Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240580

What Kind of Transformer Models to Use for the ICD-10 Codes Classification Task

Mariem MANSOUR[a], Fatma YILMAZ[a], Marko MILETIC[a] and Murat SARIYAR [a,1]
[a]*Bern University of Applied Sciences, Switzerland*
ORCiD ID: Murat Sariyar https://orcid.org/0000-0003-3432-2860

Abstract. Coding according to the International Classification of Diseases (ICD)-10 and its clinical modifications (CM) is inherently complex and expensive. Natural Language Processing (NLP) assists by simplifying the analysis of unstructured data from electronic health records, thereby facilitating diagnosis coding. This study investigates the suitability of transformer models for ICD-10 classification, considering both encoder and encoder-decoder architectures. The analysis is performed on clinical discharge summaries from the Medical Information Mart for Intensive Care (MIMIC)-IV dataset, which contains an extensive collection of electronic health records. Pre-trained models such as BioBERT, ClinicalBERT, ClinicalLongformer, and ClinicalBigBird are adapted for the coding task, incorporating specific preprocessing techniques to enhance performance. The findings indicate that increasing context length improves accuracy, and that the difference in accuracy between encoder and encoder-decoder models is negligible.

Keywords. NLP, transformer models, BERT, ClinicalLongformer, ClinicalBigBird

1. Introduction

The precise coding of patient data in hospitals has become more complex and costly with the introduction of the International Classification of Diseases (ICD)-10 and its clinical modifications (CM). This change has increased the number of diagnosis codes from about 18,000 in ICD-9 to over 155,000, significantly enlarging the challenges for medical coders. Coders play a crucial role in translating complex clinical information into standardized codes, a process that requires precision and is both time-consuming and error-prone. They often face difficulties interpreting incomplete patient records and fragmented examination reports, impacting the speed and accuracy of the coding process. This situation emphasizes the importance of accurate billing processes and medical statistics collection. Natural Language Processing (NLP) facilitates the analysis of unstructured data [1]. This is particularly relevant for free-text fields in electronic health records, which serve as a central source of information for diagnosis coding [2].

Our investigation aims to determine the optimal type of transformer models for ICD-10 classification. From a theoretical standpoint, encoder models should suffice, as they can be effectively fine-tuned for classification tasks. The CLS token in BERT, for

[1] Corresponding Author: Murat Sariyar, Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland; E-mail: murat.sariyar@bfh.ch.

example, is aptly named [3]: it represents a summary of the word embeddings of all subsequent tokens, facilitating the classification of the entire text. Nevertheless, generative language model models based on encoder-decoder transformers hold promise for classification tasks as well. This can be justified by the fact that pre-training of decoders involves the task of predicting the next token in a sequence. Considering the autoregressive nature of decoders, it could be argued that all tokens function akin to the CLS token in reverse (looking backwards). In the machine learning context, classifiers are often categorized as either discriminative or generative. While generative models such as Naive Bayes can be employed for classification tasks, more intricate tasks typically demand discriminative methods. Discriminative models prioritize the classification task without modeling the relationship between all variables simultaneously, resulting in reduced modeling complexity and enhanced efficiency. This distinction extends to transformer models as well: encoder models commonly feature a significantly smaller parameter count compared to encoder-decoder models.

In the subsequent sections, we will provide concise descriptions of the pre-trained models utilized for benchmarking: BioBERT [4], ClinicalBERT [5], ClinicalLongformer [6], and ClinicalBigBird [7]. Evaluation of ICD-10 code classification will be performed using clinical discharge summaries from the Medical Information Mart for Intensive Care (MIMIC)-IV dataset. This dataset comprises a comprehensive collection of electronic health records, encompassing data from intensive care units [8].

2. Methods

From the (MIMIC)-IV dataset, we extract clinical discharge notes along with their corresponding ICD-10 codes as target variables. To limit the complexity of the fine-tuning task, we concentrate on the top 50 ICD codes. This selection is justified as it enables direct comparison with other studies that have also concentrated on these codes. Key pre-processing steps include removing formatting characters and unnecessary spaces, converting the entire text to lowercase, and selectively eliminating numbers. The texts are further reduced by extracting the portion that starts with the "discharge" heading. Stop words are not removed to preserve contextual integrity, as models often require complete context for optimal performance. Text length is adjusted according to each model's specifications. For models with a maximum context length of 512, such as BioBERT and ClinicalBERT, the text is truncated to the first 512 tokens after the keyword "discharge." For more advanced models that support longer context lengths such as 1,024 or 4,096, like ClinicalLongformer and ClinicalBigBird, the maximum token length is defined accordingly. To optimize the finetuning of the models, we employ a constant batch size of 32 across six training epochs. The learning rate for the AdamW optimizer is set to 5e-5 [9]. While we have access to a high-performance server equipped with a Tesla V100 32 GB, access is shared with other research groups, limiting our availability during our evaluations. Following models are considered:

