Fine-Tuning Transformer Models for Adverse Drug Event Identification and Extraction in Biomedical Corpora: A Comparative Study

Chanaa Hiba(✉), El Habib Nfaoui, and Chakir Loqman

LISAC Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco
hiba.chanaa@usmba.ac.ma

Abstract. Adverse Drug Events (ADEs) are potentially fatal problems that patients can deal with only if they have a solid awareness of them. With the available amount of unstructured textual data from biomedical literature, electronic records, and social media (e.g., tweets), early detection of unfavorable reactions and sharing them with biomedical experts, pharma companies, and healthcare professionals is a necessity, as this can prevent morbidity and save many lives. The Biomedical Named Entity Recognition (BioNER) task can be considered the initial step toward resolving this issue. In this paper, we present an empirical evaluation experiment by fine-tuning pretrained language models for detecting biomedical entities (e.g., drug-names and symptoms). We fine-tuned five transformer models: BERT (Bidirectional Encoder Representations from Transformers), SpanBERT, BioBERT, BlueBERT, and SCIBERT, on two well-known biomedical datasets, CADEC and ADE-corpus. The evaluation results demonstrate that BioBERT which was pretrained on both general and domain-specific (biomedical domain) corpora outperformed all other models on both datasets and reached 90.3% and 68,73% on the F1-score in the ADE and CADEC corpora, respectively.

Keywords: Adverse Drug Event · Bio-Named Entity Recognition · Deep Learning · Natural Language Processing · Fine-tuning · Transformer Models

1 Introduction

Human health is the greatest wealth that life can bestow upon us. Unfortunately, this wealth is at risk of being spoiled by the occurrence of numerous adverse events, which currently represent one of the most worrying topics in the biomedical field. These bad events are countless and diverse and have been addressed by a huge number of scholars and practitioners; however, some of them suffer from conceptual ambiguity; for example, clinicians are frequently unable to identify or handle cases of drug-related damage, which can result in significant morbidity and mortality [1]. Therefore, researchers on

[2] used a case study of a patient who experienced many adverse drug events to identify terms such as adverse event, adverse drug reaction, medication error, and side effect.

S. Motahhir and B. Bossoufi (Eds.): ICDTA 2023, LNNS 668, pp. 957–966, 2023.
https://doi.org/10.1007/978-3-031-29857-8_95

According to the "International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use" an adverse event (AE) could be any unintended medical incident that may occur when using a medication but does not necessarily have a causal relationship with this medication. Thus, an AE can be any undesirable, accidental indication, symptom, or illness that occurs while using a medicinal product, whether or not it is regarded as related to it [3, [2]. Moreover, the "International Conference on Harmonization" defines adverse drug reactions as all noxious and unintended responses to a medicinal product related to any dose. Thus, an adverse drug reaction is an adverse event that has a direct connection to a drug. An adverse medication reaction, such as renal failure caused by ibuprofen, happens at typical doses and is brought on by the drug's effect [2].

As a potential challenge in Natural Language Processing (NLP), Named Entity Recognition (NER) has drawn much attention in recent years [4]. The primary goal of the NER task is to identify and categorize references of named entities in a given unstructured text into predefined semantic categories. These entities often relate to a person's name, nationality, and institution in traditional NER tasks [5]. In this work, we will focus on biomedical NER (BioNER), a field that seeks to extract specific medical entities from the medical literature. These entities include symptoms, diseases, drugs, etc.

In this work, we proposed to explore various pretrained language models, some of which are built using the global English domain (i.e., BERT [6] and SpanBERT [7]) and others are built on biomedical domain-specific models (i.e., BioBERT [8], SCIBERT [9] and BlueBERT [10]) to accomplish the BioNER task using ADE and CADEC corpora.

The remainder of the paper is structured as follows. Section 2 highlights the problem formulation of this study and discusses the models used during our experiments. Section 3 illustrates the dataset used in this study and discusses the obtained results. Finally, sect. 4 summarizes the conclusions of this study.

2 Materials and Methods

2.1 Problem Formulation

ADE identification is a sequence labeling problem (e.g., BioNER task). The ultimate goal of this task is to identify entities such as "Disease-name," "Drug-name," and "Symptoms," from the input sentences. Given an input sequence of words $X = \{x1, x2, ..., xi\}$, the task is to label each word xi in the sequence with a tag yi which is a part of the tag set $Y = \{y1, y2, ..., yi\}$.

