WJAETS

World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Research Article)

Check for updates

Medical Data NER and Classification Using Hybridized BERT Model

G. Jothi [1, *] and S. Clement Virgeniya [2]

[1] *Department of Computer Science, Dr. Umayal Ramanathan College for Women, Tamil Nadu, India.*
[2] *Adjunct Faculty, Department of Computer Science, Alagappa University, Tamil Nadu, India.*

World Journal of Advanced Engineering Technology and Sciences, 2024, 13(01), 040–047

Publication history: Received on 16 July 2024; revised on 30 August 2024; accepted on 01 September 2024

Article DOI: https://doi.org/10.30574/wjaets.2024.13.1.0376

Abstract

The extraction of important information from medical texts by Named Entity Recognition (NER) is a key component of advanced medical text processing. Medical practitioners rely heavily on NER's assistance with disease surveillance, clinical resolution building, and substantiation-based treatment. As the foundation of text information processing in the medical field, it guarantees precise location of data required for knowledgeable medical decisions and attentive disease surveillance. Additionally, a core goal in medical Natural Language Processing (NLP) is medical text categorization, which tries to classify short medical texts into distinct groups. Most recent work has concentrated on using pre-trained linguistic processes for text cataloging in medicine. The present work presents a novel clinical neural network architecture (NER) method that was created with a customized Rule Based BiLSTM-BERT (Bidirectional Encoder Representations from Transformers) architecture that incorporates Retrieval Augment Generation. Across several fields, including medicine, deep learning has demonstrated noteworthy advancements. These results show that, when applied to our test dataset, the BiLSTM-BERT-RAG model produced results that were almost human-like. The system proficiently recognized pertinent vocabulary representative of the intended protocol.

Keywords: Name Entity Recognition; Relation Extraction; BiLSTM; BERT; RAG

1. Introduction

Essential medical information, such as symptoms, prescriptions related to an illness, and diagnoses, can be discovered in medical texts, which can be obtained in places like electronic health check accounts and medicinal writing [2]. Medical information research is increasingly impacted by natural language processing (NLP) due to the laborious process of obtaining such expertise by human efforts [1], [21]. 21] Within this framework, Medical Text Classification (MTC) organizes brief medical texts into categories such as organ condition, disease phase, and allergy intolerance; thus, MTC is essential to Medical NLP. These categorization outcomes have a significant effect on subsequent tasks including determining challenging corrective trials and creating Clinical Decision Support Systems (CDSS).

Regarding MTC [10], previous research has highlighted that the effectiveness of machine learning and statistical methods primarily depends on the value of characteristic engineering. Conversely, deep learning mechanisms such as Convolutional Neural Networks (CNN) [7][20] or Recurrent Neural Networks (RNN) demonstrate superior performance without requiring a great deal of physical characteristic choice. That being said, regardless of the method selected, the main objective is still to create representations with conceptual meanings and then predict categories based on these representations. Since these mechanisms can only hold the semantic facts in the training data, their effectiveness is limited by the availability of labeled training datasets. As a result, the key to improving MTC performance

The significance of pre-training language models (PLMs), namely Enhanced Language Representation with Informative Entities (ERNIE), BERT [9], and ELECTRA, is highlighted in a significant amount of recent NLP research [3], [4], [8].

[*] Corresponding author: G. Jothi.

Researchers have studied these models in great detail and suggested that PLMs learn substantial amounts of prior semantic knowledge by using methods such as Masked Language Model (MLM) and Next Sentence Prediction (NSP) [6]. In order to apply this knowledge to Medical Text Classification (MTC) [5], researchers first added more categorizers to PLMs and then used a specific principle function to fine-tune both building blocks.

Natural Language Processing (NLP) is a key field of research in the ever-evolving field of artificial intelligence development [23]. Artificial intelligence technologies are slowly becoming important in many aspects of modern life and are revolutionizing how people live [1–5]. In this context, named entity recognition becomes an important element, especially with regard to proper names and phrases. Named entity identification forms the basis for many downstream operations, such as machine translation, automated question answering, knowledge graph building, and information extraction. Identified entities usually relate to unique entities with explicit references in broad text material that is not associated with any particular domain. But named entities often include things like genes, illnesses, and medications in specialist domains like medical data. Based on a number of common principles, a named entity identification system performs sequence labeling tasks to extract these entities from unstructured, unlabeled text [6–10].

