

# Using natural language processing to evaluate the impact of specialized transformers models on medical domain tasks

Soufyane Ayanouz, Boudhir Anouar Abdelhakim, Mohammed Ben Ahmed

Faculty of Sciences and Techniques, Abdelmalek Essaadi University, Tangier, Morocco

## Article Info

### Article history:

Received Apr 13, 2023

Revised Aug 15, 2023

Accepted Sep 16, 2023

### Keywords:

Deep learning

Entity recognition

Intelligent machines

Medical

Natural language processing

Relation classification

## ABSTRACT

We are presently living in the age of intelligent machines, machines are rapidly imitating humans as a result of technological breakthroughs and advances in machine learning, deep learning, and artificial intelligence. In our work, we based our approach on the idea of utilizing a specialized corpus to enhance the performance of a pre-trained language model. We utilized the following approach: (V = vocabulary domain, C1 = initial corpus, C2 = specialization corpus). We applied this approach with different combinations such as (V = general, C1 = general, C2 =  $\emptyset$ ), (V = general, C1 = general, C2 = medical), (V = medical, C1 = medical, C2 =  $\emptyset$ ), and (V = medical, C1 = medical, C2 = medical) to compare the performance of a general bidirectional encoder representations from transformers model and specialized BERT models for the medical domain. In addition, we evaluated the model's using informatics for integrating biology and the bedside, and drug-drug interaction datasets to measure their effectiveness in medical tasks.

*This is an open access article under the [CC BY-SA](#) license.*



## Corresponding Author:

Soufyane Ayanouz

Faculty of Sciences and Techniques, Abdelmalek Essaadi University

Bp416 Tangier-Morocco

Email: ayanouz.soufyane@gmail.com

## 1. INTRODUCTION

Transformers are a type of neural network architecture that has become increasingly important in recent years, particularly in the field of natural language processing (NLP) [1]. Transformers were first introduced in a paper published by Google researchers in 2017 [2] and have since become one of the most popular models for NLP tasks such as language translation and sentiment analysis. The key advantage of transformers is that they are able to process entire sequences of words at once, rather than one word at a time. This makes them much more efficient and effective for tasks that require understanding the context of words in a sentence or paragraph. The importance of transformers in this era cannot be overstated, as they have enabled significant advances in NLP and have become a critical tool for businesses and researchers working with language data.

Artificial intelligence (AI) [3] and machine learning (ML) [4] are increasingly being used in the medical field [5], to help diagnose diseases, identify risk factors, and develop personalized treatment plans. The importance of AI and ML in medicine is due to the vast amounts of data that are generated by healthcare systems and the need for more efficient and accurate ways of analyzing this data. AI and ML algorithms can process and analyze large volumes of medical data, enabling healthcare providers to make more informed decisions about patient care. This has the potential to significantly improve patient outcomes and reduce the cost of healthcare.

The use of AI and ML in the medical field has created a need for specialized models that can effectively process and understand medical language. This is where BioBert [6], BlueBert [7], and SciBert [8]

come into play. These specialized transformers have been specifically trained on biomedical text, allowing them to understand the specific terminology and context of medical language. By using these models, healthcare providers can more accurately extract information from medical records, identify risk factors for diseases, and develop new treatments and therapies. In this way, the importance of AI and ML in medicine is greatly enhanced by the development of specialized models like BioBert, BlueBert, and SciBert.

This work focuses on analyzing how the bidirectional encoder representations from transformers (BERT) model performs in the medical field under different conditions. Three key variables are being studied: the vocabulary domain used, the initial training corpus, and the specialization corpus. The aim is to investigate how these variables affect the final performance of the model, which is then evaluated through two classic biomedical tasks: medical entity recognition and textual implicature. To ensure a fair comparison, all models are trained using the same hyperparameters. The work's findings will provide valuable insights into optimizing the BERT model for use in the medical field, which could lead to more accurate diagnoses and better patient outcomes.

The objective of our research is to compare the performance of the general BERT model with a specialized one that is specifically designed for the medical field. Our study is organized as follows: In Section 2, we provide an overview of the relevant literature. Section 3 presents the methodology used in our research, while Section 4 highlights the significance of vocabulary in the BERT model. The results of our research and their interpretation are summarized in Section 5. Finally, in the concluding section, we provide a summary of our findings and draw our conclusions.