2.2 BioNER Utilizing Pretrained Language Models

Named Entity Recognition (NER) is the method of defining and categorizing entities in a given text. Biomedical NER (BioNER) is a hot topic in health coverage since it is the first step in performing relation extraction and clinical decision-making tasks.

[11]. BioNER is complicated when compared to NER in other domains since labeled data in the biomedical domain are limited in quantity and are expensive to collect, and

it necessitates the detection of complex entities that are not prevalent in other fields [12]. Deep learning techniques that use massive volumes of unstructured data, such as Bi-LSTM with CRF [13] and BERT fine-tuning [14], have recently been used to obtain State-Of-The-Art (SOTA) findings concerning the BioNER task. Lately, another work has reached SOTA performance by combining BioBERT word embeddings with BiLSTM-CNN-Char and making certain architectural adjustments: a) removing lexical characteristics such as POS tags and introducing new character level features. b) Generating token feature maps that capture information such as spelling and casing using a 1D convolution layer composed of 25 filters with kernel size 3. These added features proved to be useful while dealing with misspellings and out-of-vocabulary tokens.

The most prevalent NER annotation system is BIO tagging, where 'B' stands for Beginning of entity, 'I' stands for Inside of entity, and 'O' stands for Outside of entity. Figure 1 presents the overall flowchart of the BioNER task utilizing pre-trained language models. First, we extract the sentences from the dataset and tokenize each sentence. Second, we applied sentence tokens labelling by means of which we assign to each word a BIO (Begin, Inside, and Outside) labels to denote the position and type of the token inside an entity mention. Third, the input sentences were then processed by inserting a ([CLS]) and ([SEP]) tokens at the start and the end of the text, respectively. Then, the generated inputs were fed into the model after tokenization based on its vocabulary. Eventually, the NER function is completed by putting a further linear layer on top of the contextual representations outputted by the model to predict token tags.

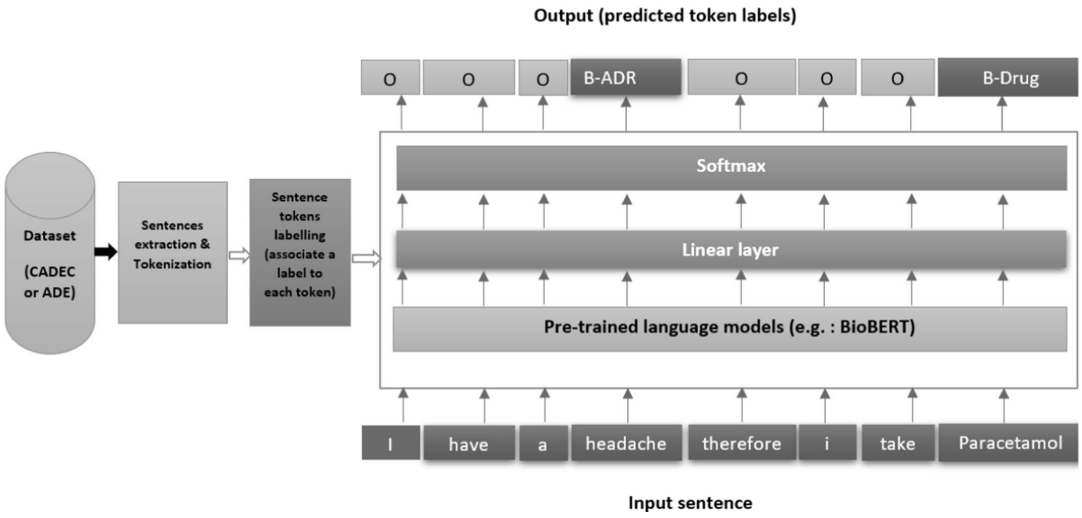

Fig. 1. The global flowchart of the BioNER task utilizing pre-trained language models.

2.3 Pre-trained Language Models

In this study we examined five pre-trained language representation models for the general and biomedical domains: BERT, SpanBERT, BioBERT, SCIBERT and BlueBERT.