This work combined the Bidirectional LSTM with the Rule Based NER in order to provide increased care disease, medications and dose entity recognition, medical notion mining, and clinical data classification. The BERT model was used to classify the NER, and the RAG model was well-tuned in this work. In order to train the Rule Based Bi-LSTM model to generate more intelligent and skilled word vectors while avoiding actual buildings, the dataset from the physical condition ground was applied.

Using a discriminative PLM in a RAG tuning framework, this paper explores a novel approach to medical text classification. It suggests that the RAG technique has potential to improve the effectiveness of pre-training models [3] in tasks particular to particular domains, which offers insightful information for researchers. Additionally, as tuning approaches are balancing in nature, the research focuses mostly on them. Here are a summary of this study's primary contributions: The remaining of this effort is planned as follows: Section 2 explains related works on EHR Analytics, section 3 illustrate the methodology of the proposed work's process in detail, section 4 argue the investigational outcome and discussion, and section 5 sum up the proposed work's conclusion.

2. Related Studies

NER, NLP, and NN's foundational principles are explained in this section of the book. The study then gives an overview of the recommended procedure for relationship extraction and named entity recognition. Valuable insights are automatically extracted from textual data using this methodology. Finally, information extraction from biomedical literature is addressed in this article, which provides an overview of the issues and solutions that are currently being discussed. Many particular entities are covered in large text collections such as scholarly papers and medical records in the realm of life sciences and health [12]. Scholars are putting out methods for NER to make the most of these materials. They are studying the effectiveness of sophisticated masked language models, especially those that use transformer architecture, when trained on various health and life science literature in the languages of French and English that cover biology, medicine, and chemistry. These models are trained first on a variety of text sources, and then they are refined with particular datasets from every language and area.

After that, conventional voting methods are employed to group the models together. These trials' outcomes show a significant improvement over BERT-based models in the way assembly models are presented, with an ideal 77% macro F1-score overall performance. A detailed study of the ensemble results is also carried out, which reveals differences in effectiveness according to the characteristics of the entities, including the frequency, length, and stability of the annotations in the corpus. Based on a variety of health and life science books, these results imply that ensembles of advanced masked language models provide a potent solution to Named Entity Recognition problems. Biomedical text mining is becoming increasingly important, as evidenced by the volume of biomedical documents that is expanding [26]. Researchers can't wait to use recent advances in natural language processing (NLP) to mine biomedical text for invaluable insights. The development of useful models for biomedical passage mining has been markedly expedited by deep learning approaches. However, because of the change in expression division from universal area corpora to biological ones, immediately applying NLP progression to this sector sometimes presents difficulties.

This paper investigates how the recently-established pre-trained language mechanism BERT can be tailored to biological corpora. Bidirectional Encoder Representations from Transformers for Biomedical Text Mining, or Bio BERT, is a field exact language model that has been pre-trained on a large number of biomedical datasets. Similar in construction to BERT, Bio BERT is tailored exclusively for biomedical texts and routinely outperforms previous state-of-the-art models as well as BERT on a range of biomedical passage mining tasks.

Although BERT performs similarly to earlier models, Bio BERT performs significantly better in three critical biomedical text mining tasks: 0.62% better at biomedical named entity detection, 2.80% better at biomedical relation extraction, and 12.24% better at biomedical query answering. Our findings highlight the huge improvement in BERT's ability to produce composite biomedical texts that results from pre-training it on biomedical data.

This research used deep learning techniques and linguistic models to automate the diagnosis prediction process based on symptoms [11]. The effectiveness of two iterations of the Medical Concept Normalization Bidirectional Encoder Representations from Transformers (MCN BERT) models in conjunction with a BiLSTM model was especially evaluated in this research. For syndrome prediction from symptom photos, each model was fine-tuned using a unique hyperparameter optimization technique. Chaichulee et al. [17] evaluated three Natural Language Processing (NLP) techniques in their research: many pre-trained BERT models, Naive Bayes-Support Vector Machine (NB-SVM), and Universal Language Model Fine-tuning (ULMFiT) using medication allergy reaction documents. Finding the symptoms and their medical causes was the main goal. When it came to less often reported symptoms, the NB-SVM model performed better than both BERT and ULMFiT, even though the BERT models were generally better at performing. Using an ensemble model to combine these algorithms produced remarkable outcomes, with the best performance being attained for the 36 most common symptoms. This improved model was then implemented into a symptom term suggestion system that performed well in prospective pharmacist trials. A reasonably good degree of contract between the model's implications and pharmacist measurements was indicated by the trials' 0.7081 Krippendorf's alpha agreement coefficient.

