Neural Processing Letters (2024) 56:177
https://doi.org/10.1007/s11063-024-11526-y

Multi-step Transfer Learning in Natural Language Processing for the Health Domain

Thokozile Manaka[1] · Terence Van Zyl[2] · Deepak Kar[3] · Alisha Wade[4]

Accepted: 8 January 2024 / Published online: 20 May 2024

Abstract
The restricted access to data in healthcare facilities due to patient privacy and confidentiality policies has led to the application of general natural language processing (NLP) techniques advancing relatively slowly in the health domain. Additionally, because clinical data is unique to various institutions and laboratories, there are not enough standards and conventions for data annotation. In places without robust death registration systems, the cause of death (COD) is determined through a verbal autopsy (VA) report. A non-clinician field agent completes a VA report using a set of standardized questions as guide to identify the symptoms of a COD. The narrative text of the VA report is used as a case study to examine the difficulties of applying NLP techniques to the healthcare domain. This paper presents a framework that leverages knowledge across multiple domains via two domain adaptation techniques: feature extraction and fine-tuning. These techniques aim to improve VA text representations for COD classification tasks in the health domain. The framework is motivated by multi-step learning, where a final learning task is realized via a sequence of intermediate learning tasks. The framework builds upon the strengths of the Bidirectional Encoder Representations from Transformers (BERT) and Embeddings from Language Models (ELMo) models pretrained on the general English and biomedical domains. These models are employed to extract features from the VA narratives. Our results demonstrate improved performance when initializing the learning of BERT embeddings with ELMo embeddings. The benefit of incorporating

✉ Thokozile Manaka
thokozilemanaka@gmail.com

Terence Van Zyl
tvanzyl@uj.ac.za

Deepak Kar
deepak.kar@wits.ac.za

Alisha Wade
Alisha.Wade@wits.ac.za

1 School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, Gauteng, South Africa

2 Institute for Intelligent Systems, University of Johannesburg, Johannesburg, Gauteng, South Africa

3 School of Physics, University of the Witwatersrand, Johannesburg, Gauteng, South Africa

4 MRC/Wits Rural Public Health and Health Transitions Research Unit, School of Public Health, University of the Witwatersrand, Johannesburg, Gauteng, South Africa

character-level information for learning word embeddings in the English domain, coupled with word-level information for learning word embeddings in the biomedical domain, is also evident.

Keywords Transfer learning · Verbal autopsy · Natural language processing · Text classification · Feature extraction · Fine tuning

1 Introduction

Most underdeveloped and developing countries lack robust death registration systems, and more than half of the 60 million annual deaths go unrecorded because they occur outside medical facilities [1, 2]. A verbal autopsy (VA) is a tool that can offer information about a cause of death (COD) in these places. Two parts make up a VA report: structured data and unstructured data. The structured data is made up of quantitative features like age and binary features, which are "yes" and "no" responses to disease symptoms. An open-ended narrative text outlining events leading up to death makes up the unstructured part.

The adoption of natural language processing (NLP) techniques in the automation of coding of textual data has advanced NLP applications in the English domain. Still, these advancements have seen slower progression in the medical domain [3]. This is caused by restricted access to health information due to patient privacy and confidentiality policies. Additionally, clinical data lacks annotation conventions and standards, as it varies across different institutions and laboratories [4].

In transfer learning, knowledge from domains, languages and tasks where data are abundant can be used in domains where data are limited via adaptation techniques of feature extraction and fine-tuning [5–8]. Kim [9] showed that language modelling has been widely adopted as a source task for transfer learning and has helped advance NLP techniques. Language models possess knowledge about how language is structured and represented, and several NLP tasks share common knowledge about linguistic representation. This shared knowledge can inform each other on semantics and syntax of language [10].

This study presents Multi-Step Transfer Learning, a framework that improves the text classification task in the health domain. The model builds upon NLP transfer learning techniques of ELMo (Embeddings from Language Models) and BERT (Bidirectional Encoder Representations from Transformers). VA embeddings learned from ELMo trained in the English domain are used to initialize the learning of VA embeddings BERT trained in the biomedical domain. The resultant embeddings are used for the downstream task of COD classification.