## 2. RELATED WORKS

### 2.1. Transformers library: a hub for natural language processing models

Transformers' is based on tensor2tensor library from the 1990s and the original source code for BERT [9], both from Google Research. Originally developed by AllenNLP, the notion of enabling simple caching for pre-trained models is now widely used. As well as neural translation and language modelling systems, such as Fairseq, the library is intimately associated with them, open NMT [10], texar [11], megatron-LM [12], and marian NMT [13] aspects of transformers that go beyond this are the addition of new user-facing features that allow for the easy downloading, caching, and fine-tuning of models, as well as a smooth transition from development to production. A degree of interoperability with these libraries may be found in transformers, which includes a tool for doing inference using models from marian NMT and Google's BERT, which is the most closely similar of the three.

There has been a long legacy of user-friendly, easy-to-use libraries for general-purpose natural language processing that have been developed throughout time (NLP). Natural language toolkit (NLTK) and NLTK2 are two core libraries that are required [14] and Stanford core NLP [15], which compile several diverse techniques for natural language processing into a single package. Recent general-purpose, open-source libraries have concentrated largely on machine learning for a range of natural language processing applications, such as Spacy, AllenNLP [16], and Stanza [17]. These libraries provide similar functionality to transformers, and Transformers provides functionality that is equivalent to it. In addition, each of these libraries now makes use of the Transformers library and model hub as a low-level framework to make their operations more convenient and efficient. In order to fulfil its role as a hub for natural language processing models, transformers is linked to well-known model hubs such as torch hub and tensor flow hub, which aggregate model parameters unique to certain frameworks for easy usage. The transformers system, in contrast to typical hubs, is domain-specific, allowing it to provide automated assistance for model analysis, usage, deployment, benchmarking, and simple replicability while being cost-effective.

It is a common practice for individuals to visit hospitals or doctors for routine checkups or even minor illnesses. This can be quite burdensome and time-consuming. To address this issue, medical chatbots have been developed to assist patients with their inquiries. These chatbots possess advanced learning capabilities and exceptional problem-solving skills. They have proven to be particularly helpful in providing guidance to patients with minor ailments. This is made possible through the utilization of natural language processing [18] technology in the chatbot design.

### 2.2. Applications of Bert in the medical field

Bert model is applied in several domains such as finance [19], marketing [20], and healthcare [21]. In the medical domain, several applications are based on the Bert model. Liu *et al.* [22] used Bert to Extract Evidence from Chinese Radiology Reports. Kim *et al.* [23] developed a bert model for analyzing natural language in Korean medicine. In [24], they have developed a role-distinguishing Bert model for a sustainable smart city's medical conversation system. In [25], they have used Bert for Recognizing Mental Health Intent in Arabic-Speaking Patients. As for [26], they proposed a Bert-based pre-training framework for named entity recognition in medical records.

### 3. METHOD

#### 3.1. Transformer model

In the paper 'Attention Is All You Need,' a groundbreaking architecture named transformer [27] is presented. As the title implies, it leverages the attention mechanism. This sequence-to-sequence model, such as the long short-term memory networks, uses the transformer architecture to convert one sequence into another using two components (encoder and decoder). However, it differs from previously introduced models as it does not make use of any recurrent networks (gated recurrent units and long short-term memory). Traditionally, recurrent networks were considered to be a highly effective method of understanding temporal relationships in sequences. However, the research team that presented the findings has shown that an architecture that solely relies on attention mechanisms and not recurrent neural networks (RNNs) can lead to better results in tasks such as translation and other activities. The team behind BERT: pre-training of deep bidirectional transformers for language understanding has introduced one such enhancement for natural language tasks as illustrated in the Figure 1.