The classification of brief medical texts into pertinent categories is a crucial task in the field of medical natural language processing [13]. The majority of recent research has been on using language models that have already been trained and tailored for this particular goal. Conversely, this approach defines additional parameters while training more classifiers. The prompt-tuning technique may improve performance on a variety of natural language processing tasks, according to recent study, by bridging the gap between pre-training objectives and subsequent challenges.

The CAC model is a text categorization approach presented by Yang et al. [16] that fuse multi-layer characteristics by combining CNN and concentration methods. This model removes specific properties and computes overall concentration, drawing inspiration from membrane computing. Experiments' results show that the CAC model performs better than models that only use concentration, with notable gains in correctness and overall presentation as compared to auxiliary models. In a novel investigation into this area, our work uses prompt-tuning to investigate the classification of medicinal passage using a discriminative pre-trained language model named ERNIE-Health [14]. To improve performance, prompt-tuning leverages the learned features of pre-trained language models to transform dual or multi-categorization tasks into mask prediction tasks. Primarily, the author uses prompt-tuning based on the ERNIE-Health pre-training task—multi token collecting. The initial text is transformed into the most recent iteration that follows a pattern in which a [UNK] token is used in place of the class label.

Afterward, the model is trained to forecast the probability distribution of possible groups. Lee et al. [18] and KafKang et al. [15]. developed a novel technique for the identification and categorization of drug-drug interactions (DDIs) that makes use of Relation Bio BERT (R-Bio BERT) and BiLSTM models [14]. Their method effectively identifies different types of medication.

3. Proposed Methodology

Big Data analytics in healthcare makes it easier to handle Electronic Health Records (EHR) for patient care and healthcare administration by utilizing cutting-edge technologies [19]. The goal of this study is to examine how big data analytics may be used in the medical field. Previous research indicates that integrating big data analytics can benefit healthcare organizations in a number of ways by allowing them to analyze both structured and unstructured data for administrative, business, and clinical elements. This emphasizes how medical institutions base their decisions on data-driven insights.

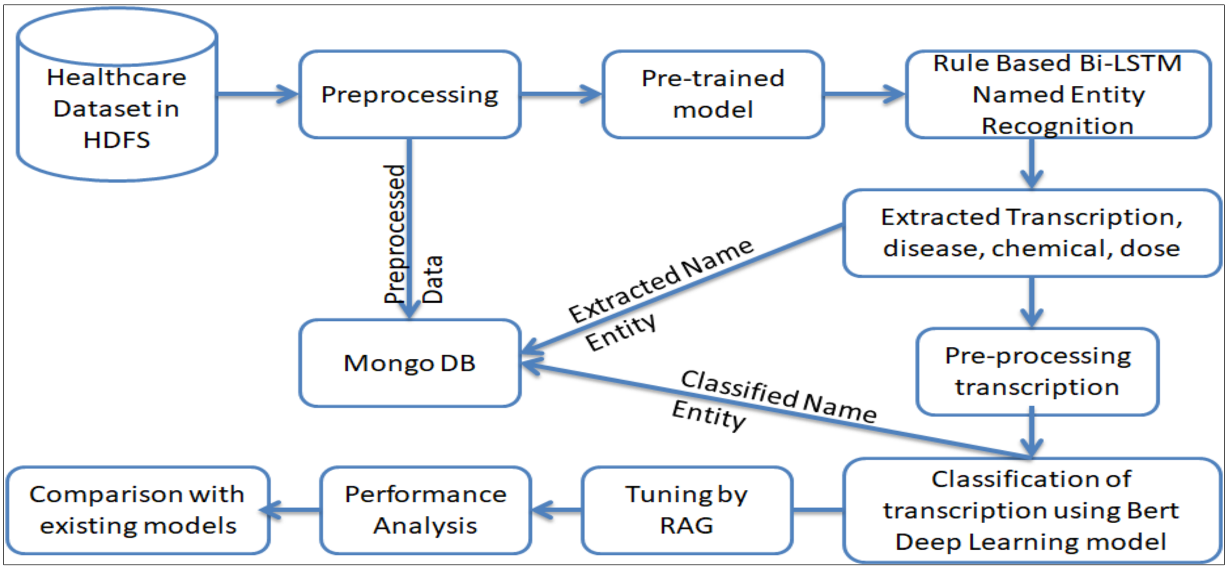
Healthcare Dataset in HDFS
Preprocessing
Pre-trained model
Rule Based Bi-LSTM Named Entity Recognition
Preprocessed Data
Extracted Transcription, disease, chemical, dose
Extracted Name Entity
Mongo DB
Pre-processing transcription
Classified Name Entity
Comparison with existing models
Performance Analysis
Tuning by RAG
Classification of transcription using Bert Deep Learning model

