Improving Medical Code Classification for Death Certificates Using Ontology-Adapted Contrastive Loss in BERT Models

Kevin Roitero[1,*], Davide Volpi[1], Riccardo Lunardi[1], Mihai Horia Popescu[1] and Vincenzo Della Mea[1]

[1]*University of Udine, Udine, 33010, Italy*

Abstract

Accurate identification of the Underlying Cause of Death (UCOD) is crucial for informed healthcare policy and planning. The World Health Organization supports the use of the ICD-10 system to standardize the coding of death certificates, a task increasingly supported by automated systems built on top of language models. This study advances the effectiveness of state-of-the-art BERT-based models for UCOD identification by incorporating a novel ontology-adapted contrastive loss function. Extensive experimentation on a dataset from the U.S. National Center for Health Statistics show that BERT models equipped with this specialized contrastive loss function outperform traditional state-of-the-art models.

Keywords

Medical Code Classification, BERT Models, Ontology-Adapted Contrastive Loss, Death Certificates

1. Introduction

The accurate classification of medical codes on death certificates is essential for public health planning and resource allocation. The International Classification of Diseases, Tenth Revision (ICD-10), is the global standard for coding diseases and death certificates [1, 2]. However, the complexity and variability of medical terminology make this task challenging and time-consuming.

Recent advances in natural language processing (NLP), particularly transformer-based models like BERT [3], have led to high accuracy in automating death certificate classification [1, 2, 4], reaching effectiveness scores up to 0.97 on some datasets [5]. Nevertheless, these models still struggle to fully capture the relationships embedded in medical texts and the ICD-10 structure.

This study proposes a novel approach to enhancing BERT models for death certificate coding by integrating an ontology-adapted contrastive loss function. By leveraging the hierarchical structure of ICD-10, this loss guides the model toward a deeper understanding of relationships among medical conditions, improving automated classification effectiveness.

The contributions of this paper are threefold: first, we propose a novel contrastive loss function that incorporates the hierarchical structure of medical classifications, which, to the best of our knowledge, has not been previously explored. We conduct extensive experiments on a large dataset of death certificates from the U.S. National Center for Health Statistics, showing that our adapted models outperform standard BERT models in ICD-10 code classification. Finally, we provide a detailed analysis of model effectiveness and errors across different adaptations, offering insights into the role of loss functions in medical text classification.

The code, models, and data needed to reproduce our results are available at: https://osf.io/4ha2k/.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

[*] Corresponding author.

kevin.roitero@uniud.it (K. Roitero); volpi.davide@spes.uniud.it (D. Volpi); riccardo.lunardi@uniud.it (R. Lunardi); mihaihoria.popescu@uniud.it (M. H. Popescu); vincenzo.dellamea@uniud.it (V. D. Mea)

0000-0002-9191-3280 (K. Roitero); 0009-0005-2651-7889 (D. Volpi); 0009-0001-5550-317X (R. Lunardi); 0000-0003-3378-0368 (M. H. Popescu); 0000-0002-0144-3802 (V. D. Mea)

Table 1
Example of a coded death certificate.

Part 1	***Condition***
1	J69.0 *Pneumonitis due to food and vomit*
2	I48.9 *Atrial fibrillation and atrial flutter, unspecified*
3	I10 *Essential (primary) hypertension*
4	....
Part 2	***Condition***
1	R53 *Malaise and fatigue*
Other	***Administrative data*** Sex: *male* Age: *85*
	Underlying cause of death I10 *Essential (primary) hypertension*

2. Background and Related Work

Many challenges in machine and deep learning are heavily dependent on the capability to effectively learn a distance metric [6]. While most traditional deep learning methods focus on enhancing architectural depth, such as BERT [3], distance-based techniques are gaining prominence in the field [7, 8, 9]. This paper specifically explores the application of deep metric learning through the use of triplet loss. Initially become popular after its usage in FaceNet [10], triplet loss has been extensively applied to a variety of tasks, demonstrating effectiveness in image classification [11, 12], image retrieval [13, 14], and object recognition [10]. Beyond its initial application in face recognition tasks, triplet loss has been effectively used in recent studies to enhance effectiveness in different recognition tasks, such as person re-identification [15, 16], action recognition [17], vehicle recognition [18], place recognition [19], 3d pose recognition [20], and speaker recognition [21].