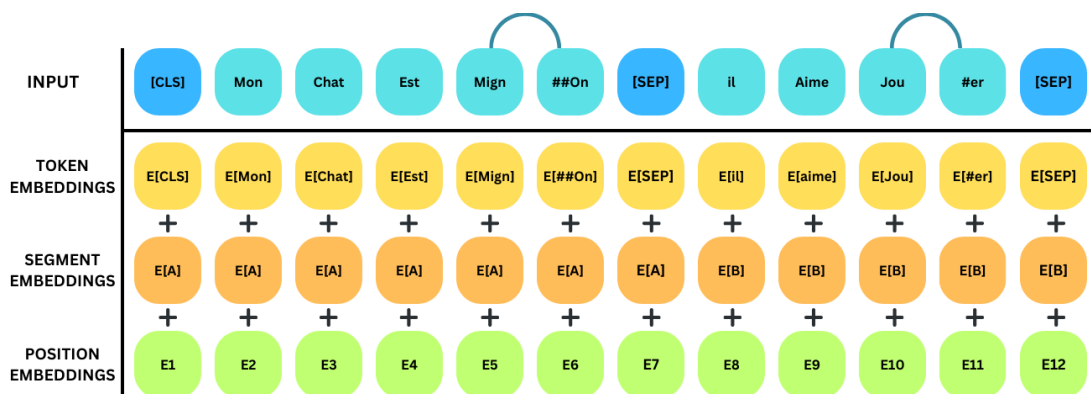


Figure 1. Representation of input in BERT

In contrast to traditional word-based approaches, BERT uses a vocabulary comprising a mixture of words and subwords called word pieces [18]. This allows it to address the problem of out-of-vocabulary words by breaking each unknown word into a sequence of subwords that are part of the vocabulary. If necessary, BERT can go down to the character level to decompose any input word.

During the pre-training phase, BERT systematically takes a pair of sentences as input. These sentences are then segmented into either words or sub-words, with a special symbol for each word. This is achieved by using a special symbol [CLS] at the beginning of the sequence and special symbols [SEP] after each phrase. Each element of the input is also represented embedding matrix. Then, in order to inject a notion of position and to distinguish more easily elements from each of the input sentences, we add a position embedding to this initial vector, this position embedding and segment embedding are added to this initial vector. The complete view of these inputs is shown in Figure 1.

Despite its modest size, positional encoding is critical to the overall structure of the model. We must somehow associate a relative location with each word or segment of our sequence, as there are no recurrent networks capable of remembering how sequences are fed into a model. This is necessary because the order of the components in a sequence is determined by the sequence's elements. By including these places, the size of each word's embedded representation is increased (which is an n-dimensional vector). For instance, take the Multi-Head Attention blocks in the model as an example.

Explanation: Multi-Head Attention (a) as shown in the Figure 2, the multi-head attention mechanism, the K-keys, V-values, and Q-queries are projected for each of the heads and are used to compute an attention scalar product. A scalar product of attention. The set of products is then concatenated and finally undergoes a linear projection. Transformer Layer (b) within the transformer layer, the output of the previous layer serves as the key, value, and query for the calculation of the multi-head attention.

Let's start with the description of the attention mechanism on the left-hand side of the screen. It isn't difficult to understand and maybe summarized by the following mathematical as (1):

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \times V \quad (1)$$

The table comprises of an inquiry, which is a vector illustration of a single word in the sequence. The keys are stored in K which are vector illustrations of all the words in the sequence. The values are stored in V which are also vector illustrations of all the words in the sequence. In the case of multi-head attention modules such as encoders and decoders, V and Q possess the same word sequence. However, V differs from the sequence represented by Q in the attention module, as it takes into account both the encoder and decoder sequences while making decisions. To simplify, we can say that the values in V are combined and added together with certain attention coefficients, where the attention coefficients are determined by the relationship of each word in the sequence (represented by Q) to all the other words in the sequence.

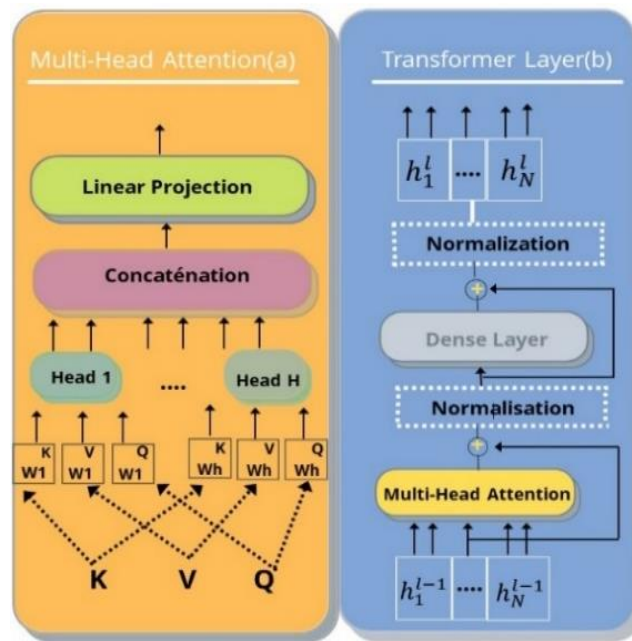


Figure 2. Schematic of a transformer layer in BERT

### 3.2. Bert model

BERT utilizes a unique approach in language modeling by implementing the bidirectional training feature of the transformer architecture, a widely used attention model. Unlike traditional methods that analyze text in a linear fashion, BERT's bidirectional training allows for a more comprehensive understanding of the context and flow of language. The introduction of a new technique called masked language modeling (MLM) further enables bidirectional training in models where it was previously impossible.