Figure 1 Architecture of the Proposed Work

The hybrid model presented in this study combines three different elements: RAG, BERT, and Rule Based BiLSTM. The model first generates rules for Named Entity Recognition (NER) using rule-based approaches. The relationships between these items are then extracted at the input layer using BiLSTM. Next, the pre-trained BERT is used to classify the Named Entities in order to obtain word embeddings containing semantic information. The final results are then run through the RAG layer, which produces an automated and optimized NER and classification model.

An enormous database of transcribed medical reports covering a wide range of medical disciplines can be accessed through the website MTSamples.com. A wide range of specializations and job types are covered by the dataset that may be found on MTSamples.com for example transcribing reports. 40 distinct specializations and areas are covered by the website's medical transcribing samples and reports. Due to the fact that Electronic Medical Record (EMR) data is frequently noisy, incomplete, and inconsistent, steps are made to improve the quality of the data by eliminating noise, adding missing data, and correcting inconsistent data.

Rather than just single sentences, medical text data frequently consists of paragraphs containing many sentences. For sentences and paragraphs, Named Entity Recognition (NER) is used. Sentence segmentation is used to split the text into sentences at the start of the process. Every sentence is then subjected to additional segmentation, preprocessing, and analysis in two main stages.

Using a combination of Natural Language Processing (NLP) techniques, entity mentions are found and recognized in the first stage [25]. The second phase focuses on picking and extracting these entities using the previous phase's insights and a set of rules that control the extraction of pertinent entities. Potential entity mentions usually appear as noun phrases, which are phrases with linguistic meaning. But occasionally, in an effort to increase the significance of identified entity mentions, additional noun phrases or combinations of noun phrases with other morphological phrases may be included in entity mentions. Rule-based NER techniques are used after preprocessing, using common terms or language details that are specific to the unique characteristics of the entities of interest. Nevertheless, the manual rule construction process, which is time-consuming and dependent on the domain, is a major drawback of these methods. The detection accuracy of medical named entities is typically compromised by standard pretrained models, especially when it comes to expressing word polysemy. This results in reduced accuracy of detection in electronic medical records. The work proposes a new model that combines two well-known model-based techniques, BiLSTM and CRF, to address this problem. Although these techniques are capable of strong generalization, they require a large amount of labeled data.

There is a lot of data available because of the exponential increase in electronic medical records due to growing medical needs. This pattern is in line with the quick advancement of deep learning in a variety of medical domains, including drug development, medical image analysis, disease detection, and clinical decision-making [1-3]. Therefore, there is great potential in investigating the use of deep learning techniques to automated illness coding, multidata source integration analysis, public health, and related fields.

This study heavily relies on BERT, a framework for training deep bidirectional transformers. Three pretraining activities are included in BERT, which serves as a general language model. To train the language model for the first challenge, randomly select words that are masked within a sentence and use that mask information. A task to predict whether two input paragraphs form continuous text is introduced in the second task, which is a sentence-level continuity prediction task. With this addition, the model's comprehension of the connections between uninterrupted text portions is improved.

The work presents a novel Rule-Based Bi-LSTM-BERT_RAG NER approach for the extraction of disease, chemical, and medicine dose information. This method is based on a small set of rules that are not tied to entity qualities, unlike the typical rule-based NER systems that depend on entity characteristics. This is a calculated move to avoid having to go through the tedious process of creating rules for every unique entity of interest. Furthermore, the paper suggests a feature fusion framework based on BERT to solve the problem of text semantic feature representation in short texts, which are characterized by a small vocabulary and sparse characteristics. Still, the NER model is refined with the help of Retrieval Augmented Generation (RAG) in order to increase efficiency. By incorporating external knowledge sources, this method takes a big step forward and is especially useful for tasks requiring a lot of information.

