Intelligent Health Systems – From Technology to Data and Knowledge
E. Andrikopoulou et al. (Eds.)

doi:10.3233/SHTI250476

Few-Shot Learning of Medical Coding Systems: A Case Study on Death Certificates with BERT and Mistral

Mihai Horia POPESCU[a,1], Vincenzo DELLA MEA[a], and Kevin ROITERO[a]
[a] *University of Udine, Department of Mathematics, Computer Science and Physics*
ORCiD ID: Mihai Horia Popescu https://orcid.org/0000-0003-3378-0368,
Vincenzo Della Mea https://orcid.org/0000-0002-0144-3802,
Kevin Roitero https://orcid.org/0000-0002-9191-3280

Abstract. Identifying the Underlying Cause of Death accurately is crucial for effective healthcare policy and planning. The World Health Organization recommends using the ICD-10 system to standardize death certificate coding, a task often supported by semi-automated systems. This paper assesses the effectiveness of BERT and Mistral language models in automating this process, focusing particularly on their handling of varied instance densities per ICD code, ranging from 1 (simulating the introduction of a new code) to 100 (representing well-established codes). Through extensive comparative experiments, we find that the finetuned Mistral model substantially outperforms BERT, especially in scenarios with limited data. Mistral's higher effectiveness, even with limited data from less commonly used codes, highlight its potential to significantly enhance automated coding systems.

Keywords. Large Language Models, Health Informatics, Death Certificates

1. Introduction

The accurate identification of the underlying cause of death (UCOD) is crucial for effective public health surveillance and policy making. The World Health Organization (WHO) has standardized the coding of death causes through the International Classification of Diseases, Tenth Revision (ICD-10) [1], which facilitates consistent and reliable data annotations across different countries. However, manual coding is labor-intensive, subject to human error, and can be inconsistent due to subjective interpretations and biases.

Recent advances in Natural Language Processing (NLP) offer promising results for automating this process. Classical [2] and BERT [3] based models [4] have demonstrated high accuracy levels, significantly increasing efficiency. However, with the emergence of instruction-based models [5] such as Llama [6] and Mistral [7], there is a renewed interest in testing their capabilities, particularly in their ability to intuitively capture richer text semantics compared to the state-of-the-art model for UCOD identification.

We focus on comparing the performance of an instruction-based model, with the state-of-the-art system, particularly examining how these models adapt to varying

[1] Corresponding Author: Mihai Horia Popescu; E-mail: mihaihoria.popescu@uniud.it.

instances per class, a common scenario in medical coding due to the uneven prevalence of different causes of death. We focus on the following Research Questions (RQ):

- **RQ1:** Can instruction-based models be more effective than BERT in identifying UCOD from death certificates?
- **RQ2**: What is the comparative performance of these models, and which model better adapts to scenarios with few instances per class?

2. Background and Related Work

Numerous studies have developed methodologies that use deep learning networks to predict the UCOD or to handle tasks involving medical classifications [4,8,9]. In particular, [4] achieved state-of-the-art performance in selecting the UCOD using BERT-like models by coding death certificates using NLP, and showing that transformers-based models outperform other families of deep-learning models that were before used for this task [10]. Similar work has been proposed in [2], which developed a modified Inception network, and [11], which compared multiple approaches. While these methods achieve high effectiveness scores in the classification task, they are limited to classifying only the labels present in the training set, which is quite large in the case of ICD codes. Other work focused on this domain, such as [12] which proposed a few-shot ICD coding approach that tackles the long-tail code distribution problem, and [13] that show that prompting LLMs with serialized tabular data in a few-shot settings.

To the best of our knowledge, this is the first work that employs instruction-based models, such as Mistral, to address these challenges in predicting the UCOD of medical certificates. Additionally, we investigate the impact and adaptability of these models with respect to the distribution of codes and the evolution of the classification system, offering insights into how these factors influence model performance.