#### 3.2.1. How does bert nlp model works

BERT is a language model that employs the transformer architecture, a mechanism that identifies the correlation between words or subwords in a text. The transformer architecture consists of two different components, an encoder that reads the input text and a decoder that produces a forecast for the task. Nevertheless, as the goal of BERT is to create a language model, only the encoder is utilized. The functioning of the transformer architecture is further outlined in this article.

#### 3.2.2. How bert predicts

Prior to inputting word sequences into BERT, a percentage (15%) of the words in each sequence are substituted with a placeholder token [MASK]. The model then attempts to infer the original value of the replaced words, based on the context provided by the remaining unmasked words in the sequence. The BERT loss function only takes into account the prediction of the masked values and ignores the prediction of the unmasked words. As a result, the model converges more slowly than directional models, a drawback that is compensated by its superior contextual knowledge. During the training process, the model is supplied with pairs of sentences as input and is trained to determine whether the second sentence in the pair is the consecutive sentence in the original document. Half of the input used during training are pairs where the second sentence is the next in the original document, while the other half consists of a random sentence from the corpus selected as the second sentence, which is assumed to be unrelated to the first sentence.

– Masked language modelling

BERT is an advanced language model that utilizes a different approach in comparison to conventional language modeling techniques, which aims to predict the next word based on the previous words. Instead, BERT is trained to predict a randomly masked word within the input text. The transformer architecture allows the model to take into account both the right and left contexts of the target word. This approach enables BERT to learn more comprehensive and contextualized representations in comparison to unidirectional models such as ELMo. ELMo generates its representations by combining unidirectional representations, while BERT's representations are generated in a bidirectional manner. In practice, target words may be replaced with a special symbol [MASK], replaced with another random word, or kept as is (as shown in Figure 3) - MLM: The original text is modified by randomly selecting words, each of which may be replaced by a special symbol [MASK], replaced by another word from the vocabulary, or kept intact as illustrated in Figure 3.

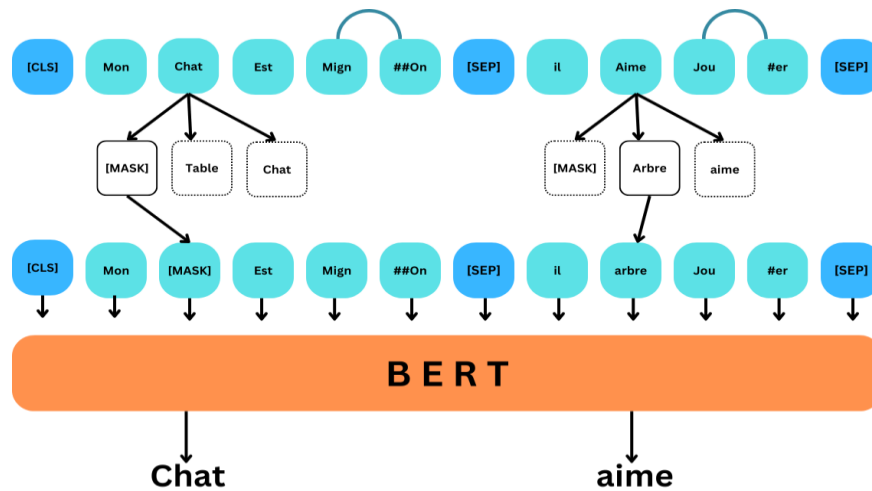


Figure 3. The masked language model of bert

– Next sentence prediction

BERT is also trained to identify whether two input sentences are consecutive in a next-sentence prediction task. The purpose of this task is to improve the model's performance on tasks that involve determining the relationship between pairs of sentences, such as textual implication. Practically, a special symbol [CLS] is used for classification of input sentence pairs and other classifications once the model is trained.

**4. IMPORTANCE OF VOCABULARY IN BERT**

The segmentation applied by BERT is done in two steps: first a "classic" segmentation into words, then a cutting into sub-words (wordpieces). During this second step, BERT cuts as many times as necessary the words out of the vocabulary until known wordpieces are found.

Thus, the choice of vocabulary should directly influence the quality of this decomposition, especially in the medical domain where technical vocabulary is widely used. by two vocabularies: one from the general domain and the other from the medical domain, we then observe that the medical vocabulary tends to decompose words much less than the general vocabulary. To be sure, we analyze the result of the segmentation of a medical corpus to ensure this, we analyze the segmentation result of a medical corpus, whether we count occurrences in the text or distinct word types as shown in Table 1.

