Hindawi
Complexity
Volume 2021, Article ID 6633213, 6 pages
https://doi.org/10.1155/2021/6633213

*Research Article*

# ABioNER: A BERT-Based Model for Arabic Biomedical Named-Entity Recognition

**Nada Boudjellal**,[1] **Huaping Zhang**,[1] **Asif Khan**,[1] **Arshad Ahmad**,[2] **Rashid Naseem**,[2] **Jianyun Shang**,[1] **and Lin Dai**[1]

[1]*School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China*
[2]*Department of IT and Computer Science, Pak-Austria Fachhochschule: Institute of Applied Sciences & Technology, Haripur, Pakistan*

Correspondence should be addressed to Huaping Zhang; kevinzhang@bit.edu.cn

Received 30 October 2020; Revised 1 March 2021; Accepted 8 March 2021; Published 15 March 2021

Academic Editor: Atif Khan



The web is being loaded daily with a huge volume of data, mainly unstructured textual data, which increases the need for information extraction and NLP systems significantly. Named-entity recognition task is a key step towards efficiently understanding text data and saving time and effort. Being a widely used language globally, English is taking over most of the research conducted in this field, especially in the biomedical domain. Unlike other languages, Arabic suffers from lack of resources. This work presents a BERT-based model to identify biomedical named entities in the Arabic text data (specifically disease and treatment named entities) that investigates the effectiveness of pretraining a monolingual BERT model with a small-scale biomedical dataset on enhancing the model understanding of Arabic biomedical text. The model performance was compared with two state-of-the-art models (namely, AraBERT and multilingual BERT cased), and it outperformed both models with 85% F1-score.

## 1. Introduction

Being in the era of digital information, where the web is being loaded with a large volume of data daily (mainly, unstructured text data), the need for information extraction and natural language processing (NLP) systems is increasing significantly. One of the most critical steps in data processing procedures is named-entity recognition (NER). NER—being the task of identifying named entities located in an unstructured text and classifying them into specific semantic classes such as person, location, organization, and disease—is the key step towards question-answering, text summarization, topic detection, and many others. This task is being tackled by many research articles for different languages, mostly English, while other languages are staying behind but progressing considerably [1].

Despite being the official language of 22 countries and being spoken by more than 400 million people around the world, the Arabic language suffers from the scarcity of language resources and NLP systems used to mine the Arabic scripts, especially in the biomedical domain. The biomedical domain has a special and complex structure for named entities as compared to other open text domains. Despite these complexities, it is still witnessing drastic progress in information extraction applications. The Arabic biomedical NER is more challenging: first, for the lack of resources as aforementioned and, second, for the special characteristics of the Arabic language. The Arabic language is morphologically different from other Latin-based languages like English, an example is, named entities can be identified by capitalization. In Arabic, this feature is absent which makes the task harder. Besides this, the Arabic language structure is highly agglutinative with a lack of vowels that are replaced with diacritics, the latter when missing, creates ambiguity [2]. Another challenge is spelling variations of transliterated words (that have no equivalent in

Arabic) coming from other languages. More details about these challenges can be found in [3].

For open domain Arabic NER, considerable work has been done using different kinds of approaches, namely, (i) rule-based approaches [4] and (ii) machine learning approaches, that combine supervised [5], semisupervised [6], and unsupervised methods, where for the medical domain Arabic NER, very scarce research has been conducted.

In this paper, we aim to build a model for Arabic biomedical NER, which can have a better understanding of Arabic clinical data and that can overcome the problem of the lack of data by using a transformer model which we, herein, will refer to as ABioNER.

The remaining of the paper is organized as follows: Section 2 gives a literature review about the different methods used to tackle the task of bioNER for different languages. Section 3 describes the approach used to build the model. The methodology is further described in Section 4 while results and discussion are detailed in Section 5. The paper is concluded in Section 6 along with, stating some future outlines.

