ISSN 1870-4069

Enhancing Biomedical NLP in Spanish: Large Language Model's Domain Adaptation for Named Entity Recognition in Clinical Notes

Rodrigo del Moral, Orlando Ramos-Flores,
Helena Gómez-Adorno

Universidad Nacional Autónoma de México,
Instituto de Investigaciones en Matemáticas
Aplicadas y en Sistemas,
Mexico

rodrigodelmoral@comunidad.unam.mx,
orlando.ramos@aries.iimas.unam.mx,
helena.gomez@iimas.unam.mx

Abstract. Named Entity Recognition (NER) plays a crucial role in extracting valuable information from clinical texts, enabling the organization of relevant data within databases and knowledge bases. However, automatic recognition of named entities in clinical notes poses significant challenges due to the diversity of writing styles and vocabularies. This paper investigates domain adaptation techniques' effectiveness of large language models in enhancing NER in Spanish clinical texts. We propose using pre-trained language models and a corpus of Mexican clinical notes, applying anonymization techniques to protect sensitive information. The study evaluates the performance of adapted models on publicly available NER datasets, Cantemist and PharmaCoNER. Results indicate a slight improvement in NER performance, showcasing the potential of domain adaptation to tailor language models for specific clinical domains. Future research directions are discussed to refine domain adaptation strategies and promote the responsible use of AI in healthcare applications.

Keywords: Large language models, domain adaptation, clinical texts, named entity recognition.

1 Introduction

The concept of a "Named Entity" refers to the mention of real-world objects within a text, which can be made using common nouns or proper names. In the context of clinical texts, named entities can encompass references to individuals, locations, diseases, symptoms, medications, treatments, among others.

The ability to identify and extract named entities from clinical texts has become increasingly relevant, as it enables the organization of relevant information within databases, such as electronic health records (EHR). When extracted accurately, this information can be linked to knowledge bases, such as taxonomic, diseases, and pharmaceutics databases. Various applications have been proposed for clinical information databases, including the development of tools for automated

pre-diagnostic assistance [8] and systems for detecting and monitoring medication side effects [17]. More specific applications have also been suggested, such as models for monitoring and predicting suicidal crises in vulnerable patients [13]. These proposals seek to leverage machine learning algorithms to generate knowledge but heavily rely on carefully extracted data for effective training. While these applications hold promise, building models that can automatically recognize, classify, and extract named entities is no trivial task. The challenges largely stem from the inconsistencies in the clinical notes under analysis. These notes, authored by different physicians from various specialties and backgrounds, exhibit an overwhelming diversity of structures, vocabularies, and styles.

Consequently, rule-based model creation becomes prohibitively excessively time-consuming and resource-intensive, requiring many rules that work in concert and require validation by medical experts [4]. Several language models based on machine learning algorithms have made advancements in this field. However, these models, such as Med-BERT [16] and BioBERT [9], have been primarily trained on English medical texts.Additionally, tools have emerged, such as IBM Watson Health, Google Cloud Healthcare API, and MS Azure Text Analytics for Health, that provide complete pipelines for recognizing certain general types of clinical entities in English.

However, progress in developing specific models for clinical texts in the Spanish language has been limited. A key reason behind this disparity is the scarcity of clinical texts in Spanish for training these models; to build systems akin to those in English, an equally substantial dataset would be required. An alternative approach to creating a language model for a specific domain as clinical texts in Spanish is domain adaptation [7]. This technique involves taking a base model pre-trained on a general domain corpus (i.e Spanish texts) and then retraining it with text from a more narrowly defined target domain (i.e., clinical texts). The volume of training data needed in domain adaptation is not as large, since the base model has already learned certain language features during its pre-training [18]. Hence, by starting with a properly pre-trained language model that has been adapted to the desired domain, training a model capable of recognizing entities specific to the clinical domain should yield superior results [5].