3. Methodology

To evaluate the performance of BERT and Mistral models in identifying UCOD, we use a dataset of death certificates from the U.S. National Center for Health Statistics[2]. This dataset contains nearly 13 million records from 2014 to 2017, including medical histories and administrative data like sex, birth date, and death date. We simulate varying instance densities, ranging from 1 to 100 instances per ICD code, to test the models' adaptability. This includes scenarios representing new ICD codes (e.g., subsets with 1, 2, or 10 instances per class). Each certificate is pre-processed into a narrative format describing the medical conditions found, following previous work that demonstrated this method improves text semantics [4]. For example, a certificate might be: "Female, 55y: (Acute myocardial infarction) due to (Hypertension) due to (Unspecified kidney failure)" with UCOD "I21.9 - Acute myocardial infarction". We sample up to 100 instances per ICD-10 code for training, using 100K stratified instances as the test set. Although the dataset is structured, it is the most comprehensive resource available for UCOD tasks.

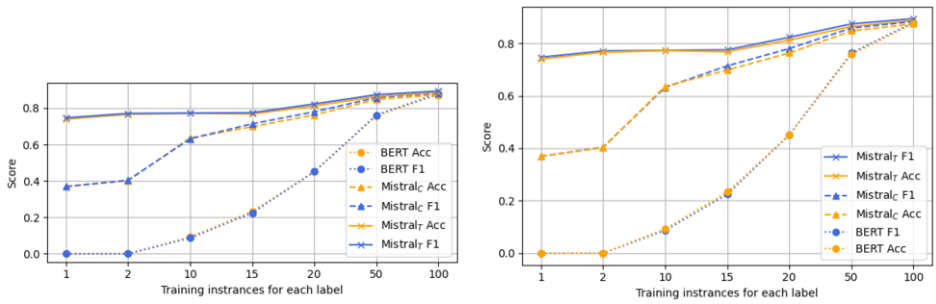
We use two models: BERT and Mistral. BERT [3] revolutionized NLP tasks through its bidirectional attention-based architecture [14], achieving state-of-the-art performance

[2] https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm

in UCOD classification [4]. We also use Mistral's variant Mistral-7B-Instruct-v0.2, based on GPT architecture [15] and optimized for instruction-based tasks [7]. To fine-tune Mistral, we apply Low-Rank Adaptation (LoRA) [16], which adapts the model by training low-rank matrices (less than 2% of parameters). We test LoRA rank values from 16 to 1024, set scaling α to 2 times the rank, and a dropout rate of 0.1, training for 1 epoch. We evaluate two Mistral versions: one predicting ICD-10 codes ($Mistral_C$) and the other generating descriptive text ($Mistral_T$). All experiments run on a server with an NVIDIA RTX A6000 48GB GPU. The code to reproduce our results can be found at: https://osf.io/k3w4b/?view_only=d7cde4616d5d450d86684641987c3fd9.

Table 1. Accuracy, Precision, Recall, and F1 of Mistral's variants. Train dataset is 100 instances per label.

		LoRA Rank						
Metric	**Model**	**16**	**32**	**64**	**128**	**256**	**512**	**1024**
Accuracy	$Mistral_C$	.680	.723	.800	.800	**.874**	.871	.859
Precision	$Mistral_C$	.823	.838	.878	.878	**.907**	.905	.899
Recall	$Mistral_C$	.680	.732	.800	.800	**.874**	.871	.859
F1	$Mistral_C$	.695	.754	.817	.817	**.883**	.877	.868
Accuracy	$Mistral_T$	.808	.827	.859	.859	**.887**	.884	.877
Precision	$Mistral_T$	.862	.884	.901	.901	**.915**	.911	.909
Recall	$Mistral_T$	.808	.827	.859	.859	**.887**	.884	.877
F1	$Mistral_T$	.819	.841	.870	.870	**.894**	.889	.884



(a) Mistral' variants on different LoRA ranks.

(b) BERT, $Mistral_C$, and $Mistral_T$ effectiveness.