## 2. Literature Review

Because of its importance, the task of NER got a lot of attention. In the biomedical domain, various methods were used to tackle this problem. A good portion of those methods was used for the English language, which is rich in resources compared with other languages. Most of the work incorporated deep learning methods that proved to be efficient in this domain with several publications doubling yearly [7].

Table 1 resumes the methods applied for Biomedical NER (BioNER), their results, and the type of semantic entities they focused on.

Table 1 shows that most of the work on bioNER is meant for the English language. Moreover, applying them to the Arabic language will need huge training data, which is a hard task to achieve. As for the unsupervised method, it does not live to the desired standard of good performance.

Deep learning (DL) methods (such as LSTM and BERT) are being widely used recently due to their outstanding performance among other methods on different tasks including NLP.

After Google released the BERT (Bidirectional Encoder Representations from Transformers) model [18], which is a language model that expands the impact of the fine-tuning approach by using a masked model that pretrained a huge amount of textual unstructured data in an unsupervised or self-supervised manner, the effectiveness of this method on downstream NLP tasks including NER becomes clear to researchers. Although BERT comes with a multilanguage model, researchers found that pretraining the BERT model on a specific language improves its performance on that language, which was the case with AraBERT [19] where the authors pretrained the BERT model solely on huge Arabic textual data. The fine-tuning results on different NLP tasks such as sentiment analysis, NER, and question-answering outperformed the original BERT multilanguage model. On the other hand in [16], to improve the model understanding of the biomedical text, the authors of BioBERT pretrained the model on original BERT data plus PubMed abstracts and PMC full-text articles. This method allowed them to enhance the performance of the model on downstream bioNLP tasks.

## 3. Approach

ABioNER model is inspired by AraBERT and BioBert models. On the one hand, BioBERT highlighted the fact that training a language representation model on domain-specific data (biomedical) enhances its performance and understanding of that domain. On the other hand, AraBERT authors proved that the single language BERT model (Arabic) improves the performance on NLP tasks for that specific language.

Since the size of available Arabic biomedical data is considerably small as compared to open domain Arabic, we decided to investigate if the use of small-sized biomedical corpora can enhance the performance of AraBERT model on the biomedical NER task.

Figure 1 illustrates an overview of the model. The model is developed following these steps:

(1) AraBERTv0.1-base weights were used to initialize the ABioNER model. AraBERT was pretrained on general domain Arabic Corpora (namely, Arabic Wikipedia, Arabic news websites articles, the 1.5 billion words Arabic Corpus, OSIAN: the Open Source International Arabic News Corpus).

(2) Pretrain the model on AraBERTv0.1-base original data along with biomedical Arabic literature corpora. The size of available Arabic biomedical data is considerably small compared to open domain Arabic. We assume that even a small domain-specific data added to the pretraining phase is enough to boost model performance in that domain.

(3) Fine-tune and evaluate the model on an optimized version of the silver standard biomedical corpus for the Arabic language [20].

Our work contributions can be resumed as follows:

(i) We show that pretraining a monolingual BERT model on a small-scale domain-specific dataset can still improve the performance of the model on it

(ii) Our model achieved better performance on the bioNER task for the Arabic language, outperforming original multilingual BERT and AraBERT models

(iii) To the best of our knowledge, this is the first work for Arabic biomedical NER of this kind

## 4. Materials and Methods

In this section, the steps of pretraining and fine-tuning of the ABioNER model are described.

Table 1: An overview of methods used for biomedical named-entity recognition.