This work is structured in the following way: A review of earlier works on the automation of COD from VA reports is presented in Sect. 2. Techniques that handle a class imbalance in NLP applications are also discussed here. Section 3 gives a comprehensive description of the data and introduces the experimental design of the proposed technique. This section also outlines the Multi-Step Transfer Learning technique's parameter settings and performance evaluation measures. Section 4 discusses the experimental findings and limitations of the study, while Sect. 5 concludes the study and outlines the planned future research directions.

2 Background

Clinical natural language processing (NLP) is rapidly advancing in healthcare and medical research. It involves applying NLP techniques to clinical and biomedical texts, such as elec-

tronic health records (EHRs), medical literature, and other healthcare documents to extract meaningful information to improve healthcare outcomes.

Medical images, like radiology and pathology images, along with their reports, also play a crucial role in clinical diagnosis and treatment [11, 12]. However, creating medical reports, typically paragraphs detailing normal and abnormal findings, can be time-consuming and error-prone for both experienced and inexperienced radiologists [13].

Liu et al. [14] show that existing medical report-generation techniques often rely on supervised approaches, requiring paired image-report data, which can be resource-intensive in the medical field. To address this, they proposed an unsupervised Knowledge Graph Auto-Encoder model that utilizes independent sets of images and reports during training. This model establishes a shared latent space through a knowledge graph, connecting visual and textual domains.

A "patient instruction" (PI) is a set of important directions given to both caregivers and patients when they leave hospital. Liu et al. [15] proposed a novel task of automatic PI generation, built a PI dataset, and presented a deep-learning approach named Re^3Writer, which imitates physicians working patterns to automatically generate a PI at the point of discharge from the hospital.

The field of question-answering (QA) has been transformed by recent advancements in large language models, but evaluating LLMs in the medical field is challenging due to a lack of standardized datasets [16]. Existing medical datasets [17–19] for LLM evaluation often have limitations of size that hinder thorough assessments. Many are sourced from potentially biased online forums customer service feedback surveys and lack diversity, especially in non-English languages due to resource inequality in NLP [20–22].

Overall, the insufficiency of well-curated evaluation datasets has impeded the evaluation of LLMs in the medical domain. In response to this, Liu et al. [23] introduced CMExam, a dataset derived from the Chinese National Medical Licensing Examination, serving as a benchmark for LLM performance in medical question-answering tasks, including answer prediction and reasoning.

Liu et al. [24] emphasizes that the dependence of the majority of neural networks on supervised learning means their effectiveness is impacted by the accessibility and quality of labeled data. This poses a particular challenge for rare conditions such as emerging pandemics. The Medical multi-modal large language model (Med-MLLM) was introduced as a solution for learning radiograph representations from unlabeled data. Experiments of it on COVID-19 datasets showed it's adaptability to rare diseases, and it's efficiency in handling both visual (X-rays, CT scans) and textual (medical reports) information [24].

With transfer learning [25], information learned from domains with large datasets can be applied to tasks, languages, or domains with smaller datasets. There are two steps to transfer learning: pretraining, where general-purpose language representations are learned, and adaptation, where the learned features are applied to a new task or domain. Clinical text representation models based on transfer learning have been developed in the health domain to boost NLP tasks of mortality prediction and hospital readmission. These models include ClinicalBERT [26], which models hospital readmission from clinical notes, MeDAL [27], a huge medical text sample compiled for abbreviation extraction for medical domain pretraining and the Publicly Available Clinical BERT Embeddings [28]. There is also the BERT model applied on clinical notes and discharge summaries, BioBERT [29] and BioELMo [30], the ELMo model applied to biomedical literature and SciBERT [31], the BERT model which is trained on scientific literature.

The datasets used to train these models include records from clinics and hospitals, MIMIC-II and MIMIC-III (Medical Information Mart for Intensive Care) [32]. Other datasets are

from PubMed,[1] a biomedical literature database that provides abstracts of published articles and PubMed Central (PMC),[2] a full-text repository provides the full text of the database's publications. The target tasks include disease prediction, diagnosis, prediction of hospital readmission, and prediction of mortality. These tasks are related to patient information from hospital care, such as vital signs, clinical features, medications, and investigations, as documented by clinical care providers

The differences and similarities between the source and target tasks have been identified as critical properties influencing the performance of domain adaptation approaches involving feature extraction and fine-tuning. While both techniques have been found to perform similarly in the English domain, their performance changes when the training objectives and target tasks are either relatively similar or significantly different [33]. In the health domain, the same holds true, where both BioELMo and BioBERT representations have demonstrated effectiveness in biomedical tasks such as named entity recognition (NER) and natural language inference (NLI). In these tasks, BioELMo has exhibited superior performance as a feature extractor compared to BioBERT [34]. Jin et al. [34] attribute this to BioELMo's efficacy in encoding entity types and biomedical relationship details, such as correlations between symptoms and diseases. In the general domain, ELMo has proven to be a superior feature extractor for cases involving similar source and target tasks, while BERT (fine-tuning) excels when dealing with distinct source and target tasks [33].