BERT: (Bidirectional Encoder Representations from Transformers) [6] is an embeddings model trained with bidirectional transformers [15]. It was trained on large unlabeled text corpora to design deep bidirectional embeddings to better represent sequences of text. It employs two main training objectives: masked language modeling objective and next sentence prediction. Therefore, bidirectional representation (instead of unidirectional) of text sequences is critical to better encode words in natural language [6]. Thus, because of complicated interactions between biomedical terms we anticipate that BERT bidirectionality is essential in biomedical text analytics [15].

BioBERT: Is the first language representation model that has been pre-trained for the biomedical field [8]. Because biomedical documents contain a substantial number of domain-specific words that are largely understood by biomedical professionals. Consequently, NLP models that were built for general-purpose language comprehension frequently perform poorly in biomedical text mining applications. Therefore, BioBERT had previously been trained on corpora from the field of biomedicine (PubMed abstracts and PMC full-text articles). It revealed good results on three well-known NLP tasks (NER, RE, and QA), however in this study, we will illustrate its effectiveness on the BioNER task using both ADE and CADEC datasets.

SciBERT: Is a deep neural network training tool for NLP applications that requires a large amount of labeled data. It was pretrained from scratch using a random sample of 1.14 M documents from Semantic Scholar [16]. This corpus contains 18% publications from the subject of computer science and 82% papers from the broad biomedical sector. They used the entire text of the papers rather than simply the abstracts. The average length of a paper is 154 sentences (2,769 tokens), resulting in a corpus size of 3.17 billion tokens, which is similar to the 3.3 billion tokens on which BERT was trained [9].

SpanBERT: Is a pre-trained model for improving the representation and prediction of text spans. It differs from the original BERT in two points: 1) The masking mechanism: instead of masking random individual tokens it masks an entire contiguous span of tokens. 2) It uses a new span-boundary objective to train the model to predict the entire masked span based on tokens in the boundary [7].

BlueBERT [10]: Is a biomedical domain-specific model. It was initialized with BERT weights and then pre-trained utilizing a large clinical and biomedical domain (i.e., PubMed abstracts and clinical notes MIMIC- III).

3 Experiments and Results

3.1 Data

In this paper, we explored two corpora, ADE (Adverse Drug Event) and CADEC (CSIRO Adverse Drug Event Corpus). Table 1 shows the statistics of both corpuses.

They follow the same basic annotation scheme, whereas the CADEC corpus has more entities than the ADE corpus.

Table 1. Overall statistics of both CADEC and ADE corpuses.

Corpus	Origin	Type	Size	Entities
ADE [17]	MEDLINE	Literature	400 Abstracts (6821)	Disease, Adverse effects
CADEC[19]	AskaPatient	Medical Forum	1253 posts (7398 sentences)	Drug, adverse effect, disease, symptom, finding

ADE-Corpus (version 2) [17] respects many features that should be taken into account, the most significant ones are the corpus's domain appropriateness and the target user population. For the biomedical domain, medical case reports are the most relevant data sources because they contain essential information about individual patients' symptoms, signs, diagnosis, therapy, and so on. More significantly, case reports might act as an early warning signal for drugs with unreported or unexpected side effects. MEDLINE articles were utilized because of their free public availability. As a result, the ADE corpus is a subset of MEDLINE case reports. This dataset is used to determine whether a sentence is related to an ADE or not.

CADEC dataset [19] is a novel, richly annotated corpus of patient-reported ADEs from medical forums. The dataset is derived from social media posts and comprises content that is mostly written in colloquial English and frequently veers away from traditional English grammar and punctuation norms. Annotations include references to terms such as medications, negative effects, symptoms, and diseases that are associated with terms with the same names in restricted vocabularies. This corpus is crucial for studies on information extraction, or much more broadly text analytics, from social networks to extract potential adverse drug events from first-hand patient experiences (NER task).

3.2 Experiments

Further examination of the ADE corpus reveals that sentences are frequently repeated to identify different combinations of drugs and adverse reactions. This is not suitable in an NER setting because if we assigned one set of token labels per row in this dataset as-is, we would end up labeling the same tokens differently in the same sentences. This would confuse the model during fine-tuning, so we must first consolidate all of the variations provided for each unique sentence before labeling all known entities. For the CADEC corpus, as the text documents are separated from the annotations, we had to manually annotate each sentence in a given document based on the indices presented in the associated annotations file.