In this paper, we investigate the effectiveness of domain adaptation techniques for named entity recognition in clinical texts using pre-trained language models in the Spanish language. We explore the challenges associated with the scarcity of clinical data in Spanish and propose strategies to enhance the performance of named entity recognition models for Spanish clinical texts. The rest of the paper is structured as follows. Section 2 provides a review of the current literature on language models and their performance in biomedical and clinical NLP tasks.

Section 3 details the creation of a clinical notes corpus along with anonymization techniques. Section 4 discusses the selection of RoBERTa-based models and the domain adaptation process. Section 5 explains the setup for evaluation using the Cantemist and PharmaCoNER datasets. Section 6 presents the performance of language models in NER tasks, comparing different base models with domain-adapted models. Section 7 highlights the potential of domain adaptation and outlines future research directions. The paper ultimately offers insights into adapting language models for specific domains and improving NER in clinical texts.

2 Related Work

In the study conducted by [11], a comprehensive set of experiments was carried out using diverse datasets, spanning both scientific and clinical domains, to address common modeling tasks. These included Named Entity Recognition (NER) tasks as well as text anonymization (also referred to as de-identification). In addition, to tasks that involved relation extraction, multi-class and multi-label classification, and Natural Language Inference (NLI)-style tasks.

The authors compared five publicly-available language models to gain a representative understanding of the state-of-the-art in biomedical and clinical NLP. Additionally, they pre-trained new models on their curated corpora and examined key design factors that affect downstream performance in BioNLP tasks. Specifically, they focused on three criteria: i) the impact of model size on downstream performance, ii) the influence of the pre-training corpus on downstream performance, and iii) the significance of tokenizing with a domain-specific vocabulary on downstream performance.

Their experimental approach largely followed the pre-training methodology of [12]. The findings revealed that RoBERTa-large consistently outperformed RoBERTa-base, despite both models having access to the same training corpora. Furthermore, among the publicly available models they experimented with, BioBERT demonstrated the best performance. The newly introduced models in their study exhibited strong performance, achieving superior results on 17 out of the 18 tasks compared to the existing models, often by a significant margin.

Moreover, a team of researchers from the Barcelona Supercomputing Center (BSC) addressed the language gap in Spanish by pre-training two Transformer-based language models from scratch [2]. They compiled biomedical and clinical corpora of various sizes and sources. This included an Electronic Health Record (EHR) corpus containing 95M tokens from over 514k clinical cases, as well as a biomedical corpus with 1.1B tokens across 2.5M documents. The RoBERTa base architecture, featuring 12 self-attention layers, was employed for pre-training, focusing solely on Masked Language Modeling (MLM) using Subword Masking (SWM) as the training objective.

Tokenization was done with the Byte-Pair Encoding (BPE) algorithm, resulting in a vocabulary of 50,262 tokens. The authors produced two RoBERTa models: bsc-bio-es, trained with biomedical resources only, and bsc-bio-ehr-es, which utilized both the biomedical and the EHR corpus. They evaluated the models by fine-tuning for Named Entity Recognition (NER) on three distinct datasets: PharmaCoNER [3], CANTEMIST [14], and ICTUSnet[1].

When comparing their models with various benchmarks —including general domain and domain-specific models— they achieved significant improvements over general-domain models and matched or outperformed domain-specific models across all tasks.

[1] ictusnet-sudoe.eu/es/

3 Pre-Training Corpora

In order to perform domain adaptation on pre-trained models, we compiled a corpus from a set of clinical notes penned by doctors based in Mexico City. This corpus was initially comprised of 86,392 notes taken from Electronic Health Records (EHRs), wich contained records mentioning the Interrogation, Physical Examination, Symptoms, Vital Signs, Auxiliary Studies, Diagnosis, Management Plan, and Treatment Plans of patients.

However, a significant portion of these clinical notes contained irrelevant information; for instance, many fields within the notes were filled with words such as “empty”, “no”, and similar. Moreover, these notes held sensitive information, which must be anonymized to be used in the training of a model without jeopardizing the privacy of those involved [6]. It should be noted that around 98% of these notes originate from emergency services within Mexican hospitals, a factor that merits consideration when analyzing the performance of models trained with this data.