2.1 Multi-source Domain Adaptation

Zhao et al. [35] indicated that transferring models directly between different domains causes severe performance decline due to domain shift. Domain shift [36, 37] refers to the situation where there are differences in the joint probability distributions of observed data and labels between two domains.

Domain adaptation is a paradigm that aims to mitigate the effects of domain shifts between the source and target domains. One way it does this is by aligning the source and target domains. Multi-source domain adaptation (MSDA) is a powerful extension of this concept, and it leverages knowledge from multiple sources with diverse distributions to an unlabeled target domain.

Having multiple source domains available for training in real-life applications is not unexpected. Thus, in this case, utilizing only one source domain for training would seem inefficient. The typical approach is to treat all the sources as a single source, disregarding their differences. Another alternative is to train a classifier for each source and then combine these classifiers [38]. The study also demonstrated that applying risk minimization principles allows for assigning weights to base models, enabling the combination of multiple base models to enhance the performance accuracy in a new domain [39].

Reimer et al. [40] showed that this task poses a challenge due to the substantial domain shift that occurs not only between the target and source domains but also among the various source domains. These variations can potentially interfere with each other during the learning process.

Sun et al. [39] also showed that while using abundant training data can benefit systems, conflicting properties in data sources can lower a single model's performance. The authors suggested that while training separate systems would be ideal, data sources often have shared characteristics despite their differences. A single system hides differences, while separate

[1] https://pubmed.ncbi.nlm.nih.gov/.

[2] https://www.ncbi.nlm.nih.gov/pmc/.

systems ignore similarities. Given that the source and target domains differ and contain domain-specific and common features, the authors demonstrated that it is possible to establish mappings from the original feature space to a latent feature space shared between the domains.

Sun et al. [39] surveyed earlier MSDA methods that mainly focused on shallow models, and the models were grouped into those that learned a latent feature space for various domains [41], and those that combined pre-learned classifiers [42].

The latest deep learning MSDA can be categorized into two groups depending on the techniques employed for alignment: latent space transformation and intermediate domain generation. Latent space transformation techniques align the target and source feature features to make them appear similar to discriminators. The primary goal of these methods is to confuse the discriminator, preventing it from accurately determining whether the features originated from different sources or were sampled from the same distribution [35]. Other latent space transformation techniques directly quantify the differences between latent spaces, representing features across domains. They accomplish this by optimizing specific discrepancy losses, such as maximum mean discrepancy and the Renyi-divergence [43].

Intermediate domain generation techniques aim to overcome the limitation of feature-level alignment, particularly in computer vision. This has been demonstrated in their ability to align only high-level information, which may not be adequate for precise predictions like pixel-wise semantic segmentation [44]. Zhao et al. [35] illustrated that this challenge can be overcome by generating an intermediate adapted domain using GANs and achieving pixel-level alignment.

Existing deep learning architectures for MSDA have predominantly concentrated on scenarios involving a single-target domain. Li et al. [45] also demonstrated that existing CNN-based methods were primarily designed for single-task applications. The tasks of image segmentation [46] and landmark localization [47] are significant in diagnosing and treating knee-related illnesses. Given the intricate nature of the 3D knee MRI analysis problem, encompassing both image segmentation and landmark localization tasks, which play critical roles in diagnosing and treating knee diseases, these techniques were found to be insufficient. The authors designed a Spatial Dependence Multi-Task Transformer, which incorporates spatial encoding into the features and introduces a multi-head attention mechanism that combines tasks. This attention mechanism comprises two types of attention heads: inter-task attention heads, which manage spatial interdependence between tasks, and intra-task attention heads, which handle correlations within individual tasks.