Table 1. Wordpieces segmentations from vocabularies of different domains

Reference term	Medical vocabulary	General vocabulary
Gastroentérite	(gastro, ##enterite)	(ga, ##st, ##ro, ##en, ##ter, ##ite)
Hypertension	(hyper, ##tension)	(hy, ##per, ##ten, ##sion)
Myocardite	(myo, ##cardite)	(my, ##oc, ##ard, ##ite)

## 5. EXPERIMENTATION

The approach generally adopted is to train specialized versions of BERT from the original model (general domain) by simply continuing the pre-training procedure on specialized texts. To evaluate the relevance of this strategy, we train several models by varying the following parameters: vocabulary domain (medical vs. general), initial corpus (medical vs general vs a mixture of both), and specialization corpus (none vs. medical). We then compare the performance of these models across different tasks and domains to assess the effectiveness of the specialization strategy. Additionally, we conduct thorough analysis of the models' transfer learning capabilities to determine their adaptability and generalization to new domains.

### 5.1. Evaluation tasks

In the field of natural language processing, evaluation tasks are critical for determining the effectiveness of different machine learning models and algorithms. Two of the most commonly used datasets for these evaluations are i2b2 and DDi. The i2b2 dataset focuses on extracting information from clinical text, such as identifying medical conditions and treatments. It has been widely used in research on clinical natural language processing. The DDi dataset, on the other hand, is focused on drug-drug interactions and aims to identify whether a given pair of drugs has a potential interaction. Both of these datasets are challenging and require sophisticated machine learning techniques to achieve high accuracy. As a result, they are often used as benchmarks for evaluating the effectiveness of different approaches in natural language processing.

#### 5.1.1. Entity recognition on i2b2 dataset

Medical entity recognition is a natural language processing task that involves identifying and extracting medical concepts such as symptoms, diseases, treatments, and tests from text. Here are some examples of the three types of medical concepts that are commonly identified in medical entity recognition:

- **PROBLEM:** This category includes medical concepts related to symptoms, conditions, or diseases that a patient might be experiencing. Examples of PROBLEM medical concepts include "headache," "asthma," "diabetes," "hypertension," and "cancer."
- **TREATMENT:** This category includes medical concepts related to the treatments or therapies that a patient might be receiving for a particular condition or disease. Examples of TREATMENT medical concepts include "oxycodone," "insulin," "radiation therapy," "chemotherapy," and "surgery."
- **TEST:** This category includes medical concepts related to the diagnostic tests or procedures that a patient might undergo to diagnose or monitor a particular condition or disease. Examples of TEST medical concepts include "MRI," "CT scan," "blood test," "urine test," and "biopsy."

#### 5.1.2. Relation classification on DDI dataset

Drug-drug interaction (DDI) can occur in a variety of ways and can have significant impacts on a patient's health outcomes. Being aware of the four categories of DDI is crucial for healthcare professionals to effectively manage patients' medications. By understanding how drugs can interact with each other, healthcare providers can make informed decisions about medication regimens and avoid potential adverse effects, here are some examples of the four categories of DDI:

- **DDI-advise:** This category involves drugs that have a low risk of interaction, but caution is still advised when taken together. For example, a doctor might advise a patient taking aspirin to avoid drinking alcohol, as the combination can increase the risk of stomach bleeding.
- **DDI-effect:** This category involves drugs that interact with each other to produce an effect that is different from the intended therapeutic effect. For example, taking a combination of opioids and benzodiazepines can lead to respiratory depression, which can be life-threatening.
- **DDI-mechanism:** This category involves drugs that interact with each other at the molecular level, altering the pharmacokinetics or pharmacodynamics of one or both drugs. For example, rifampin, an antibiotic, can increase the metabolism of warfarin, an anticoagulant, leading to a decrease in warfarin's effectiveness.
- **DDI-int:** This category involves drugs that interact with each other but the mechanism and clinical significance of the interaction are not well understood. For example, there is limited data on the interaction between cannabidiol, a component of cannabis, and other drugs, so caution is advised when taking them together.

### 5.2. BERT configurations

In the following, we use the BERT (base, uncased) architecture which consists of  $L = 12$  transformer layers with for each,  $H = 12$  heads. Our models are all learned from English texts and produce contextualized plots of dimension 768. We denote each configuration by a triplet corresponding to the different parameter values: (V = vocabulary domain, C1 = initial corpus, C2 = specialization corpus)

(V = general, C1 = general, C2 =  $\emptyset$ ) for a fair comparison, we train our model for the general domain. Training this model ourselves ensures consistency in training conditions for all models being compared. However, we use the same vocabulary as the original BERT model: a vocabulary built from English Wikipedia and Book Corpus corpora [27]. During pre-training, we use a general corpus (see Table 2) built from English Wikipedia as well as a portion of the OpenWebText corpus [28].