Consequently, we designed a pre-processing algorithm comprising several steps. Initially, we eliminated all notes that solely contained vital signs measurements. Additionally, we removed discharge and follow-up notes that contained repeated or redundant text to circumvent potential memorization issues [10].

To anonymize any names found within the notes, we developed an algorithm that consistently substitutes these. We retrieved names from both the ‘patient’s name’ and ‘doctor’s name’ fields in the clinical notes database, and subsequently generated sets from these. We supplemented these sets using lists of common Spanish names provided by public statistical services institutes in both Mexico and Spain. These produced sets distinguished between predominantly male and female names, and also included a category for gender-neutral names. The resulting anonymization sets included 468 male names, 563 female names, 97 gender-neutral names, and 1130 surnames. The discrepancy in the number of names and surnames can be attributed to the fact that compound names were split and then handled individually.

Following the creation of the names sets, we begin the anonymization process. The algorithm processes each note, normalizing it to lowercase and stripping away accents and special characters. It then seeks exact matches with each name and surname present in the names sets. If a match is found, it is substituted, preserving the accents and special characters and adapting it to match the original case. For each name that is matched, the substitution is recorded and applied consistently to any subsequent matches in the same note. Additionally, to account for possible spelling errors, a search using Levenshtein distance is performed for the correct patient’s and doctor’s names in the original database. Moreover, the algorithm to generate random names is designed not to repeat names within the same note.

As part of our analysis of the clinical notes, we found instances where doctors had recorded sensitive personal information, such as the patient’s address, phone number, and other potentially sensitive data. Consequently, any lines of text containing phone numbers, email addresses, and physical addresses were eliminated. We also masked any alphanumeric codes that could potentially be identifiers, such as patient numbers, professional IDs, and social security numbers.

At the end of this process, we had amassed a corpus of 78,616 unique and anonymized clinical notes. We identified and anonymized 47,237 names and 63,982 surnames, with at least one present in 80% of the notes. We discarded 244 phone numbers and alphanumeric codes, as well as 1,697 addresses. The final token count for the retraining process is 39,680,624.

4 Domain Adaptation of Models

To generate an adapted language model tailored to the domain of Mexican clinical notes, we chose models based on the RoBERTa architecture as a starting point. This architecture is preferred for domain-specific model training because it facilitates the learning of new vocabularies through the Byte Pair Encoding (BPE) tokenization algorithm [15]. Specifically, we started by selecting the most prominent language model for the biomedical domain in Spanish as our base-model: the RoBERTa-based bsc-bio-ehr-es[2]. This model achieves the best results for Named Entity Recognition (NER) tasks in the Spanish biomedical domain using a simple fine-tuning architecture [2].

Even though the BSC model is trained entirely on biomedical texts, only 16.17% of the training corpus belongs to the clinical domain. Out of this percentage, 48% of the tokens belong to clinical notes; and the remaining tokens in the clinical domain are comprised of 'clinical cases', which fall within the same domain of knowledge but differ slightly in writing style. These differences in writing style, vocabulary, and carried-out medical procedures suggest that enriching the training process with the Mexican clinical notes corpus could enhance model performance for tasks specifically dealing with such texts. This approach is advantageous not only for addressing the unique attributes present in Mexican clinical notes but also for capturing the specific nuances inherent to Spanish-language clinical text.

To get the Mexican clinical notes ready for training, we set up a careful preparation process. First, we split the notes into batches of 512 tokens each. Then, we performed the tokenization using the BPE tokenizer from the BSC model trained with biomedical text. Each batch was generated following the full-sentence scheme described in the original RoBERTa article [12], in which the batches are filled with tokens from complete sentences regardless of crossing the document boundary. In total, 39,680,624 tokens from 78,616 notes were obtained for model training. The original training parameters from the BSC model were maintained, except for the batch size, which was reduced to 32 samples. No warm-up steps were performed; we started training with a rate of 5E-5 which was linearly decreased to 0.