Wan and Jiang et al. [48] presented TransCrispr. This hybrid deep neural network comprises four components: Embedding, CNN, Transformer, and a Multilayer Perceptron (MLP) with Fully Connected layers for the prediction of CRISPR/Cas9 single guide RNA cleavage efficiency. MSDA has also been shown to enhance the COD classification task in VA reports [49]. In their study, Manaka et al. [49] demonstrated that better performance could be achieved by combining the English and biomedical domains for learning representations from the VA corpus. They incorporated character-level information for learning VA embeddings in the English domain and word-level information for learning VA embeddings in the biomedical domain, resulting in improved results compared to using embeddings from these domains separately.

2.2 ELMo

ELMo is a context-dependent language model that generates word embeddings by considering the word's context in both directions by using a shallow bidirectional LSTM architecture.

ELMo tends to learn more generic linguistic features such as syntax, semantics and some contextual information, but it might not capture fine-grained contextual details as effectively as BERT.

ELMo makes use of character-level information in addition to word-level information. It uses character embeddings to represent each character in a word and employs a CNN to process the sequences of character embeddings. The CNN operates over a fixed-sized sliding window across the character embeddings. This window captures local patterns and interactions between characters. These character embeddings are combined to form word representations, which are then used as inputs for the bidirectional language model [50].

Character-level information has been shown to improve text classification models compared to word-level information and has been used for some time in developing word embeddings. ELMo is one such model that uses character-level information. ELMo, with its variants like BioELMo, therefore, can capture syntactic, morphological, and orthographic information at the character level, enhancing the model's generalization on both frequent and unseen words. Additionally, they offer the benefit of representing out-of-vocabulary words and misspelled terms [51]. They can also learn long-term context dependency, critical for VA reports where each case typically comprises three to five related sentences.

2.3 BERT

The core make-up of a BERT model is based on the Transformer architecture. The vanilla Transformer [52] is a sequence-to-sequence model and consists of an encoder and a decoder, each of which is a stack of L identical blocks. Each encoder block comprises a multi-head self-attention module and a position-wise feed-forward network (FFN). For building a deeper model, a residual connection [53] is employed around each module, followed by a layer normalization [54] module. Compared to the encoder blocks, decoder blocks insert cross-attention modules between the multi-head self-attention modules and the position-wise FFNs. Furthermore, the self-attention modules in the decoder are adapted to prevent each position from attending to subsequent positions.

In contrast to ELMo, BERT's training objective is a masked language modeling, which involves randomly replacing words in a phrase with a particular token and using a transformer to create a prediction for the token, taking into account the unmasked words surrounding it [6]. The other pretraining task it uses is next-sentence prediction, which can be thought of as a form of sentence modeling. For BERT, the Transformer architecture strictly utilizes the workpiece tokenization method instead of ELMo, which combines character and word tokenizations. The BERT model family consists of BERT Experts, which are made up of eight models that all feature the BERT-base architecture but offer a selection of different pretraining domains to better align with the target task. The significant benefits achieved by BERT led several modern representation models to adopt the Transformer architecture as their main building element. Compared to ELMo, which uses a shallow bidirectional architecture, BERT uses a deep bidirectional architecture.

Combining character-level and word-level information word embeddings has been proven to boost the performance of various text classification tasks. Character-level CNNs were employed on a dataset of news articles and a dataset of online reviews to show that the character-level model outperformed the word-level one on the classification task [55]. In a similar study, character-level and word-level information embeddings were merged with padding and a Long Short-Term Memory (LSTM) language model to produce better perplexity scores than comparable word-based models [56]. An LSTM was utilized in the health

domain to learn character embeddings that were later coupled with pretrained word embeddings to retrieve information about cancer, results of which successfully showed that coupling character and word-based techniques for COD is effective on the medical domain [57].

Recent studies have shown that character-level information can improve text classification models, especially in cases with numerous spelling errors and variants, such as the VA text [58]. Yan et al. [58] introduced two CNN based methods, namely embedding concatenation and model combination to combine word and character embeddings [58]. With these methods, the authors demonstrated that information about characters can overall improve the COD classification task for VAs and datasets that are relatively smaller in size. They further showed that an added benefit to character-based models is their smaller vocabulary size, which causes the input representations to have small variations. This trait is especially useful for very small datasets like VA narratives.