| Paper | Language | NE | Method | Results |
|---|---|---|---|---|
| [8] | English | Gene protein | SVM | Best balanced F1-score = 0.79 |
| [9] | English | Chemical mentions | Hybrid (CRF + dictionary) | F1-score = 68.1 |
| [10] | English | Problem treatment test protein DNA RNA cell type cell line | Unsupervised learning | Overall performance (exact micro-F) Pittsburgh dataset: 53.1 GENIA dataset: 39.5 |
| [11] | English | Disease | Hybrid (stacked ensemble + fuzzy matching) | F1-score = 89.12% |
| [12] | English | Disease | Multiple label convolutional neural networks | F1-score NCBI corpus: 85.17% CDR corpus: 87.83% |
| [13] | English | Document-level chemical NER | Hybrid (attention-based BiLSTM-CRF) | F1-score HEMDNER corpus: 91.14% CDR corpus: 92.57% |
| [14] | English | Genes diseases protein DNA RNA cell type cell line | n-Gram character and word embeddings via convolutional neural network | F1-score: NCBI dataset: 87.26% Biocreative II dataset: 87.26% JNLPBA dataset: 72.57% |
| [15] | English | Chemicals<br>Diseases<br>Species<br>Gene<br>Protein | Transfer learning | F1-score:<br>88.21<br>82.09<br>87.01<br>83.09 |
| [16] | English | Diseases<br>Species | Bidirectional encoder representations from transformers | Best F1-score<br>89.71<br>75.31 |
| [17] | Spanish/ Swedish | Spanish: disease/drug Swedish: body part/disorder/finding | Bidirectional long short-term memory network | Avg. F1-score: Spanish: 75.25 Swedish: 76.04 |
| [2] | Arabic | Disease diagnosis symptoms treatment methods | Bayesian belief network (BBN) | Avg. F1-score: 71.05% |

*4.1. Pretraining ABioNER.* AraBERTv0.1-base was pretrained on a set of 2.7 billion words collected from different general domain Arabic corpora. ABioNER model uses the same set of corpora in addition to Arabic biomedical literature corpora which we collected manually from different sources:

(i) PubMed: we queried all available and relevant articles and abstracts written in the Arabic language. The number of available Arabic medical articles in PubMed is very negligible.

(ii) MedlinePlus Health Information in Arabic [21]: we extracted Arabic clinical text from provided articles.

(iii) Journal of Medical and Pharmaceutical Sciences [22], Arab Journal for Scientific Publishing [23]: extracting Arabic text from these journals was a hard task and time-consuming; due to the language errors occurring during converting PDF to text using OCR (optical character recognition). These errors need to be corrected manually.

(iv) Eastern Mediterranean Health Journal [24]: abstracts in the Arabic language were extracted.

All corpora were combined in one text file where each line has one sentence and a line separating the documents from each other. The whole corpora constructed a set of around 500k words. The same original configuration of AraBERT was used which is the same as the one used for BERT. For validation, 15% of the dataset was used.

*4.2. Fine-Tuning for NER.* As there are no existing annotated corpora for Arabic biomedical Relation Extraction or Question Answering, ABioNER was only fine-tuned on the NER task.

Fine-tuning is a way to speed up the training and overcome the problem of a small-sized dataset, which is the case for the Arabic language that suffers from annotated biomedical corpora. It serves the purpose of continuing training on the existing pretrained model on a specific small dataset, which is in this case, an optimized version of the silver standard biomedical tagged corpus for the Arabic language. As far as we know, there are no other freely available annotated datasets to use for evaluation.

*4.3. Dataset.* The silver standard biomedical corpus used in this work is originally labelled with 13 different semantic types: Disease or Syndrome, Therapeutic or Preventive Procedure, Diagnostic procedure, Sign or Symptom, Antibiotic, Virus, Enzyme, Hormone, Clinical Drug, Mental or Behavioral Dysfunction, Injury or Poisoning, Bacterium, and Gene or Genome.

The corpus is tagged with the IOB2 encoding format. In this paper, only "Disease or Syndrome" and "Therapeutic or Preventive Procedure" semantic types were used. Therefore, we only selected sentences with those two mentions. If any other entity tag is found in any of the selected sentences, it was set to "O," which are, respectively, (B-DS, I-DS) and (B-TPP, I-TPP). The split of 80% and 20% was used for training and test datasets, respectively.