4. Results and Discussion

Using rule-based approaches, named entities are identified from medical texts and assigned rules that are valid for certain tasks but not for others. They are coupled with the Bi-LSTM model for Named Entity Recognition (NER) and Relation Extraction (RE) in order to address this. Table 1 provides specifics on NER and RE's performance in terms of accuracy measures. The next step in this study is to classify these named things after they have been identified and relations have been extracted. A performance analysis is used to assess how well this classification task performed. Universal language representations can be learned effectively through pre-training language models. Modern language model pre-training techniques like BERT have shown impressive outcomes on a range of language understanding challenges. A thorough solution for BERT tuning is presented in this study, which involves conducting extensive experiments to investigate various RAG tuning approaches of BERT for the text categorization job. Finally, the suggested methodology outperforms new state-of-the-art findings on recently widely used passage classification datasets.

Table 1 Performance Analysis Report for NER & Classification

S. No	Models	Fine-Tuning	Prompt-Tuning	RAG-Tuning
1.	Rule Based Bi-LSTM	0.854	0.871	0.890
2.	Rule Based Bi-LSTM+BERT	0.912	0.937	0.945
3.	Rule Based Bi-LSTM+BERT+RAG	0.953	0.966	0.983

For the evaluation purpose the proposed work, fine tuning and prompt tuning are compared with proposed RAG tuning task. The performance results reported that RAG is performed well than other two tuning.