While triplet loss has demonstrated high effectiveness in different tasks, particularly when negative samples are selected in an informative way, achieving this optimal selection strategy can be practically challenging. In fact, the key to maximizing the benefits of triplet loss in deep metric learning lies in the strategic selection of triplets. Consequently, various methods for triplet selection have been developed across different tasks [22, 11, 23]. For instance, Ge [24] introduced a novel approach called hierarchical triplet loss, which utilizes an adaptively updated hierarchical tree to select informative triplets by encoding global context. In another study, Sumbul et al. [25] proposed a two-step triplet sampling method for a deep learning-based retrieval system, starting with the selection of diverse anchor images followed by the choice of positive and negative images based on relevance, difficulty, and diversity. Moreover, Yu et al. [12] highlighted the potential pitfalls of selecting the hardest triplets, which can lead to poor local minima. To address this, they introduced a new variant of triplet loss designed to minimize bias in triplet selection by adaptively correcting the distribution shift of the chosen triplets. In this paper, we present a new selection method tailored to the hierarchical structure of the ICD-10 classification, exploring its potential to enhance the learning process in a NLP task.

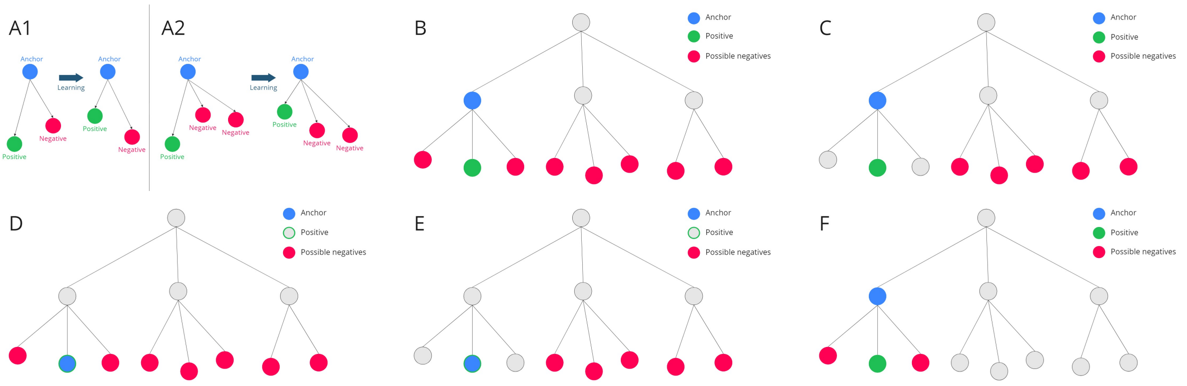
3. Methodology

3.1. Data, Models, and Metrics

For evaluating our model's effectiveness in identifying the underlying cause of death (UCOD), we use a dataset of death certificates from the U.S. National Center for Health Statistics.[1] The dataset contains nearly 13 million records from 2014 to 2017, including detailed medical histories and demographic data. Our experiments follow a two-phase structure. In Phase 1, we apply stratified sampling to create sets of 50,000 training and 5,000 testing instances. In Phase 2, the best-performing models, along with the baseline, are evaluated at scale on 500,000 training and 100,000 testing instances. Following Della Mea et al. [1], each record is pre-processed by converting coded entries into narrative text to better

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

Figure 1: Model loss variations. First image adapted from [10].



capture underlying semantics. For example, the certificate in Table 1 is transformed into: *Male, 85y old: Pneumonitis due to food and vomit due to Atrial fibrillation and atrial flutter, unspecified due to Essential (primary) hypertension in the context of Malaise and fatigue.*

To classify medical codes in death certificates, we leverage BERT [3], a transformer-based model that introduced bidirectional contextualization and set a new standard in natural language understanding. Its self-attention mechanism enables effective modeling of long-range dependencies, making it particularly suitable for the nuanced and context-rich language typically found in death certificate narratives. BERT has achieved state-of-the-art results across many NLP tasks, including death certificate encoding [1, 2].

To assess the effectiveness of our model variations, we use standard classification metrics: we report Accuracy, as well as Precision, Recall, and the F1-Score. To address class imbalance, we compute all metrics using both *micro* (aggregating over all instances) and *macro* (averaging per class) strategies.