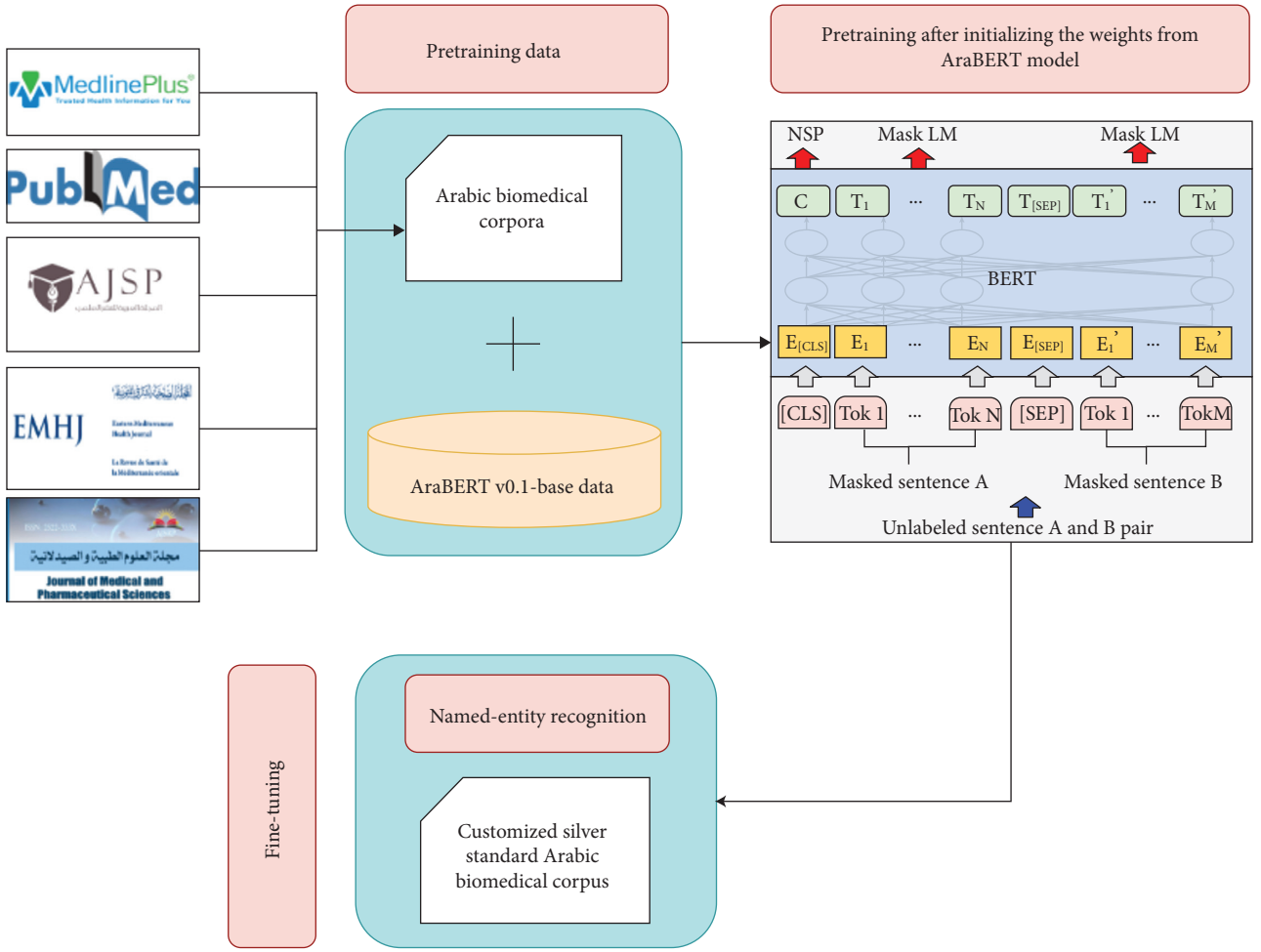


Figure 1: Overview of ABioNER model.

## 5. Results and Discussion

To prove the effectiveness of ABioNER, we compared it with AraBERT and BERT multilingual cased models. All models were fine-tuned on the optimized corpus for the biomedical named-entity recognition task.