BioBERT: Biomedical Bidirectional Encoder Representations from Transformers is based on Google's BERT encoder and optimized for biomedical text mining. It consists of 110 million parameters and was trained on large biomedical datasets, including PubMed summaries and full texts from PMC. Operating with a maximum context length of 512, it allows for the analysis of text segments of this magnitude. Its bidirectional structure enables understanding the context of a word in both the left and right text

sequences, leading to precise text interpretation. It is primarily used for text mining tasks such as named entity recognition, relation extraction, and question answering.

ClinicalBERT: is also based on Google's BERT and has 110 million parameters. It can also process up to 512 tokens per input. ClinicalBERT differs from its predecessor BioBERT in its pretraining data source. While BioBERT was pretrained on PubMed abstracts and full-text articles, ClinicalBERT was pretrained on the MIMIC-III dataset, which also contains clinical notes from intensive care units as MIMIC-IV does.

ClinicalLongformer: represents a specialized adaptation of the Longformer model, specifically designed for analyzing large clinical text sequences. As an encoder model trained on MIMIC-III clinical notes it has the capability to handle input sequences of up to 4,096 tokens, which constitutes a significant expansion compared to standard encoder models. This enhanced capacity enables the ClinicalLongformer to effectively capture intricate details and nuances present in extensive clinical texts, making it a valuable tool for various healthcare-related natural language processing tasks.

ClinicalBigBird: represents an advanced adaptation of the Transformer architecture, specifically designed for efficiently handling long sequences. At its core, ClinicalBigBird aims to convert the quadratic memory overhead typically associated with traditional Transformer models into a linear one, thereby significantly enhancing scalability. This optimization is achieved through an innovative sparse attention mechanism, which utilizes a combination of localized sliding windows and selective global attention. As an encoder-decoder model, ClinicalBigBird has been pretrained on MIMIC-III clinical notes. It offers the capability to process inputs of up to 4,096 tokens.

3. Results

Table 1 summarizes the results from the evaluation of different models with varying context lengths. Across the board, it's evident that increasing the context length leads to improvements in accuracy. For instance, ClinicalLongformer achieves an accuracy of 0.927 with a context length of 512, which increases to 0.93 with a context length of 1024, and further improves to 0.94 with a context length of 4096. This trend suggests that longer context lengths allow the models to capture more information, resulting in better performance in classifying ICD-10 codes. Similarly, ClinicalBigBird also demonstrates this pattern, with an accuracy of 0.928 with a context length of 512, which improves to 0.94 with a context length of 1024. Due to constraints on our high-performance computer, we did not run ClinicalBigBird with a context length of 4096. The trend suggests that this would likely yield the best result, as we are already achieving comparable performance with a context length of 1024 tokens.

Furthermore, training times vary significantly across models and context lengths. While ClinicalBERT exhibits the shortest training time of 5 minutes and 53 seconds, BioBERT and ClinicalLongformer take around 21 to 22 minutes for training with a context length of 512. However, as the context length increases, the training time significantly escalates, exemplified by ClinicalLongformer's training time of 2 hours and 57 minutes for a context length of 4096. This suggests that longer context lengths not only improve accuracy but also require substantially more computational resources and time for training. Overall, these results underscore the trade-off between accuracy and training time, highlighting the need for careful consideration when selecting context lengths for model training in real-world applications. For efficiency purposes, it may be

more practical to compress the text to a suitable size through preprocessing rather than working with the original size and providing a large context length.

The observed increase in accuracy as the context length increases from 512 to 1024 and then to 4096 is marginal. Despite the significant increase in context length, the improvement in accuracy is relatively minor. This phenomenon can be attributed to several factors. Firstly, the initial context length of 512 may already capture a substantial amount of relevant information within the clinical text data. Therefore, increasing the context length further may not significantly enhance the model's ability to extract additional meaningful insights. Secondly, the law of diminishing returns may apply, indicating that beyond a certain point, increasing the context length may yield diminishing improvements in accuracy. Lastly, other factors such as model architecture, training data quality, and hyperparameter optimization may play crucial roles in determining model performance, overshadowing the impact of context length alone. Overall, while longer context lengths may theoretically allow for capturing more information, in practice, the marginal gains in accuracy may not justify the substantial increase in computational resources and training time required.