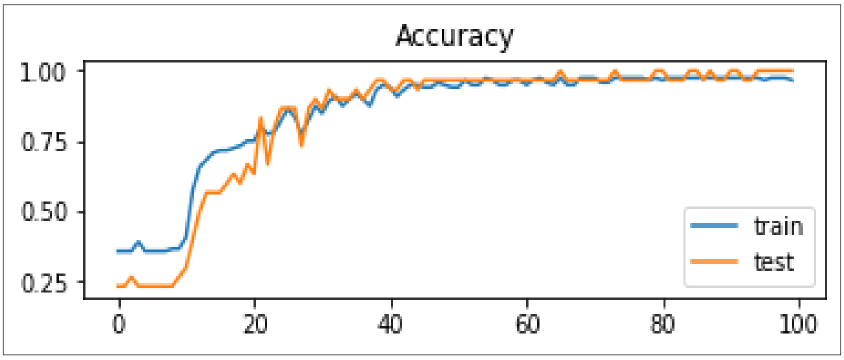

Figure 2 Accuracy of Rule Based Bi-LSTM

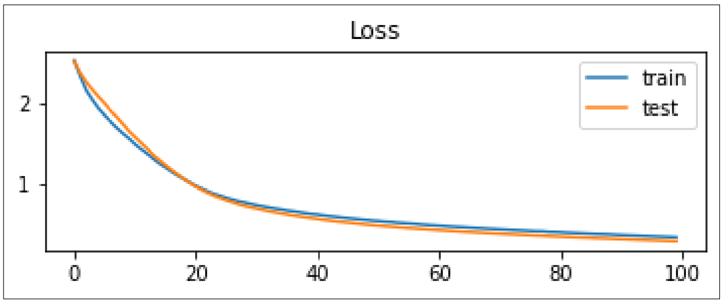

Figure 3 Loss of Rule Based Bi-LSTM

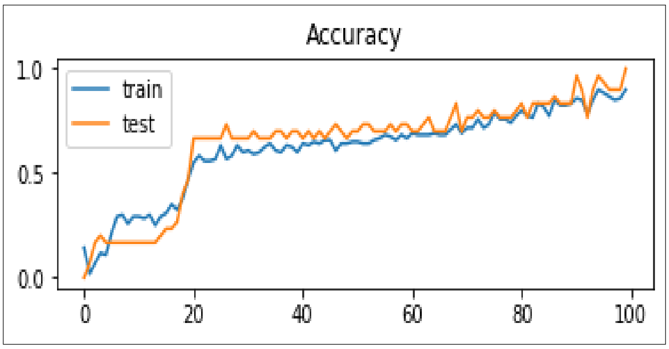

Figure 4 Accuracy of Rule Based Bi-LSTM + BERT

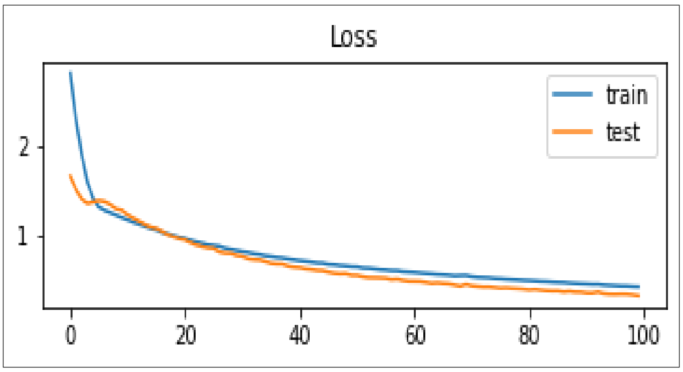

Figure 5 Loss of Rule Based Bi-LSTM + BERT

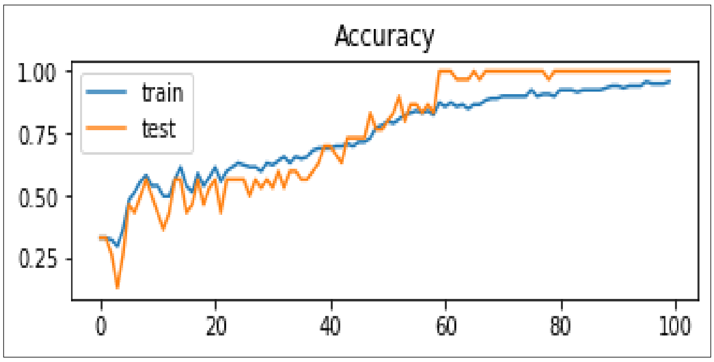

Figure 6 Accuracy of Rule Based Bi-LSTM + BERT+RAG

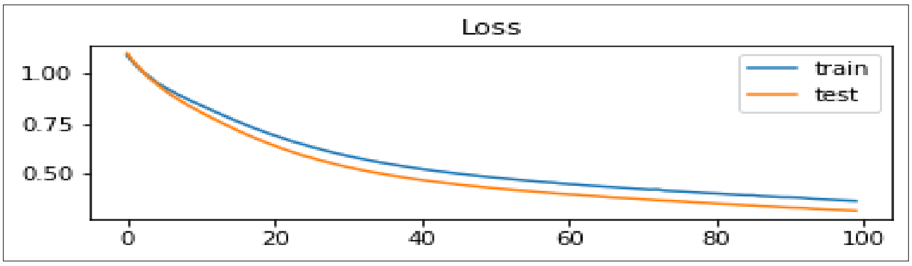

Figure 7 Loss of Rule Based Bi-LSTM + BERT+RAG

Table 2 Performance Analysis Comparison of NER & Classification

Authors	Year	Existing Work	Technique	Metrics	Value
Jinhyuk Lee	2020	[11]	Bio-BERT	Recall	94.29%
Nona Naderi	2021	[12]	BERT	F1-score	85.00%
Yu Wang	2023	[2]	ERINE-Health	Accuracy	86.66%
Esraa Hassan	2024	[13]	MCN-BERT	Accuracy	97.03%

5. Conclusion

Despite medical professionals' best efforts to provide accurate diagnoses and treatments, recommendations frequently rely on subjective clinical experiences, increasing the risk of misdiagnoses and overlooked conditions. A Medical Decision Support System empowers healthcare providers to receive guidance on treatment strategies grounded in factual data; further development and implementation of this system could greatly aid medical professionals in diagnosing diseases, especially those with less clinical experience.

Medical Decision Support and Disease Risk Prediction demand significant effort to ensure that physicians are well-informed about patients' treatment plans. A large number of healthcare facilities are actively promoting the use of medical decision support systems and enhancing their hospital information systems. Furthermore, the creation of risk prediction models can help doctors assess the chance of a condition worsening or getting better, improving patient treatment even in situations where resources are scarce. In addition, individuals can make knowledgeable choices about acquiring health insurance to reduce medical costs.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than English: opportunities and challenges. Journal of biomedical semantics, 9, 1-13.

[2] Wang, Y., Wang, Y., Peng, Z., Zhang, F., Zhou, L., & Yang, F. (2023). Medical text classification based on the discriminative pre-training model and prompt-tuning. Digital Health, 9, 20552076231193213.