In the first stage, for both datasets we apply the NER sequence labeling task in which each token in the sequence is assigned a predetermined IOB tag, where "B" corresponds to the beginning of an entity, "I" means inside an entity, and "O" represents all other nonentity words. This results in 11 possible classes for each token in the CADEC dataset which are the following: ('O' - outside any entity we care about, 'B-ADR' - 'I- ADR'

- 'B-Drug' - 'I-Drug' - 'B-Symptom' - 'I-Symptom' - 'B-Finding' - 'I-Finding' - 'B-Disease' - 'I-Disease'), and in 5 possible classes for the ADR corpus ('O' - 'B-Drug' - 'I- Drug', 'B-Effect' - 'I-Effect). Second, the input sentences were tokenized by the default tokenizer using the vocabulary of the pre-trained language models and adding the special [CLS] token at the start of the sentence and a [SEP] token at the end and then fed as input into the model. Finally, the NER process is completed by predicting token labels using an additional linear classification layer on top of the pre-trained language model Fig. 2 shows an example of a sequence labeling performed by the fine-tuned BioBERT model). To handle the out-of-vocabulary (OOV) issue, transformer models often break original words into many pieces of sub-words (WordPiece tokenization) by inserting a specific tag "##" in front of the following sub-words. The Hugging Face transformers library [18] (implemented in PyTorch) was used to obtain all transformer and tokenizer models. Table 2 shows the optimal hyperparameters that were utilized for all the models to obtain good performance based on a simple grid search with a small search space.

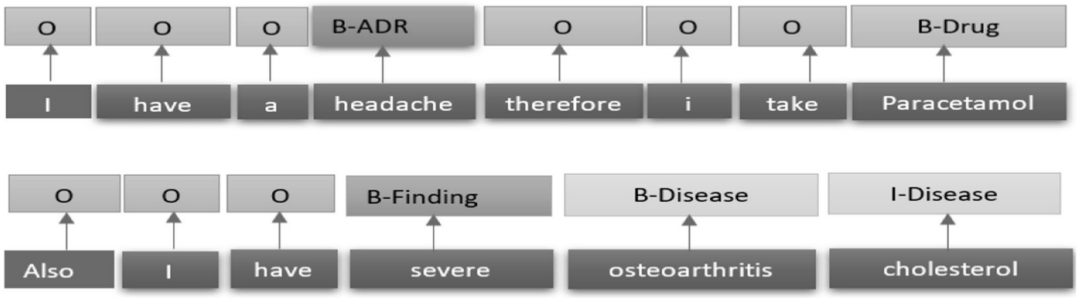

Fig. 2. An example of sequence labeling performed by the fine-tuned BioBERT model.

Table 2. Hyperparameters employed for all models.

Hyperparameters	Search space	Optimal Value
training epochs	[5–15, 15]	8
Learning rates	[1e-6-1e-2]	1e-5
Training Batch size	[4, 8, 16, 32, 64]	8

The effectiveness of all transformer-based NER models was assessed using four evaluation metrics, including precision, recall, F1-score [19] and accuracy measures. Formally:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

$$Precision = TP/(TP + FP) \tag{2}$$

$$Recall = TP/(TP + FN) \tag{3}$$

$$\text{F1\_}score = 2 \times (\text{Pr}\,ecision \times Recall)/(\text{Pr}\,ecision + Recall) \quad (4)$$

where TP (True Positives) is the number of data points properly categorized as positives; TN (True Negatives) refers to the number of data points properly classified as negatives; FP (False Positives) is the data points misclassified as positives; and FN (False Negatives) is the number of data points misclassified as negatives. Because the dataset's categories are unbalanced, we utilized the F1-score measure, which better indicates performance in this case.

3.3 Results

Table 3 shows the overall performance of fine-tuning the five pre-trained language models for the BioNER task on the CADEC dataset using Precision, Recall, F1-score and Accuracy measures. Among all models, we observe that BioBERT achieved the highest score on precision with an improvement of 7.6%. It also acquires the best outcomes with F1-score and accuracy by reaching up to 68.73%, and 91.9%, respectively. On the other hand, SpanBERT outperforms all the other models on the recall metric.