The training was conducted over five epochs, saving a checkpoint at the end of each epoch to track the progress of the model. The retraining process took approximately 4 hours on two RTX-A5000 cards. Additionally, we carried out the same retraining process using the Mexican clinical notes corpus in two additional separate instances: first by leveraging a model[3] that had been previously fine-tuned from the original BSC model using the Chilean Waiting List (CWL) corpus [1]; and second, starting the

[2] huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es

[3] huggingface.co/plncmm/roberta-clinical-wl-es

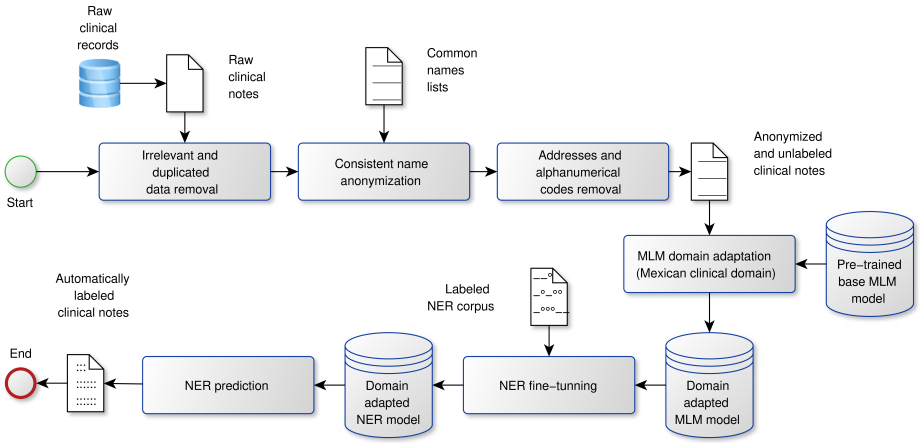

Fig. 1. General algorithm of the paper's code implementation.

multilingual general-purpose model, xlm-roberta-base[4]. Additonally to this description, in Figure 1, we show the complete pipeline of both the retraining of the MLM model using Mexican clinical notes, and the NER training step for evaluating the model.

5 NER Task and Datasets

We evaluated the performance of the domain-adapted language models on two Clinical Named Entity Recognition (NER) task. The evaluation was carried out using publicly available tagged datasets, the PharmaCoNER set[5] (Drugs and chemical entities) and the Cantemist set[6] (Tumors). This evaluation datasets were chosen due to the similarity of the texts they contain to the texts we used to retrain the model on the MLM task.

However, it is important to note there are still notable differences between the domains of the retraining texts and the NER evaluation datasets. While the retraining was grounded in a corpus of Mexican clinical notes primarily derived from real emergency episodes in hospitals — generally characterized by a certain urgency and spontaneity in documentation — the NER datasets exhibit more structured, meticulously crafted notes stemming from a diverse array of medical specialties.

This difference underscores a potential variation in complexity and style between the training and evaluation materials, and should be considered in the interpretation of the evaluation results. The PharmaCoNER set consists of 1,000 clinical cases written in Spanish and extracted from the Spanish Clinical Case Corpus (SCCC)[7], from a variety of clinical areas such as oncology, urology, cardiology, pneumology, among others. These clinical cases are annotated with 4 categories of entities: Normalizable Chemicals, Non-normalizable Chemicals, Proteins, and Others.

[4] huggingface.co/xlm-roberta-base
[5] huggingface.co/datasets/PlanTL-GOB-ES/pharmaconer
[6] huggingface.co/datasets/PlanTL-GOB-ES/cantemist-ner
[7] github.com/PlanTL-SANIDAD/SPACCC

Table 1. Results of the NER model trained with the Cantemist dataset. Precision (P), Recall (R), and F1 (F) scores on each dataset are reported. The best scores are in bold.