[3] Richter-Pechanski P, Geis NA, Kiriakou C et al. Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models. Digital Health 2021; 7: 20552076211057662

[4] Saad, E., Sadiq, S., Jamil, R., Rustam, F., Mehmood, A., Choi, G. S., & Ashraf, I. (2022). Predicting death risk analysis in fully vaccinated people using novel extreme regression-voting classifier. Digital Health, 8, 20552076221109530.

[5] Mujtaba, G., Shuib, L., Idris, N., Hoo, W. L., Raj, R. G., Khowaja, K., ... & Nweke, H. F. (2019). Clinical text classification research trends: systematic literature review and open issues. Expert systems with applications, 116, 494-520.

[6] Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics, 53, 196-207.

[7] Yahia, H. S., & Abdulazeez, A. M. (2021). Medical text classification based on convolutional neural network: a review. International Journal of Science and Business, 5(3), 27-41.

[8] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

[9] Qasim, R., Bangyal, W. H., Alqarni, M. A., & Almazroi, A. A. (2022). A fine-tuned BERT-based transfer learning approach for text classification. Journal of healthcare engineering, 2022.

[10] Guo, Y., Ge, Y., Yang, Y. C., Al-Garadi, M. A., & Sarker, A. (2022, August). Comparison of pretraining models and strategies for health-related social media text classification. In Healthcare (Vol. 10, No. 8, p. 1478). MDPI.

[11] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

[12] Naderi, N., Knafou, J., Copara, J., Ruch, P., & Teodoro, D. (2021). Ensemble of deep masked language models for effective named entity recognition in multi-domain corpora. medRxiv, 2021-04.

[13] Hassan, E., Abd El-Hafeez, T., & Shams, M. Y. (2024). Optimizing classification of diseases through language model analysis of symptoms. Scientific Reports, 14(1), 1507.

[14] Kocaman, V., & Talby, D. (2022). Accurate clinical and biomedical named entity recognition at scale. Software Impacts, 13, 100373.

[15] KafiKang, M., & Hendawi, A. (2023). Drug-drug interaction extraction from biomedical text using relation BioBERT with BLSTM. Machine Learning and Knowledge Extraction, 5(2), 669-683.

[16] Yang, H., Zhang, S., Shen, H., Zhang, G., Deng, X., Xiong, J., ... & Sheng, S. (2023). A Multi-Layer Feature Fusion Model Based on Convolution and Attention Mechanisms for Text Classification. Applied Sciences, 13(14), 8550.

[17] Chaichulee, S., Promchai, C., Kaewkomon, T., Kongkamol, C., Ingviya, T., & Sangsupawanich, P. (2022). Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language processing. PloS one, 17(8), e0270595.

[18] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

[19] Batko, K., & Ślęzak, A. (2022). The use of Big Data Analytics in healthcare. Journal of big Data, 9(1), 3.

[20] Engineering, J. O. H. (2023). Retracted: Named Entity Recognition of Medical Text Based on the Deep Neural Network. Journal of healthcare engineering, 2023, 9805297.

[21] Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data processing and text mining technologies on electronic medical records: a review. Journal of healthcare engineering, 2018.

[22] Pagad, N. S., Pradeep, N., Almuzaini, K. K., Maheshwari, M., Gangodkar, D., Shukla, P., & Alhassan, M. (2022). Clinical text data categorization and feature extraction using medical-fissure algorithm and neg-seq algorithm. Computational Intelligence and Neuroscience, 2022.

[23] Zia, A., Aziz, M., Popa, I., Khan, S. A., Hamedani, A. F., & Asif, A. R. (2022). Artificial intelligence-based medical data mining. Journal of Personalized Medicine, 12(9), 1359.

[24] Virgeniya, S. C. (2024). Digital Twins and Predictive Analytics in Smart Agriculture. In Intelligent Robots and Drones for Precision Agriculture (pp. 87-100). Cham: Springer Nature Switzerland.

[25] Uma Maheswari, S., & Dhenakaran, S. S. (2020). Opinion exploration of tweets and amazon reviews. Int. J. Sci. Res.(IJSTR), 1-9.

[26] Vasantharajan, C., Tun, K. Z., Thi-Nga, H., Jain, S., Rong, T., & Siong, C. E. (2022, November). Medbert: A pre-trained language model for biomedical named entity recognition. In 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 1482-1488). IEEE.