Table 3. Overall performance results on the test set of the CADEC dataset.

Models	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Accuracy (%)**
BERT	60.1	68.9	64.2	91.0
SpanBERT	65.4	**72.3**	68.71	91.7
BioBERT	**65.8**	71.8	**68.73**	**91.9**
SciBERT	61.4	70.9	65.8	91.5
BlueBERT	58.2	68.7	63.03	91.3

Table 4 shows the detailed results of fine-tuning the five pre-trained language models for the BioNER task on the ADE dataset using Precision, Recall, F1-score and Accuracy measures. BioBERT considerably surpasses all other models, ranking highly on all measures. Particularly, it surpasses the competing models by 3% on recall and achieves 90.3% and 96.3 on F1-score and accuracy, respectively. All measures show a considerable improvement, with the exception of accuracy, where the SpanBERT model obtains a score equivalent to BioBERT.

3.4 Discussion

In this study, we conducted experiments with a variety of pre-trained language models as previously stated to evaluate the effectiveness of fine-tuning them to address the BioNER challenge on biomedical corpora. Experimental analysis demonstrates the relevance of all five embedding pre-trained models. Indeed, the deep architecture of the pre-trained

Table 4. Overall performance results on the test set of the ADE dataset.

Models	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
BERT	81.6	89.4	85.3	91.0
SpanBERT	**88.4**	91.6	90.0	95.8
BioBERT	**88.4**	**92.4**	**90.3**	**96.3**
SciBERT	87.0	90.7	88.8	96.1
BlueBERT	84.1	89.9	86.9	95.6

models (which contain 12 transformer encoder blocks, each of which involves fully connected layers, multi-head attention layers and output layer) allows them to capture contextual information between words in the given sequence of text. As mentioned above, BERT and SpanBERT were pre-trained utilizing global English corpuses from English Wikipedia and Books Corpus. On the other hand, BioBERT and BlueBERT are built on top of BERT and have been fine-tuned with additional biomedical domain corpora. SciBERT, however, was trained from scratch utilizing biomedical corpora that provide relevant and representative word embeddings for domain-specific tasks. We hypothesize that the nature of the study corpora (ADE and CADEC) which include both general and biomedical domains is the main reason why BioBERT reveals good results over other competing models since it was fine- tuned based on original BERT utilizing biomedical corpora (PubMed Abstracts, PMC Full-text articles Number). As shown in Tables 3, 4 SpanBERT slightly reveals better performance than SciBERT and BlueBERT. The first reason might be the mechanism of masking spans of tokens (instead of individual tokens as in the original BERT) and then predicting the entire masked span based on the tokens at the span's boundary. The two corpora exhibited remarkable variance in performance, demonstrating the intrinsic distinctions between them (e.g., F1 scores of 68.73% and 90.3% on CADEC and ADE datasets, respectively, for the same BioBERT model). Eventually, this study highlights the utility of fine tuning large pre-trained models to achieve satisfactory results on the BioNER task using CADEC and ADE corpuses that are approximately close to state- of-the art [20],which is based on the BiLSTM-CNN-Char architecture [21] combined with BioBERT embeddings.

4 Conclusion and Future Work

In this study, we demonstrate the effectiveness of fine-tuning pre-trained language models to accomplish the NER task on biomedical corpora. The findings show that all models obtain satisfactory performance, with BioBERT outperforming the other models. This is because the pre-training corpora involved both general and biomedical domains. Furthermore, disparities in performance between the CADEC and ADE corpora were observed, which indicates that the natural dataset had a significant role in achieving good results. In the future, we plan to apply these models to other NLP tasks such as relation extraction and combine them with other pertinent models to obtain the best results.