(V = general, C1 = general, C2 = medical) here we seek to replicate the classic approach of continuing to train a general domain model on specialized texts. Specifically, while retaining the general vocabulary, we continue training the previous model on a medical corpus consisting of clinical notes from MIMIC-III [29] and abstracts of medical scientific articles from PubMed [30].

(V = medical, C1 = medical, C2 =  $\emptyset$ ) Unlike previous models, this version is trained directly on medical texts. Moreover, here we use a medical vocabulary that we build from the medical corpus (see Table 2) through the theSentencePiece library, which implements the BPE algorithm [31].

(V = medical, C1 = medical, C2 = medical) From the model trained directly on the medical corpus, we perform a second full training on the same corpus in order to arrive to a comparable version in terms of model training time (V = general, C1 = general, C2 = medical).

Table 2. Details of the corpus used for BERT pre-training

Corpus	Composition	Number of documents	Number of words
General	Wikipedia	12,906,740	2,127,653,315
	Open Web Text	32,102,000	1,284,308,223
Medical	MIMIC-111	4,154,115	504,856,155
	Pub Med	4,532,218	520,316,200

## 6. RESULTS AND DISCUSSION

In Table 3, the performance of i2b2/VA 2010 is calculated in terms of strict F1 on the entities to be detected. The performance of DDI is calculated in terms of micro-F1 measurement. The best performance is affected in bold.

Table 3 presents the results of the evaluation of various models trained using different corpora and vocabularies for biomedical tasks. The table highlights the importance of the corpus and vocabulary used in training the BERT models. The results show that the general model, which was trained on a general domain corpus without any medical or additional corpus, achieved lower F1 scores than the medical model, which was trained on a medical corpus with a medical vocabulary. This finding is expected as the medical corpus contains domain-specific terminologies that are useful for biomedical tasks. Moreover, the use of an additional medical corpus in the training process further improved the performance of the medical model, highlighting the importance of additional corpus in training biomedical language models.

Another interesting finding in the table is that the combination of corpora did not always result in better performance than the purely medical model. The model trained on a combination of medical and general domain corpus performed worse than the medical model, but better than the general model. This finding suggests that adding a general corpus to a medical corpus could be useful for some biomedical tasks but not all. Therefore, the choice of corpus to use in training a BERT model should be based on the specific task at hand.

In conclusion, the table demonstrates the importance of corpus and vocabulary choice in training BERT models for biomedical tasks. The use of a medical corpus with a medical vocabulary result in better performance compared to a general domain corpus without a medical vocabulary. Furthermore, adding an additional medical corpus to the training process further improves the performance of the model. However, the combination of corpora did not always lead to better performance than the purely medical model. Therefore, selecting the appropriate corpus for a specific task is crucial in training BERT models for biomedical applications.

Table 3. Result of the evaluation of the models

Model	V	C1	C2	Evaluation Task	
				i2b2/VA 2010	DDI
General	General	$\emptyset$		85.68	76.61
General	General	Medical		89.05	78.72
Medical	Medical	$\emptyset$		88.96	79.42
Medical	Medical	Medical		89.60	81.11
BERT (base)				85.12	78.89
BLUE BERT (base) [7]				88.90	78.89



## 7. CONCLUSION

In conclusion, this research evaluated the performance of several models using different configurations of vocabulary and corpus on a variety of evaluation tasks. The results showed that the model trained on a medical corpus with a medical vocabulary consistently performed better than its general domain counterpart. However, the combination of corpora did not lead to a significant improvement compared to the purely medical model. Furthermore, we found that the models trained on a second corpus had similar performances and that no configuration was systematically better for all the tasks. One of the main findings of the study is that the models with medical vocabulary seem to be favored in the case of biomedical tasks (DDI). This difference may be due to the type of task, as these tasks are related to the biomedical domain, rather than the clinical domain. However, it is important to note that the study also revealed that the combination of corpora never leads to a significant improvement compared to the purely medical model. It is also important to note that while the results of this research are promising, there are some limitations to consider. One limitation is that the study only evaluated the performance of the models on a limited number of tasks and that further research is needed to evaluate the models on a broader range of tasks. Additionally, the study only used two corpora, a general corpus and a medical corpus, and it would be beneficial to evaluate the models using other types of corpora to see if the results would differ. Given the limitations of this research, there are a number of potential avenues for future research. One avenue would be to evaluate the performance of the model on a wider range of datasets and tasks, in order to further investigate the generalizability of the findings. Another avenue would be to evaluate the model on other languages, to determine whether the results are applicable to other languages. Additionally, it would be valuable to evaluate the performance of other types of models, to investigate if the results are applicable to other models. Furthermore, studying the impact of corpus size and other pre-training parameters on the performance of the model is an interesting perspective to explore. Finally, it would be intriguing to examine the effect of fine-tuning parameters on the performance of the model.