Cantemist			
Model	**P**	**R**	**F**
bsc-bio-ehr-es	0.8142	0.8581	0.8356
bsc-bio-ehr-es + cwl	0.8280	**0.8570**	0.8423
bsc-bio-ehr-es + ours	**0.8295**	0.8565	**0.8427**
bsc-bio-ehr-es + cwl + ours	0.8194	0.8509	0.8349
xlm-roberta-base	0.7958	0.8467	0.8205
xlm-roberta-base + ours	0.7995	0.8373	0.8179

Normalizable Chemicals refer to chemicals that can be linked to the SNOMED-CT knowledge base. The corpus contains 7,624 tagged entities in total. The Cantemist set consists of 1,301 clinical cases extracted from SCCC but exclusively from the oncology area. These clinical cases have been manually annotated by specialist doctors, identifying all mentions of cancerous tumors in one category: tumor morphology.

The dataset is balanced considering the age and gender of the patients, as well as the types of tumors. The corpus contains 16,030 tagged entities in total. We used a simple architecture to generate the NER model from the pre-trained RoBERTa model. A linear classification layer and a Softmax function were added after the output of the original model's hidden states. This model was trained for 10 epochs with an initial learning rate of 5E-5 and a linear decay rate. The batch size for training was set to 16 samples.

6 Evaluation and Results

Performance of the trained language models was evaluated in the NER task with the Cantemist and PharmaCoNER datasets. In addition, results from the base BSC language model (bsc-bio-ehr-es), the base XLM-RoBERTa (xlm-roberta-base), and the model enriched with the CWL notes (bsc-bio-ehr-es + cwl) are included for comparison. To measure the effectiveness of these NER models, we employed three universally acknowledged scores for NER tasks: precision, recall, and F1 score.

Precision indicates the proportion of correctly identified entities among all the entities that the model labeled, although it disregards the false negatives. In contrast, recall accounts for the proportion of correctly identified entities out of all actual entities, overlooking false positives. The F1 score is the harmonic mean of precision and recall, serving as a single metric that reflects a good balance of both.

All of these metrics were obtained using the micro-averaging method, thus giving equal weight to each entity in the final scores. These results can be found in Table 1 and Table 2, where the highest scores for each metric are highlighted to facilitate the comparison between models. The results in Table 1 show a slight improvement in the model's performance in the Cantemist task, which deals with finding mentions of cancerous tumors.

Table 2. Results of the NER model trained with the PharmaCoNER dataset. Precision (P), Recall (R), and F1 (F) scores on each dataset are reported. The best scores are in bold.

PharmaCoNER			
Model	**P**	**R**	**F**
bsc-bio-ehr-es	**0.8921**	0.9081	0.9000
bsc-bio-ehr-es + cwl	0.8703	0.9119	0.8906
bsc-bio-ehr-es + ours	0.8826	0.9119	0.8970
bsc-bio-ehr-es + cwl + ours	0.8876	**0.9146**	**0.9009**
xlm-roberta-base	0.8541	0.8788	0.8663
xlm-roberta-base + ours	0.8651	0.8891	0.8769

The model that stands out in this task is the one that was retrained with our dataset from the base BSC model (bsc-bio-ehr-es + ours) which presents a slight improvement in the precision score and F1 score. On the other hand, the bsc-bio-ehr-es + cwl model improves the performance slightly the recall score and in general is the best model compared to the base bsc-bio-ehr-es model. However, we note that when this model is subjected to an additional retraining stage (bsc-bio-ehr-es + cwl + ours), the performance drops slightly.

For this task, the worst models were the xlm-roberta-base and xml-roberta-base + ours. On the other hand, in the results described in Table 2 of the PharmaCoNER task, the performance slightly decreases when the base model is retrained, independently of the corpora. However, with the language model that has went through two sequential adaptations (bsc-bio-ehr-es + cwl + ours), the NER performance gets a minor uplift. One hypothesis for the drop in performance is that in the PharmaCoNER clinical cases, chemicals such as proteins are mentioned, which are not mentioned frequently within the clinical notes of our dataset.