References

1. Classen, D.C., Pestotnik, S.L., Evans, R.S., Classen, C.: Computerized surveillance of adverse drug events in hospital patients*. Qual Saf Heal. Care **14**, 221–226 (2005). https://doi.org/10.1136/qshc.2002.002972
2. Schroeder, S.A.: How Many hours is enough? an old profession meets a new generation. Ann. Intern. Med. **140**(10), 838–839 (2004). https://doi.org/10.7326/0003-4819-140-10-200405180-00017
3. Agency, E.M.: ICH E2A - clinincal safety data managements: definitions and standards for expedited reporting. Drug News **23**(1), 71 (2010)
4. Zhang, R., Zhao, P., Guo, W., Wang, R., Lu, W.: Medical named entity recognition based on dilated convolutional neural network. Cogn. Robot., **12,** 13–20, (2022) https://doi.org/10.1016/j.cogr.2021.11.002
5. Sundheim, B., Road, G., Diego, S., Grishman, R., York, N.: Message U n d e r s t a n d i n g C o n f e r e n c e - 6: A Brief History Ocean Surveillance Center Evaluation Division (NRaD) Short-term subtasks Portability.
6. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.,: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Oct. 2018 http://arxiv.org/abs/1810.04805
7. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: Improving pre-training by representing and predicting spans. Trans. Assoc. Comput. Linguist. **8**, 64–77 (2020). https://doi.org/10.1162/tacl_a_00300
8. Lee, J., et al.: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020). https://doi.org/10.1093/bioinformatics/btz682
9. Beltagy, I., Lo, K., Cohan, A.: SCIBERT: A pretrained language model for scientific text. In: EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th International Joint Conference Natural Language Processesing Proceedings Conference, pp. 3615–3620, (2019) https://doi.org/10.18653/v1/d19-1371
10. Peng, Y., Yan, S., Lu, Z.: Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In: BioNLP 2019 - SIGBioMed Work. Biomed. Nat. Lang. Process. Proc. 18th BioNLP Work. Shar. Task, no. iv, pp. 58–65, (2019). https://doi.org/10.18653/v1/w19- 5006
11. De Bruijn, B., Martin, J.: Getting to the (c)ore of knowledge: mining biomedical literature. Int. J. Med. Inform. **67**(1–3), 7–18 (2002). https://doi.org/10.1016/S1386-5056(02)00050-3
12. Dai, X.: Recognizing complex entity mentions: A review and future directions. In: ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Student Res. Work., pp. 37–44 (2018) https://doi.org/10.18653/v1/p18-3006
13. Li, F., Zhang, M., Tian, B., Chen, B., Fu, G., Ji, D.: Recognizing irregular entities in biomedical text via deep neural networks. Pattern Recognit. Lett. **105**, 105–113 (2018). https://doi.org/10.1016/j.patrec.2017.06.009
14. Sharma, R., Chauhan, D., Sharma, R.: Named Entity Recognition System for the Biomedical Domain. In: Proc. 17th Conf. Comput. Sci. Intell. Syst FedCSIS 2022, vol. 30, pp. 837–840 (2022) https://doi.org/10.15439/2022F63
15. Vaswani, A., et al.: Attention Is All You Need, Jun. 2017 http://arxiv.org/abs/1706.03762
16. Cariello, M.C., Lenci, A., Mitkov, R.: A Comparison between Named Entity Recognition Models in the Biomedical Domain. 76–84 (2022). https://doi.org/10.26615/978-954-452-071-7_009
17. Ammar, W. et al.: Construction of the literature graph in semantic scholar. In: NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 3, pp. 84–91, (2018) https://doi.org/10.18653/v1/n18-3011

18. Gurulingappa, H., Rajput, A.M., Roberts, A., Fluck, J., Hofmann-Apitius, M., Toldo, L.: Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports J. Biomed. Inform. **45**(5) 885–892 (2012)https://doi.org/10.1016/J.JBI.2012.04.008
19. Karimi, S., Metke-Jimenez, A., Kemp, M., Wang, C.: Cadec: A corpus of adverse drug event annotations. J. Biomed. Inform. **55**, 73–81 (2015). https://doi.org/10.1016/j.jbi.2015.03.010
20. Wolf, T., et al.:Transformers: State-of-the-Art Natural Language Processing. pp. 38–45 (2020) https://doi.org/10.18653/v1/2020.emnlp-demos.6
21. Goutte, C., Gaussier, E.: A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 345–359. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-31865-1_25
22. Ul Haq, H., Kocaman, V., Talby, D.: Mining Adverse Drug Reactions from Unstructured Mediums at Scale. 2022, Accessed: Oct. 03 2022. www.aaai.org
23. Kocaman, V., Talby, D.: Biomedical Named Entity Recognition at Scale. vol. 19958 (2019)