3.2. Ontology-Aware Triplet Loss Functions

Triplet loss is a learning objective used to train models to distinguish similarities and differences between inputs [26, 24, 12]. It involves three elements: an anchor (A), a positive example (P), and a negative one (N). The goal is to ensure that the distance between A and P is smaller than that between A and N by at least a margin (see Figure 1A1). This encourages the model to learn fine-grained distinctions. Formally, the triplet loss is defined as: $L(A, P, N) = \max\left(0, d(A, P) - d(A, N) + \text{margin}\right)$, where $d(\cdot, \cdot)$ denotes a distance metric, and the "margin" allows controlling the class separation. We develop the following variants of triplet loss specifically tailored for ontology-based medical classification (Figure 1).

- **Version 1**: The anchor is the textual description of the node's parent label. The positive example shares the same label as the input; the negative is randomly selected from different labels. (Figure 1B).
- **Version 2**: Same as Version 1, but the negative example is selected from non-sibling nodes, avoiding overly similar negatives and promoting better inter-class discrimination (Figure 1C).
- **Version 3**: The anchor and positive are both nodes with the same label, emphasizing exact label matching. The negative remains randomly selected from different labels (Figure 1D).
- **Version 4**: Similar to Version 3, but negatives are explicitly chosen to not be siblings of the input label (Figure 1E).
- **Version 5**: The anchor is again the parent's textual description; positives match the label; negatives are sampled from sibling nodes. This deliberately challenges the model to distinguish among closely related classes (Figure 1F).
- **Version 6**: Extends to a quartet structure: the anchor is the parent's description, the positive matches the input label, and two negatives are selected from non-sibling, non-related nodes. This variation exposes the model to a broader range of harder negative examples, enhancing discrimination across distant categories (Figure 1A2).

Table 2
Metrics on whole dataset.

	Whole Dataset										Category Level									
Variant	Acc	F1$_{\mathrm{M}}$	P$_{\mathrm{M}}$	R$_{\mathrm{M}}$	F1$_{\mathrm{m}}$	P$_{\mathrm{m}}$	R$_{\mathrm{m}}$	F1$_{\mathrm{W}}$	P$_{\mathrm{W}}$	R$_{\mathrm{W}}$	Acc	F1$_{\mathrm{M}}$	P$_{\mathrm{M}}$	R$_{\mathrm{M}}$	F1$_{\mathrm{m}}$	P$_{\mathrm{m}}$	R$_{\mathrm{m}}$	F1$_{\mathrm{W}}$	P$_{\mathrm{W}}$	R$_{\mathrm{W}}$
BERT	.870	.363	.342	.416	.870	.870	.870	.838	.819	.870	.892	.458	.451	.494	.892	.892	.892	.871	.860	.892
V1	.872	.379	.355	.437	.872	.872	.872	.842	.824	.872	.892	.476	.468	.515	.892	.892	.892	.873	.863	.892
V2	.874	.382	.359	.436	.874	.874	.874	.843	.824	.874	.896	**.491**	.486	**.525**	.896	.896	.896	**.877**	**.867**	.896
V3	.872	.370	.347	.424	.872	.872	.872	.840	.820	.872	.896	.477	.471	.510	.896	.896	.896	.875	.864	.896
V4	**.876**	**.388**	**.367**	**.440**	**.876**	**.876**	**.876**	**.845**	**.827**	**.876**	**.897**	.486	**.481**	.520	**.897**	**.897**	**.897**	**.877**	.866	**.897**
V5	.872	.359	.337	.414	.872	.872	.872	.841	.822	.872	.893	.455	.451	.492	.893	.893	.893	.872	.862	.893
V6	.874	.378	.356	.431	.874	.874	.874	.843	.823	.874	.896	.463	.452	.501	.896	.896	.896	.876	.864	.896

4. Results

The evaluation metrics for the whole dataset are summarized in Table 2. Overall, we can see that, all model variants display robust accuracy, with scores ranging from 0.870 to 0.876. Notably, V4 achieves the highest overall accuracy at 0.876. This suggests a slight improvement in generalizing capabilities over the baseline BERT model, which scores 0.870.