For fine-tuning, we used a single NVIDIA GPU, with the following setup: learning rate: $5e-5$; batch size: 16; and training epochs: 20. This setup was selected after many trials, and it gave the best results for all models. In the following tables and figures of results, DS and TPP refer to "Disease or Syndrome" and "Therapeutic or Preventive Procedure," respectively. Table 2 shows the results obtained from fine-tuning all models on the optimized biomedical silver standard corpus for "Disease or Syndrome" and "Therapeutic or Preventive Procedure" entities. And, Table 3 shows the corresponding macro-F1-scores. Best scores are shown in bold.

Figure 2 shows NER test results graphs. AbioNER reached the highest performance on epoch 3 (Figure 2(c)). AraBERT performance as expected was better than the BERT multilingual cased model. AraBERT F1-score results for DS named-entity are similar to ABioNER, but ABioNER has better recall and outperforms both models on TPP named entity. Overall, ABioNER model proved to be more effective for the Arabic BioNER task as shown in Figure 3, overcoming the problem of data scarcity.

Table 2: NER detailed test results for ABioNER, AraBERT, and BERT multilingual cased models.

| Model | NE type | Precision | Recall | F1-score |
|---|---|---|---|---|
| BERT multilingual cased | DS | 0.89 | 0.87 | 0.88 |
| | TPP | 0.74 | 0.78 | 0.76 |
| AraBERT | DS | **0.90** | 0.89 | 0.89 |
| | TPP | 0.77 | 0.78 | 0.78 |
| ABioNER | DS | 0.88 | **0.90** | **0.89** |
| | TPP | **0.82** | **0.81** | **0.82** |

Table 3: Results for DS and TPP entities combined.

| Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| BERT multilingual cased | 96.88 | 81.43 | 82.83 | 82.12 |
| AraBERT | 97.48 | 83.69 | 83.69 | 83.69 |
| ABioNER | **97.64** | **85.47** | **85.83** | **85.65** |