A BERT-ELMO-based deep learning neural network architecture that utilised a bidirectional LSTM (BiLSTM) as the primary building block, together with a conditional random fields layer was used for the name entity recognition (NER) task by Affi and Latiri [59]. The authors initialized word vectors using pretrained ELMo and BERT embeddings and fed the output into a BiLSTM network. They reported improved results compared to existing state-of-the-art (SOTA) systems on Conference on Natural Language Learning 2003 shared task (CoNLL-2003) (95.56% F1-Score). Building upon a similar concept, characterBERT, a BERT version that consults the characters of words to represent them using a character CNN module, was introduced for the NER task in the health domain [51].

While the self-attention mechanism of the Transformer has proven to be effective for many language models, it does come with some challenges. The self-attention used by the Transformer is known to be complex [60], resulting in the attention module becoming a bottleneck when dealing with long sequences. The second challenge pertains to structural priors. Compared to CNNs and RNNs, which come with predefined biases for spatial or temporal patterns in data, the self-attention mechanism used in the Transformer lacks specific structural biases and assumptions about data. Even the order information needs to be learned from the training data. Consequently, the Transformer's design is more flexible and is better equipped to handle diverse tasks effectively. However, this flexibility comes at the cost of potential overfitting on smaller datasets [60].

Several works have set out to mitigate this challenge by improving attention. This includes techniques that utilize a prior distribution for attention. These techniques investigate supplementing or substituting the standard attention mechanism with prior attention distributions. Combining these two attention distributions typically entails computing a weighted total of the scores linked to the prior and the generated attention and applying a softmax function [60].

Text data has been shown to favour locality strongly, and this property can be encoded as prior attention. Lin et al. [60] shows that utilising a Gaussian distribution over positions is the most straightforward approach. This means the resulting attention distribution could be multiplied by a Gaussian curve's density and adjusted to maintain proper proportions. This adjustment is akin to introducing a bias term to the initial attention scores, where larger values suggest a greater inherent likelihood that the i-th input should prioritize attending to the j-th input. Gaussian Transformers by Guo et al. [61] and Yang et al. [62] explored this approach.

BERT's combination of bidirectional context, self-attention mechanisms, positional encodings, deep architecture, and pretraining all contribute to its ability to handle long-range dependencies in language, making it effective at capturing relationships between words or tokens that are distant from each other in a text sequence [6]. The integration of CNNs with

Table 1 A verbal autopsy narrative

Narrative
The deceased started illness on the left leg where she was scratched on the left leg on her toe. The toe became swollen and rotten. She suffered with blurred. The chest pain and headache still worse. She was taken to a special doctor. The treatment was tablets. Bandage was use on her leg. Illness became worse where she was taken to matikwana hospital. Admitted for almost a month. Treatment was water drip, tablets and bandage. She was told that she had sugar diabetes. Urine was red and her toe became swollen and rotten. She was taken to mapulaneng hospital, where her leg was amputated. Then one month after she complained about diarrhea for 3 weeks and it did not stop. She started coughing for one months. She became difficult breathing for 2 days and she died at home. She had trouble seeing for about 2 years and did not stop until death.
Diagnosis: Death due to uncontrollable hyperglycaemia

Transformer architectures to enhance performance is a topic that has been explored in various works, including those in the biomedical domain. CNNs effectively capture the text data's unique characteristics, local patterns, and features. Both CNNs and Transformers have their strengths, and combining them can lead to improved performance by capturing different types of features and patterns.

In Chinese text, spacing between words is not as clear as in English, making boundaries between terms less distinct. Constructing a Chinese entity involves various symbols, characters, and abbreviated forms. Moreover, the structure of Chinese grammar is intricate, leading to instances where a single term can signify distinct types of entities within different contexts [63, 64].

To extract fine-grained semantic features of Chinese characters for the Chinese clinical named entity recognition (NER) task, Wang et al. [63] combined a dynamic fusion transformer layer with the Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking (RoBERTa-wwm) and 1-dimensional CNNs. Kong et al. [64] introduced an innovative approach that integrates multi-level CNN layers with an attention mechanism for the Chinese clinical NER task. Through this approach, they demonstrated the development of a data augmentation technique without relying on external information while also utilizing multi-modal character embeddings to delve into a wider range of semantic details.

2.4 Verbal Autopsy

A verbal autopsy (VA) report is a research tool that enables a better understanding of COD. Unlike clinical notes such as hospital discharge summaries or biomedical text from medical literature, by the nature of its collection and compilation, narrative text from VA reports does not possess clinical or biomedical knowledge. VA reports are performed by lay interviewers and later coded by physicians for COD. Many errors are made while translating the local languages and converting handwritten documents to electronic mediums. Numerous accounts frequently contain grammar and spelling mistakes, inconsistent pronouns, sentence fragments, improper punctuation, transcription problems, and the frequent usage of terminology in the local vernacular [58]. Table 1 shows a sample of a VA narrative.