Figure 1. Comparison of Mistral variants and other models.

4. Results

We start by investigating the differences between BERT and Mistral for the UCOD classification task. To allow for a fair comparison, seek at which LoRA configuration yields the best Mistral model variation. Table 1 and Figure 1a show the evaluation scores of the fine-tuned models when varying the LoRA ranks. We can see that the best results are achieved when employing a LoRA rank of 256. This behavior is consistent for both Mistral variations i.e., $Mistral_C$ and $Mistral_T$. Furthermore, it is always the case that $Mistral_T$ achieves slightly higher effectiveness scores than $Mistral_C$, for all metrics, reaching an accuracy and F1 scores of respectively 0.887 and 0.894. Figure 1a shows the same trend in a graphical format. The figure shows that there is an increase in both accuracy and F1 scores from rank 16 up to rank 256 where they peak, after which the

scores begin to slightly decline. This trend is consistent with other metrics not shown and aligns with findings from literature studies [17] which observed a non-monotonic effectiveness of LoRA. Moreover, the plot shows that, at equivalent ranks, the $Mistral_T$ models consistently outperform the $Mistral_C$ models. This is more evident at lower LoRA ranks, whereas this difference becomes more subtle both at higher and at the optimal rank of 256, where the model scores almost match.

The effectiveness scores of the best Mistral model are compared with those of BERT, the state-of-the-art model for the UCOD classification task. Table 2 and Figure 1b show such effectiveness scores across the seven dataset splits we created, which are made by sampling from 1 to 100 instances per each ICD code. We see that $Mistral_T$ consistently outperforms all the other models, both $Mistral_C$ and BERT. The best effectiveness scores are reached for all models when considering the maximum number of training instances per code (100 instances per label, last part of Table 2), where BERT achieves and accuracy and F1 scores respectively of 0.875 and 0.880, $Mistral_C$ of 0.874 and 0.883, and $Mistral_T$ of 0.887 and 0.894.

The differences between the models become more pronounced with fewer instances per code. Scenarios with 1, 2, and 10 instances simulate the addition or modification of an ICD code, where experts can provide a few annotated examples to train the model. The upper part of Table 2 reflects the models' adaptability to changes in the knowledge base for death certificate annotations. This trend becomes more evident when inspecting Figure 1b, which plots Table 2 data. Again, the other metrics not reported for space issue show a similar trend. The metrics are color-coded, and models are distinguished by line styles. We can see that it becomes evident $Mistral_T$ demonstrates higher effectiveness scores, particularly in datasets with fewer instances such as 1, 2, and 10 instances per code. The plot highlights the cold-start issue, showing the minimum instances needed for high effectiveness. BERT requires 20-50 cases per label to surpass 0.6 accuracy and F1, while $Mistral_C$ needs 2-10 instances. In contrast, $Mistral_T$ handles cold-start better, achieving over 0.6 with just one instance. The performance gap narrows as the number of instances increases, with BERT showing the largest delta (0.8), $Mistral_C$ (0.5), and $Mistral_T$ (0.1). These results show $Mistral_T$'s strong adaptability to updates in medical coding systems and its effective management of cold-start scenarios, demonstrated by the model's ability to maintain high accuracy with minimal training examples.

Table 2. Accuracy, Precision, Recall, and F1 of BERT and Mistral LoRA Rank 256 (highest from Table 1).

		Dataset						
Metric	**Model**	**1**	**2**	**10**	**15**	**20**	**50**	**100**
Accuracy	BERT	.000	.000	.093	.233	.451	.757	.875
Precision	BERT	.000	.000	.189	.494	.663	.837	.901
Recall	BERT	.000	.000	.093	.233	.451	.757	.875
F1	BERT	.000	.000	.087	.224	.451	.763	.880
Accuracy	$Mistral_C$	.370	.402	.636	.698	.763	.847	.874
Precision	$Mistral_C$	.446	.528	.726	.793	.845	.896	.907
Recall	$Mistral_C$	.370	.402	.636	.698	.763	.847	.874
F1	$Mistral_C$	.369	.404	.632	.714	.780	.858	.883
Accuracy	$Mistral_T$	**.740**	**.766**	**.772**	**.768**	**.812**	**.864**	**.887**
Precision	$Mistral_T$	**.800**	**.836**	**.839**	**.845**	**.873**	**.906**	**.915**
Recall	$Mistral_T$	**.740**	**.766**	**.772**	**.768**	**.812**	**.864**	**.887**
F1	$Mistral_T$	**.747**	**.771**	**.773**	**.775**	**.823**	**.874**	**.894**