In terms of F1 scores, which balance precision and recall, the macro and micro-averaged results are also indicative of effectiveness differences between the variants. The macro F1 scores, important for evaluating effectiveness on imbalanced datasets, vary from 0.35 to 0.388. Again, V4 stands out with the highest macro F1 score at 0.388, indicating superior effectiveness also in terms of less populated classes compared to other versions and the baseline model. Micro F1 scores, which give a better indication of overall effectiveness, closely mirror the accuracy metrics, underscoring consistent effectiveness across versions and peaking with V4. Further analysis into precision and recall (weighted, macro, and micro) confirms the trends observed in the F1 scores, with V4 consistently leading in most metrics, suggesting an optimal balance of recall and precision among all variants.

When shifting focus to category-level effectiveness, the metrics show similar trends but with generally higher scores, reflecting the models' capabilities to adapt to specific category characteristics. Accuracy figures remain high, with all variants performing above 0.892, and V4 again showing higher effectiveness scores reaching an accuracy of 0.897. Macro and micro F1 scores at the category level are also higher than those observed in the whole dataset evaluation, highlighting the effectiveness of the models on specific categories. Notably, V2) and V4 perform exceptionally well in this regard, with macro F1 scores of 0.491 and 0.486, respectively. These improvements suggest that adaptations made in these versions are particularly effective at enhancing classification in more narrowly defined categories.

We also evaluated our models on two dataset variations and a larger dataset to test their robustness under different conditions. Version 1 removes "unspecified" from the text, and Version 2 excludes ICD-10 codes ending in ".9", both aimed at reducing label ambiguity; in these settings, our adapted models, particularly V2 and V6, maintained or improved effectiveness. On the larger dataset, V2 consistently outperformed the BERT baseline across all metrics—achieving an accuracy of 0.966 vs. 0.964 and a macro F1 score of 0.443 vs. 0.438 at the whole-dataset level, and 0.970 vs. 0.969 in accuracy and 0.575 vs. 0.567 in macro F1 at the category level. Overall, the results confirm that the tailored contrastive loss functions integrated into the BERT models significantly improve classification accuracy and precision across both broad and specific categories over the baseline model. This suggests the efficacy of the ontology-adapted loss function in dealing with the task of medical code classification for death certificates.

5. Error Analysis

By examining the types and frequencies of errors, we assess model biases and areas for improvement. Our analysis focuses on identifying frequent misclassifications across BERT-based models using ontology-adapted contrastive loss, aiming to evaluate whether these specialized losses reduce classification errors compared to the baseline BERT. We also compute the average ontology distance between incorrect predictions and the true labels, measuring whether errors occur between semantically close or distant

Table 3
Errors on the whole dataset.

Code	Description	BERT	V1	V2	V3	V4	V5	V6
J44.9	Chronic obstructive pulmonary disease, unspecified	10	8	7	9	10	9	5
I25.1	Atherosclerotic heart disease	7	8	5	6	5	6	–
J18.9	Pneumonia, unspecified	6	6	8	6	4	6	5
C97	Malignant neoplasms of independent (primary) multiple sites	6	7	6	7	7	5	4
K70.4	Alcoholic hepatic failure	3	–	–	–	–	–	–
G30.9	Alzheimer disease, unspecified	–	6	–	–	–	–	–
I21.9	Acute myocardial infarction, unspecified	–	–	5	–	–	–	–
E43	Unspecified severe protein-energy malnutrition	–	–	–	5	–	–	–
E14.7	Unspecified diabetes mellitus with multiple complications	–	–	–	–	3	–	–
I26.9	Pulmonary embolism without mention of acute cor pulmonale	–	–	–	–	–	4	4
I10	Essential (primary) hypertension	–	–	–	–	–	–	3

concepts.

On the whole dataset, the baseline BERT model yields an average error distance of 5.926, while variants like V1 and V2 reduce it to 5.905 and 5.813, respectively, indicating improved clustering of related medical conditions. On the larger dataset, V2 maintains strong performance with an error distance of 6.064, slightly higher but still below the BERT baseline, demonstrating effective scalability. These results confirm that the ontology-adapted contrastive loss not only improves classification accuracy but also steers models toward semantically meaningful misclassifications within the ontology.

In addition to qualitative analysis, we performed a quantitative study of model errors. Table 3 shows that different model variants exhibit distinct error patterns across medical codes. For example, while all models struggled with chronic obstructive pulmonary disease (Code J44.9), error rates decreased in specialized variants like V6. However, performance varied across conditions: errors on atherosclerotic heart disease (I25.1) and pneumonia (J18.9) differed significantly among models, highlighting how model adaptations influenced the recognition of specific medical concepts.

