 123111, 2024. [http://dx.doi.org/10.1016/J.ESWA.2023.123111]
[87] "Named Entity Recognition (NER) on BC4CHEMD", Available from: https://paperswithcode.com/sota/named-entity- recognition-on-bc4chemd
[88] AI Datasets. Available from: https://www.ncbi.nlm.nih.gov/research/bionlp/Data/
[89] "BC5CDR (BioCreative V CDR corpus)", Available from: https://paperswithcode.com/dataset/bc5cdr