## REFERENCES




- [1] E. Ramos-Pérez, P. J. Alonso-González, and J. J. Núñez-Velázquez, "Multi-transformer: a new neural network-based architecture for forecasting S&P volatility," *Mathematics*, vol. 9, no. 15, p. 1794, Jul. 2021, doi: 10.3390/math9151794.
- [2] D. Luitse and W. Denkena, "The great Transformer: Examining the role of large language models in the political economy of AI," *Big Data & Society*, vol. 8, no. 2, p. 205395172110477, Jul. 2021, doi: 10.1177/20539517211047734.
- [3] P. Aghion, B. Jones, and C. Jones, "Artificial intelligence and economic growth," Cambridge, MA, MA, Oct. 2017. doi: 10.3386/w23928.
- [4] J. Bell, "What is machine learning?," in *Machine Learning and the City*, Wiley, 2022, pp. 207–216. doi: 10.1002/9781119815075.ch18.
- [5] A. K. Feeny *et al.*, "Artificial intelligence and machine learning in arrhythmias and cardiac electrophysiology," *Circulation: Arrhythmia and Electrophysiology*, vol. 13, no. 8, Aug. 2020, doi: 10.1161/CIRCEP.119.007952.
- [6] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.
- [7] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on Ten benchmarking datasets," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, Stroudsburg, PA, USA: Association for Computational Linguistics, Jun. 2019, pp. 58–65. doi: 10.18653/v1/W19-5006.
- [8] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," Mar. 2019, [Online]. Available: <http://arxiv.org/abs/1903.10676>
- [9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [10] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "OpenNMT: Open-source toolkit for neural machine translation," Jan. 2017, [Online]. Available: <http://arxiv.org/abs/1701.02810>
- [11] Z. Hu *et al.*, "Texar: A modularized, versatile, and extensible toolbox for text generation," in *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, Stroudsburg, PA, USA, PA, USA: Association for Computational Linguistics, 2018, pp. 13–22. doi: 10.18653/v1/W18-2503.
- [12] M. Shoyebi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training multi-billion parameter language models using model parallelism," Sep. 2019, [Online]. Available: <http://arxiv.org/abs/1909.08053>
- [13] M. Junczys-Dowmunt *et al.*, "Marian: Fast Neural Machine Translation in C++," in *Proceedings of ACL 2018, System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 116–121. doi: 10.18653/v1/P18-4020.
- [14] E. Loper and S. Bird, "NLTK," in *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics -*, Morristown, NJ, USA, NJ, USA: Association for Computational Linguistics, 2002, pp. 63–70. doi: 10.3115/1118108.1118117.
- [15] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Stroudsburg, PA, USA, PA, USA: Association for Computational Linguistics, 2014, pp. 55–60. doi: 10.3115/v1/P14-5010.
- [16] M. Gardner *et al.*, "AllenNLP: A deep semantic natural language processing platform," in *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, Stroudsburg, PA, USA, PA, USA: Association for Computational Linguistics, 2018, pp. 1–6. doi: 10.18653/v1/W18-2501.
- [17] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 101–108, Mar. 2020.
- [18] Y. Wu *et al.*, "Google's neural machine translation system: bridging the gap between human and machine translation," Sep. 2016, [Online]. Available: <http://arxiv.org/abs/1609.08144>