6. Conclusions and Future Work

This study demonstrates the potential of enhancing BERT models for medical code classification on death certificates by integrating an ontology-adapted contrastive loss function. By leveraging the hierarchical structure of ICD-10, our models improve classification accuracy, reduce semantic errors, and contribute to advancing health informatics applications, with possible extensions to diagnostic imaging and patient history analysis.

Future work will focus on extending this approach to other medical documents, such as patient records and discharge letters, and on testing generalization across diverse data sources like electronic health records. Further adaptations to other medical ontologies, such as SNOMED-CT, will also be explored. Additionally, developing specialized modules for handling unspecified labels could enhance model robustness while maintaining high accuracy.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to check grammar and spelling. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] V. Della Mea, M. H. Popescu, K. Roitero, Underlying cause of death identification from death certificates using reverse coding to text and a nlp based deep learning approach, Informatics in

Medicine Unlocked 21 (2020).

[2] L. Falissard, C. Morgand, S. Roussel, C. Imbaud, W. Ghosn, K. e. a. Bounebache, A deep artificial neural network- based model for prediction of underlying cause of death from death certificates: algorithm development and validation, Medical informatics 8 (2020).

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL, 2019.

[4] V. Della Mea, M. H. Popescu, K. Roitero, Underlying cause of death identification from death certificates via categorical embeddings and convolutional neural networks, in: ICHI, 2020.

[5] K. Roitero, B. Portelli, M. H. Popescu, V. D. Mea, Dilbert: Cheap embeddings for disease related medical nlp, IEEE Access 9 (2021) 159714–159723.

[6] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, Journal of Machine Learning Research (2009) 207–244.

[7] Y. Duan, J. Lu, J. Feng, J. Zhou, Deep localized metric learning, IEEE TCSVT (2018).

[8] G. Dai, J. Xie, F. Zhu, Y. Fang, Deep correlated metric learning for sketch-based 3d shape retrieval, AAAI CAI 31 (2017).

[9] Z. Li, J. Tang, Weakly supervised deep metric learning for community-contributed image retrieval, IEEE TMM (2015).

[10] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: IEEE CVPR, 2015, pp. 815–823.

[11] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, Learning fine-grained image similarity with deep ranking, in: CCVPR, 2014.

[12] B. Yu, T. Liu, M. Gong, C. Ding, D. Tao, Correcting the triplet selection bias for triplet loss, in: ECCV, 2018.

[13] J. Huang, R. Feris, Q. Chen, S. Yan, Cross-domain image retrieval with a dual attribute-aware ranking network, in: 2015 IEEE ICCV, 2015.

[14] B. Zhuang, G. Lin, C. Shen, I. Reid, Fast training of triplet-based deep binary embedding networks, in: CVPR, 2016.

[15] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, in: CVPR, 2016.

[16] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, Joint learning of single-image and cross-image representations for person re-identification, in: CVPR, 2016.

[17] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rossenberg, L. Fei-Fei, Learning semantic relationships for better action retrieval in images, in: CVPR, 2015.

[18] H. Liu, Y. Tian, Y. Wang, L. Pang, T. Huang, Deep relative distance learning: Tell the difference between similar vehicles, in: CVPR, 2016.

[19] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, J. Sivic, Netvlad: Cnn architecture for weakly supervised place recognition, IEEE TPAMI (2018).

[20] P. Wohlhart, V. Lepetit, Learning descriptors for object recognition and 3d pose estimation, in: CVPR, 2015.

[21] A. Wirdiani, S. N. Machetho, I. K. G. D. Putra, M. Sudarma, R. S. Hartati, H. A. Ferdian, Improvement model for speaker recognition using mfcc-cnn and online triplet mining, IJAEIT 14 (2024).

[22] Y. Cui, F. Zhou, Y. Lin, S. Belongie, Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop, in: CVPR, 2016.

[23] L. Wang, Y. Li, S. Lazebnik, Learning deep structure-preserving image-text embeddings, in: CVPR, 2016.

[24] W. Ge, Deep metric learning with hierarchical triplet loss, in: ECCV, 2018.

[25] G. Sumbul, M. Ravanbakhsh, B. Demir, Informative and representative triplet selection for multilabel remote sensing image retrieval, IEEE TGSR (2022).

[26] X. Dong, J. Shen, Triplet loss in siamese network for object tracking, in: ECCV, 2018.