Table 1. Results of our Benchmark on the MIMIC-IV dataset using the top 50 ICD-10 codes.

Model	Context Length	Accuracy	Training Time	Model
BioBERT	512	0.926	0:21:29	BioBERT
ClinicalBERT	512	0.926	0:05:53	ClinicalBERT
ClinicalLongformer	512	0.927	0:21:31	ClinicalLongformer
ClinicalLongformer	1024	0.93	0:43:33	ClinicalLongformer
ClinicalLongformer	4096	0.94	02:57:28	ClinicalLongformer
ClinicalBigBird	512	0.928	0:24:10	ClinicalBigBird
ClinicalBigBird	1024	0.94	0:34:17	ClinicalBigBird

4. Discussion and Conclusions

In this study, we conducted a comprehensive analysis of various models to assess their efficacy in automatically assigning ICD-10 codes. Specifically tailored models for medical text data, such as BioBERT and ClinicalBERT, exhibited technical limitations when processing lengthy texts, particularly in medical discharge summaries, which often contain extensive volumes of text. This observation underscores the importance of the token capacity of these models, particularly regarding the processing and analysis of extensive medical text documents. However, we have also observed that the increase in accuracy is not always substantial enough to justify the much higher computational demands of models with longer context lengths. Similar to Naive Bayes in the machine learning domain, both BioBERT and ClinicalBERT can therefore be considered as plain vanilla methods for classification based on medical text data. It is only when accuracy is notably low, and there is a necessity to process long texts, that consideration should be given to more powerful models, which predominantly consist of decoder models.

The decision to retain or remove stopwords depends on the context constraints of each model and the specific requirements of the text mining task. In our study, we intentionally chose not to remove stopwords. This choice was driven by the importance of preserving context, which is fundamental in various scientific investigations and analytical processes. By retaining stopwords, the model can capture the complete contextual nuances present in the text data, facilitating a more comprehensive

understanding of underlying patterns and relationships. Furthermore, this approach aligns with established practices in in the literature, ensuring consistency and comparability. We conducted an internal small sub-study to examine whether reducing stopwords would lead to an improvement in performance. The results were inconclusive, with no definitive conclusion reached. This variability underscores the need for caution in making decisions regarding stopwords [10].

Enhancing preprocessing techniques offers a promising avenue to further improve the already satisfying results, which surpass those reported in existing literature [11]. Firstly, optimizing preprocessing entails implementing advanced techniques to enhance data quality and extract medical information more effectively. This includes refining text cleaning processes, enhancing accuracy in removing irrelevant data, and improving the handling of medical terminologies Additionally, meticulous analysis and selection of training data will be crucial to enhance the accuracy and relevance of the models. Expanding the training data pool by integrating additional datasets and broadening the scope of ICD codes will bolster the robustness and generalizability of the models. Finally, in terms of future prospects, it's important to explore the potential of LoRA for adapting larger models [12]. LoRA offers an efficient way to optimize models for deployment on GPUs with limited capacity, alongside other methods such as quantization or pruning. Selecting an appropriate technique will ensure optimal model performance on hardware with constrained resources, thereby maximizing efficiency.

References

[1] Kunz S, Zgraggen C, Sariyar M. Mapping SNOMED CT Codes to Semi-Structured Texts via an NLP Pipeline. Stud Health Technol Inform 2022; 295: 390–393.

[2] Chen P-F, Wang S-M, Liao W-C, et al. Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning. JMIR Med Inform 2021; 9: e23230.

[3] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, pp. 4171–4186.

[4] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinforma Oxf Engl 2020; 36: 1234–1240.

[5] Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. 28 November 2020. DOI: 10.48550/arXiv.1904.05342.

[6] Zheng H, Zhu Y, Jiang LY, et al. Making the Most Out of the Limited Context Length: Predictive Power Varies with Clinical Note Type and Note Section. 13 July 2023. DOI: 10.48550/arXiv.2307.07051.

[7] Liu L, Perez-Concha O, Nguyen A, et al. Automated ICD coding using extreme multi-label long text transformer-based models. Artif Intell Med 2023; 144: 102662.

[8] Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023; 10: 1.

[9] Llugsi R, Yacoubi SE, Fontaine A, et al. Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito. In: 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM), pp. 1–6.

[10] Siino M, Tinnirello I, La Cascia M. Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. Inf Syst 2024; 121: 102342.

[11] Li Y, Wehbe RM, Ahmad FS, et al. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. Epub ahead of print 15 April 2022. DOI: 10.48550/arXiv.2201.11838.

[12] Hu EJ, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models. 16 October 2021. DOI: 10.48550/arXiv.2106.09685.