Interestingly, this decline in performance dissipates when training incorporates data from both additional sources. Another reason is the clinical notes used for training our language model primarily belong to emergency services (98.54%) and the rest of the services (1.46%) are underrepresented. Besides, these notes lack normalization regarding typographical error correction (incorrect typing, omissions, duplications, inversions, replacements, white space omissions, accent omissions, and punctuation errors). We drop normalization at this particular point due to two reasons.

The first reason is the complexities of the medical domain, where technical terms, acronyms, abbreviations, and specialized jargon are prevalent, and addressing these linguistic intricacies requires expert knowledge. The second and more relevant reason is that we intend to use the trained LM to address downstream tasks with the same type of data, i.e., clinical medical notes direct from the real world.

Regarding the retraining of the xlm-roberta-base model, we observe a modest performance boost in the PharmaCoNER task when adapted to our corpus domain; though the results in the Cantemist task do not exhibit the same consistency. These fluctuations could potentially be attributed to the changes in the vocabulary size introduced by the multilingual model.

The vocabulary size of the xlm-roberta-base tokenizer is nearly five times larger than that of the BSC tokenizer, and has been trained on text spanning over 100 languages. As such, there's a compelling case for additional experimentation with models of this magnitude to uncover more detailed insights.

7 Conclusions and Future Work

This research paper analyzes the efficacy of domain adaptation techniques to enhance named entity recognition (NER) in Mexican Spanish clinical texts using pre-trained language models. Even though the downstream tasks for evaluation were not specifically tailored using the same specific domain (i.e. Mexican Clinical Notes), we observed an improvement in the performance on the NER tasks. This highlights the value of customizing MLMs to align them to the unique characteristics and nuances of the target domain.

The approach leverages RoBERTa-based language models and a corpus of Mexican clinical notes, which are anonymized to protect sensitive information. The study finds that domain adaptation through retraining on clinical notes can lead to a slight improvement in NER performance. Specifically, the model retrained with the generated clinical notes corpus shows promising results in recognizing cancerous tumor mentions in the Cantemist dataset.

For the PharmaCoNER dataset, an increased recall hints at the model's enhanced ability to extract more entities, potentially leveraging richer context from the clinical domain, albeit at the expense of precision; a decrease in the latter might suggest that while the model has become adept at analyzing clinical texts, it may still lack sufficient training data representing specialized biochemical substances, which are less frequently mentioned in the emergency notes dominant in our corpus.

The results suggest the potential of domain adaptation for refining language models to specific clinical domains and demonstrate the significance of sufficient domain-specific data for optimal performance. Furthermore, we examined the retraining of a large language model, specifically the xlm-roberta-base model, but the results were slightly inconsistent. This may be due to the considerable change in size from the model's pre-training corpus.

Future research directions are highlighted to explore more sophisticated domain adaptation techniques and to further refine the model's capabilities for NER in clinical texts. For the validation phase, we want to use downstream tasks based on texts from the specific domain we adapted, i.e. Mexican clinical notes. We also want to generate more diverse sets of evaluation tasks, both in the types of entities in NER, and in the variety of downstream tasks.

Additionally, we aim to conduct experiments on performance variations when adapting larger, multilingual models like the xlm-roberta-base model. Finally, as we see more AI systems being developed for healthcare applications, we must continue creating better ways to protect sensitive data. And as such, we want to continue improving our anonymization algorithms to ensure that healthcare data is used responsibly in our AI research.