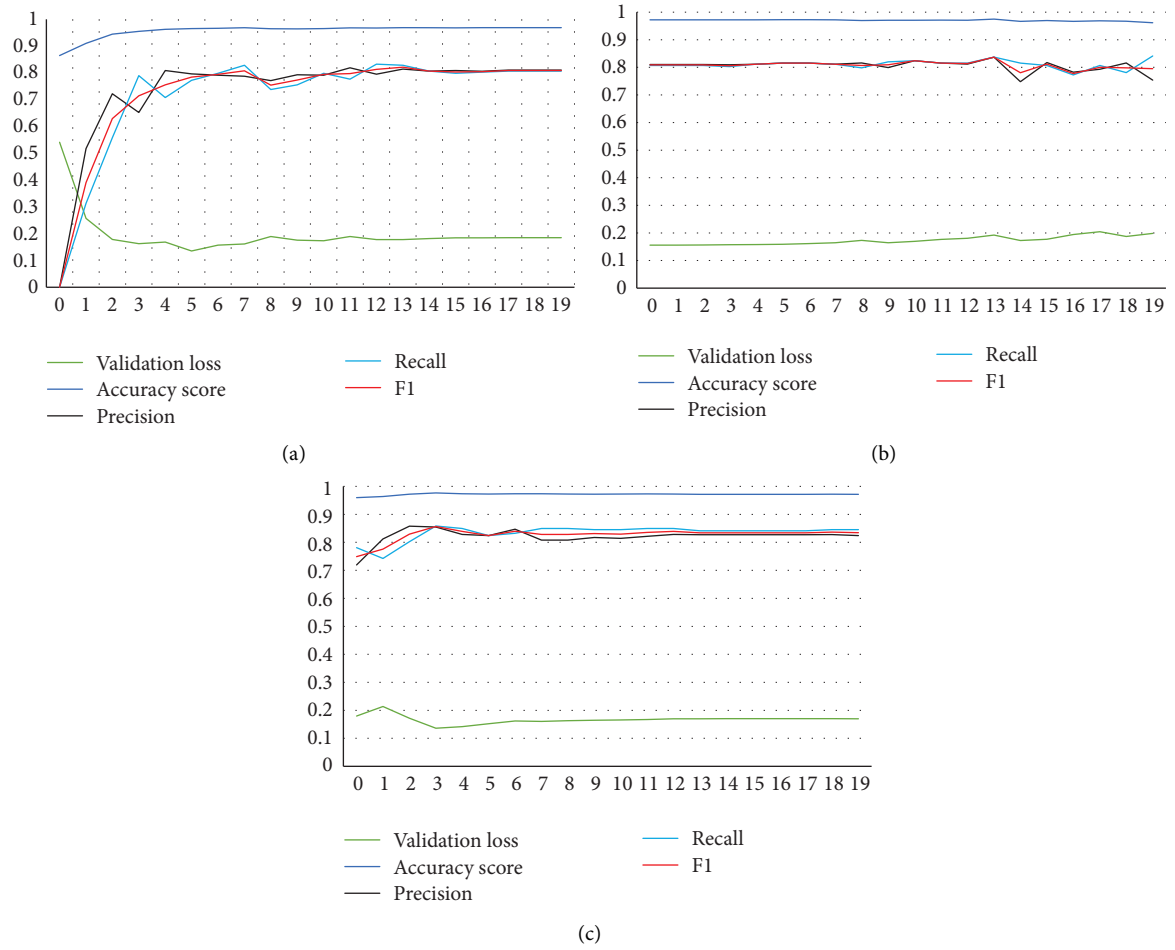


Figure 2: Test results for bioNER. (a) BERT multilingual cased. (b) AraBERT. (c) ABioNER.

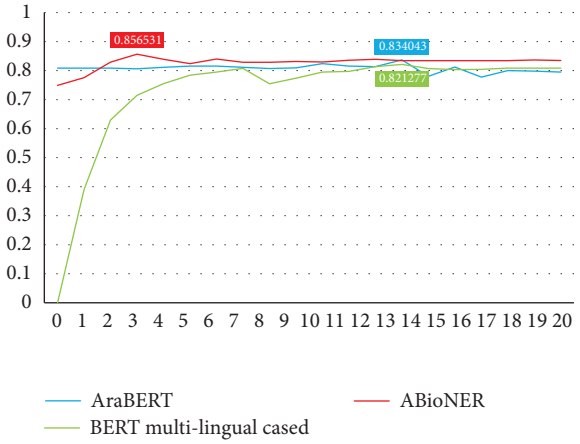


Figure 3: F1-score result for all models.

## 6. Conclusions

In this article, we presented ABioNER, a model for Arabic biomedical named-entity recognition, inspired by two state-of-the-art BERT-based models (AraBERT and BioBERT). The model was pretrained on AraBERT original data plus medical Arabic literature collected from different sources. The results of fine-tuning the model on the bioNER task outperformed those of AraBERT and BERT multilingual cased models on "Disease or Syndrome" and "Therapeutic or Preventive Procedure" semantic type. This work proved that pertaining a monolingual BERT model on small-scale biomedical data can improve the model understanding on biomedical domain data. One of the study limitations is that the model was only tested on two entity types. Moreover, it was not evaluated for relation extraction or other downstream NLP tasks due to the lack of datasets for these tasks. We believe that the performance can get better if pretrained on a larger biomedical dataset, which will be the goal of future work along with testing on more named-entity types. We believe this work can help researchers interested in domain-specific text mining.

## Data Availability

The data used in this article are available from the first and corresponding authors upon request.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

The research work was funded by the Beijing Municipal Natural Science Foundation (Grant no. 4212026), National Science Foundation of China (Grant no. 61772075), and National Key Research and Development Project of China (Grant no. 2018YFC0832304). The authors are thankful to them for their financial support.

## References

[1] N. Boudjellal, H. Zhang, A. Khan, and A. Ahmad, "Biomedical relation extraction using distant supervision," *Scientific Programming*, vol. 2020, Article ID 8893749, 9 pages, 2020.