- [19] L. Zhao, L. Li, X. Zheng, and J. Zhang, "A BERT based sentiment analysis and key entity detection approach for online financial texts," in *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, IEEE, May 2021, pp. 1233–1238. doi: 10.1109/CSCWD49262.2021.9437616.
- [20] M. G. Sousa, K. Sakiyama, L. de S. Rodrigues, P. H. Moraes, E. R. Fernandes, and E. T. Matsubara, "BERT for stock market sentiment analysis," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, Nov. 2019, pp. 1597–1601. doi: 10.1109/ICTAI.2019.00231.
- [21] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, "A fine-tuned BERT-based transfer learning approach for text classification," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–17, Jan. 2022, doi: 10.1155/2022/3498123.
- [22] H. Liu *et al.*, "Use of BERT (Bidirectional Encoder Representations from Transformers)-based deep learning method for extracting evidences in Chinese radiology reports: development of a computer-aided liver cancer diagnosis framework," *Journal of Medical Internet Research*, vol. 23, no. 1, p. e19689, Jan. 2021, doi: 10.2196/19689.
- [23] Y. Kim *et al.*, "A pre-trained BERT for Korean medical natural language processing," *Scientific Reports*, vol. 12, no. 1, p. 13847, Aug. 2022, doi: 10.1038/s41598-022-17806-8.
- [24] S. Wang, S. Wang, Z. Liu, and Q. Zhang, "A role distinguishing BERT model for medical dialogue system in sustainable smart city," *Sustainable Energy Technologies and Assessments*, vol. 55, p. 102896, Feb. 2023, doi: 10.1016/j.seta.2022.102896.
- [25] R. Mezzi, A. Yahyaoui, M. W. Krir, W. Boulila, and A. Koubaa, "Mental health intent recognition for Arabic-speaking patients using the Mini International Neuropsychiatric Interview (MINI) and BERT Model," *Sensors*, vol. 22, no. 3, p. 846, Jan. 2022, doi: 10.3390/s22030846.
- [26] N. Liu, Q. Hu, H. Xu, X. Xu, and M. Chen, "Med-BERT: A pretraining framework for medical records named entity recognition," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5600–5608, Aug. 2022, doi: 10.1109/TII.2021.3131180.
- [27] H. Wang, D. Yu, K. Sun, J. Chen, and D. Yu, "Improving pre-trained multilingual model with vocabulary expansion," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Stroudsburg, PA, USA, PA, USA: Association for Computational Linguistics, 2019, pp. 316–327. doi: 10.18653/v1/K19-1030.
- [28] A. Gokaslan, V. Cohen, E. Pavlick, and S. Et Tellex, "Openwebtext corpus (2019)," 2019.
- [29] A. Johnson, T. Pollard, and R. Mark, "MIMIC-III Clinical Database (version 1.4)," *PhysioNet*, no. June, 2015, doi: 10.13026/C2XW26.
- [30] N. Fiorini, R. Leaman, D. J. Lipman, and Z. Lu, "How user intelligence is improving PubMed," *Nature Biotechnology*, vol. 36, no. 10, pp. 937–945, Nov. 2018, doi: 10.1038/nbt.4267.
- [31] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with Subword units," Aug. 2015, [Online]. Available: <http://arxiv.org/abs/1508.07909>

## BIOGRAPHIES OF AUTHORS






**Soufyane Ayanouz**    is currently a Ph.D. student at the Faculty of Sciences and Techniques of Tangier. He is also an engineer at the field of computer science since 2018. He received the engineering degree from Abdelmalek Essaadi University and he is the co-author of several papers published in IEEE Explorer, ACM, and in high indexed journals and conferences. He published several book chapters on Springer series, and is part of the reviewer's community. His key research relates to Artificial intelligence and deep learning applied to the medical field. He can be contacted at email: [ayanouz.soufyane@gmail.com](mailto:ayanouz.soufyane@gmail.com).



**Boudhir Anouar Abdelhakim**    is currently an associate professor at the Faculty of Sciences and Techniques of Tangier. Actually, he was the president of the Mediterranean Association of Sciences and Technologies. He is an adviser at the Moroccan union against dropping out of school and IEEE Member since 2009. He received the HDR degree from Abdelmalek Essaadi University and he is the co-author of several papers published in IEEE Explorer, ACM, and in high indexed journals and conferences. He co-edited a several books published on Springer series and he is a co-founder of a series of international conferences. He can be contacted at email: [boudhir.anouar@gmail.com](mailto:boudhir.anouar@gmail.com).



**Mohamed Ben Ahmed**    is a full professor at the University Abdelmalek Essaadi. He received the PhD degree in computer sciences and Telecommunications in 2010. He is currently a full Professor in computer sciences department at the Faculty of Sciences and Techniques-Morocco. He supervised several theses and conducted several international research projects about smart cities. He has authored more than 30 papers published in international journals and conferences. He co-edited a several books published on Springer series and he is a co-founder of a series of international conferences. He can be contacted at email: [m.benahmed@gmail.com](mailto:m.benahmed@gmail.com).