Acknowledgments. This work has been carried out with the support of DGAPA UNAM-PAPIIT project number TA101722, support of Secretaría de Educación, Ciencia, Tecnología e Innovación de la Ciudad de México (Resource Allocation Agreement SECTEI/201/2021) in collaboration with Secretaría de Salud de la Ciudad de México (SEDESA), DGAPA-UNAM postdoctoral scholarship, and the CONAHCYT scholarship program (CVU: 1148113). The authors also thank CONAHCYT for the computing resources provided through the Deep Learning Platform for Language Technologies of the INAOE Supercomputing Laboratory.

References

1. Báez, P., Villena, F., Rojas, M., Durán, M., Dunstan, J.: The chilean waiting list corpus: A new resource for clinical named entity recognition in spanish. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop, pp. 291–300 (2020) doi: 10.18653/v1/2020.clinicalnlp-1.32
2. Carrino, C. P., Llop, J., Pàmies, M., Gutiérrez-Fandiño, A., Armengol-Estapé, J., Silveira-Ocampo, J., Valencia, A., Gonzalez-Agirre, A., Villegas, M.: Pretrained biomedical language models for clinical NLP in spanish. In: Proceedings of the 21st Workshop on Biomedical Language Processing, pp. 193–199 (2022) doi: 10.18653/v1/2022.bionlp-1.19
3. Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., Krallinger, M.: Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pp. 1–10 (2019) doi: 10.18653/v1/D19-5701
4. Gorinski, P. J., Wu, H., Grover, C., Tobin, R., Talbot, C., Whalley, H., Whiteley, W., Alex, B.: Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches. In: Second UK Healthcare Text Analytics Conference (2019) doi: 10.48550/arXiv.1903.03985
5. Grangier, D., Iter, D.: The trade-offs of domain adaptation for neural language models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 3802–3813 (2022) doi: 10.18653/v1/2022.acl-long.264
6. Ishihara, S.: Training data extraction from pre-trained language models: A survey. In: Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing, pp. 260–275 (2023)
7. Kalyan, K. S., Rajasekharan, A., Sangeetha, S.: AMMU: A survey of transformer-based biomedical pretrained language models. Journal of Biomedical Informatics, vol. 126 (2022) doi: 10.1016/j.jbi.2021.103982
8. Latif, J., Xiao, C., Tu, S., Rehman, S. U., Imran, A., Bilal, A.: Implementation and use of disease diagnosis systems for electronic medical records based on machine learning: A complete review. IEEE Access, vol. 8, pp. 150489–150513 (2020) doi: 10.1109/ACCESS.2020.3016782
9. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., Kang, J.: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, vol. 36, no. 4, pp. 1234–1240 (2020) doi: 10.1093/bioinformatics/btz682
10. Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., Carlini, N.: Deduplicating training data makes language models better. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 8424–8445 (2022) doi: 10.18653/v1/2022.acl-long.577

11. Lewis, P., Ott, M., Du, J., Stoyanov, V.: Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop, pp. 146–157 (2020) doi: 10.18653/v1/2020.clinicalnlp-1.17
12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach (2019) doi: 10.48550/arXiv.1907.11692
13. Metzger, M. H., Tvardik, N., Gicquel, Q., Bouvry, C., Poulet, E., Potinet-Pagliaroli, V.: Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: A French pilot study. International Journal of Methods in Psychiatric Research, vol. 26, no. 2 (2016) doi: 10.1002/mpr.1522
14. Miranda-Escalada, A., Farré-Maduell, E., Krallinger, M.: Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In: Proceedings of the Iberian Languages Evaluation Forum, pp. 303–323 (2020)
15. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
16. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine, vol. 4, no. 1, pp. 1–13 (2021) doi: 10.1038/s41746-021-00455-y
17. Shinozaki, A.: Electronic medical records and machine learning in approaches to drug development. Artificial Intelligence in Oncology Drug Discovery and Development (2020) doi: 10.5772/intechopen.92613
18. Wiese, G., Weissenborn, D., Neves, M.: Neural domain adaptation for biomedical question answering. In: Proceedings of the 21st Conference on Computational Natural Language Learning, pp. 281–289 (2017) doi: 10.18653/v1/K17-1029