Such capabilities make $Mistral_T$ particularly suitable for real-world applications, enabling reliable medical coding with both many and few data points.

5. Conclusions and Future Work

This study compared the performance of the Mistral and BERT models in identifying the UCOD from death certificates, highlighting Mistral's capabilities in scenarios with varying instance densities. We show that Mistral, particularly when using its instruction-based variant trained with LoRA, showed effective handling of the UCOD identification task, especially in the case of few instances per class. This study is limited by its focus on a specific U.S. dataset and static ICD coding system. Additionally, the performance of Mistral and LoRA may vary with different datasets, coding guidelines, or model configurations not explored here. Thus, future work will expand on these findings by applying the models to additional medical documentation types, and explore the potential for these models to adapt to the evolution of classification systems.

References

[1] WHO. ICD-10 : international statistical classification of diseases and related health problems : tenth revision. 2nd ed. Geneva: World Health Organization; 2004.
[2] Falissard L, Morgand C, Roussel S, Imbaud C, Ghosn W, Bounebache K, et al. A Deep Artificial Neural Network-Based Model for Prediction of Underlying Cause of Death From Death Certificates: Algorithm Development and Validation. JMIR Medical Informatics. 2020 Apr;8(4).
[3] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171-86.
[4] Della Mea V, Popescu MH, Roitero K. Underlying cause of death identification from death certificates using reverse coding to text and a NLP based deep learning approach. Informatics in Medicine Un-locked. 2020;21:100456.
[5] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Advances in NIPS. 2022;35:27730-44.
[6] Touvron H, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv:230709288. 2023.
[7] Jiang AQ, Sablayrolles A, et al. Mistral 7B. arXiv preprint arXiv:231006825. 2023.
[8] Remmer S, Lamproudis A, Dalianis H. Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. In: Proceedings of RANLP 2021. p. 1158-66.
[9] Amin S, Neumann G, Dunfield K, Vechkaeva A, Chapman KA, Wixted MK. MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. In: CLEF; 2019. p. 1-15.
[10] Della Mea V, Popescu MH, Roitero K. Underlying Cause of Death Identification from Death Certificates via Categorical Embeddings and Convolutional Neural Networks. In: ICHI 2020; 2020. p. 1-6.
[11] Fang X, Huang S, Yin Y, Chen T, Liao Z, Zhong W. Advancing Underlying Cause of Death Inference Through Wide and Deep Model. China CDC Weekly. 2024;6(21):487-92.
[12] Yang Z, Kwon S, Yao Z, Yu H. Multi-Label Few-Shot ICD Coding as Autoregressive Generation with Prompt. Proceedings of AAAI CAI. 2023 Jun;37(4):5366-74.
[13] Hegselmann S, Buendia A, Lang H, Agrawal M, Jiang X, Sontag D. TabLLM: Few-shot Classification of Tabular Data with Large Language Models. In: Proceedings of The 26th ICAIS; 2023. p. 5549-81.
[14] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Advances in NIPS. vol. 30. Curran Associates, Inc.; 2017.
[15] Radford A, et al. Improving language understanding by generative pre-training. OpenAI. 2018.
[16] Hu EJ, et al. LoRA: Low-Rank Adaptation of Large Language Models. In: ICLR; 2022.
[17] Biderman D, et al. Lora learns less and forgets less. arXiv:240509673. 2024.