[2] S. Alanazi, B. Sharp, and C. Stanier, "A named entity recognition system applied to Arabic text in the medical domain," *International Journal of Computer Science Issues*, vol. 12, no. 3, 2015.

[3] R. E. Salah and L. Qadri binti Zakaria, "A comparative review of machine learning for Arabic named entity recognition," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, no. 2, pp. 511–518, 2017.

[4] M. Oudah and K. Shaalan, "Nera 2.0: improving coverage and performance of rule-based named entity recognition for Arabic," *Natural Language Engineering*, vol. 23, no. 3, pp. 441–472, 2017.

[5] N. Omar and N. F. Mohammed, "Arabic named entity recognition using artificial neural network," *Journal of Computer Science*, vol. 8, no. 8, pp. 1285–1293, 2012.

[6] M. Althobaiti, *Minimally-supervised Methods for Arabic Named Entity Recognition*, University of Essex, Colchester, UK, 2016.

[7] S. Wu, K. Roberts, S. Datta et al., "Deep learning in clinical natural language processing: a methodical review," *Journal of the American Medical Informatics Association*, vol. 27, no. 3, pp. 457–470, 2020.

[8] T. Mitsumori, S. Fation, M. Murata, K. Doi, and H. Doi, "Gene/protein name recognition based on support vector machine using dictionary as features," *BMC Bioinformatics*, vol. 6, no. 1, pp. 1–10, 2005.

[9] T. Rocktäschel, M. Weidlich, and U. Leser, "ChemSpot: a hybrid system for chemical named entity recognition," *Bioinformatics*, vol. 28, no. 12, pp. 1633–1640, 2012.

[10] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: experiments with clinical and biological texts," *Journal of Biomedical Informatics*, vol. 46, no. 6, pp. 1088–1098, 2013.

[11] B. Bhasuran, G. Murugesan, S. Abdulkadhar, and J. Natarajan, "Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases," *Journal of Biomedical Informatics*, vol. 64, pp. 1–9, 2016.

[12] Z. Zhao, Z. Yang, L. Luo et al., "Disease named entity recognition from biomedical literature using a novel convolutional neural network," *BMC Medical Genomics*, vol. 10, no. 5, 2017.

[13] L. Luo, Z. Yang, P. Yang et al., "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition," *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, 2018.

[14] Q. Zhu, X. Li, A. Conesa, and C. Pereira, "GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text," *Bioinformatics*, vol. 34, no. 9, pp. 1547–1554, 2018.

[15] J. M. Giorgi and G. D. Bader, "Transfer learning for biomedical named entity recognition with neural networks," *Bioinformatics*, vol. 34, no. 23, pp. 4087–4094, 2018.

[16] J. Lee, W. Yoon, S. Kim et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[17] R. Weegar, A. Pérez, A. Casillas, and M. Oronoz, "Recent advances in Swedish and Spanish medical entity recognition in clinical texts using deep neural approaches," *BMC Medical Informatics and Decision Making*, vol. 19, no. 7, pp. 1–14, 2019.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the NAACL HLT 2019*, vol. 1, pp. 4171–4186, Minneapolis, MN, USA, June 2019.

[19] W. Antoun, F. Baly, and H. Hajj, "AraBERT: transformer-based model for arabic language understanding," in *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, February 2020.

[20] N. Boudjellal, H. Zhang, A. Khan, A. Ahmad, R. Naseem, and L. Dai, "A silver standard biomedical corpus for Arabic language," *Complexity*, vol. 2020, Article ID 8896659, 7 pages, 2020.

[21] "Health information in Arabic: MedlinePlus," 2021, https://medlineplus.gov/languages/arabic.html.

[22] "Journal of medical and pharmaceutical sciences," 2021, https://www.ajsrp.com/journal/index.php/jmps.

[23] "Arab journal for scientific publishing," 2021, https://www.ajsp.net/.

[24] WHO EMRO, "EMHJ home," *Eastern Mediterranean Health Journal*, vol. 27, 2021, http://www.emro.who.int/emh-journal/eastern-mediterranean-health-journal/home.html.