The language and dialects in VA data might not be well-represented in standard pretrained language models, and these language models might not understand the characteristics and

vocabulary of the specific language used in the VA data. VA narratives also contain domain-specific terms related to cultural practices, local beliefs, and regional medical conditions that may not be covered by standard health domain pretrained models. The data is also narrated by individuals who are not medical experts. As a result, the language used may not conform to the formal medical terminology used in health domain pretrained models.

To mitigate these challenges, we propose a hybrid transfer learning framework that offers cross-linguistic adaptation; The ELMo language model pretrained in the English domain serves as an intermediate bridge between the language used in the VA corpus and the biomedical domain of BERT. The ELMo embeddings capture linguistic characteristics that are present in the VA data and not covered by traditional health domain pretrained models.

Our framework also offers a domain-specific initialization; By using the ELMo-initialized BERT model, we are effectively initializing the model's parameters to understand both the linguistic characteristics of the VA corpus and the biomedical terminology from the pretrained BERT model. This dual initialization enhances the model's ability to handle the unique VA language.

This framework also recognizes the multilingual nature of the VA corpus and incorporates knowledge from multiple domains. The ELMo-initialized BERT model is designed to learn from both English and biomedical contexts, enabling it to better capture the features of the VA corpus language.

We relied on expertly annotated VA data. A pediatrician with expertise in type-1 diabetes coded the data by examining VAs to identify features indicative of diabetes or uncontrolled hyperglycemia. A colleague of the pediatrician, experienced in adult internal medicine, diabetes, and endocrinology, reviewed cases where the reviewing physician was uncertain, and a consensus was reached. This contribution helped guide the hybrid model's initialization, aligning the model with the biomedical knowledge present in the VA language.

2.5 Data Class Imbalance

Due to the high dimensionality of the numerical vectors generated from text, current sampling techniques, such as the synthetic minority oversampling technique (SMOTE) and its variants, like SMOTE-Tomek Links, do not perform well with text data. Although BERT is capable of handling imbalanced classes without the need for additional data augmentation, it has been shown that the model does not generalize well when the training and testing datasets are different, such as news sources whose subjects change over time. To address this challenge, Madabushi et al. [65] suggest adding cost weighting to BERT.

Wei et al. [66] showed that data augmentation strategies for NLP, such as synonym substitution and random insertion, deletion and swapping of words from a sentence with a predetermined frequency, did not produce significant gains when using pretrained models. To test this assertion, Madabushi et al. [65] used various data augmentation techniques including synonym replacement, random deletion of words in a sentence, and the random oversampling of cases from the minority class, for the sentence classification task using the BERT language model. According to the authors, except for oversampling, BERT without data augmentation approaches outperformed BERT with those techniques. In contrast to synonym insertion and random word deletion, which inject noise into the data, oversampling does not, according to the authors. In the case of natural language data, this type of noise may modify a sentence's meaning. The cost-sensitive classification was then presented as a more reliable technique for weighing samples of imbalanced data.

Cost-sensitive classification

Cost-sensitive learning solves the problem of class imbalance by changing the cost function of the model such that making incorrect classifications of training samples from the minority class are more costly.

If x_i is a single prediction and j a class, the cross entropy (CE) loss for the class is given by

$$\mathrm{CE} = -\frac{1}{N}\sum_{i}\sum_{j\epsilon\{0,1\}} y_{ij} log p_{ij} \tag{1}$$

where x_i is a member of a set of training examples X and is related to a label y_i, which is a member of the set {0, 1}. The predicted probabilities of the classes is p_i, and it is a member of [0, 1].

One can adjust the cross-entropy loss to take into account an array $weights$, the i^{th} member of which gives the weight of the i^{th} class to be

$$\text{Weighted CE} = -\frac{1}{N}\sum_{i}\alpha_i \sum_{j\in\{0,1\}} y_{ij} log p_{ij} \tag{2}$$

where α_i, a member of a set [0, 1] is set by the inverse class frequency.

Although the weighted cross entropy loss can offer some relief, the improvement is not as significant, according to Xiaoya [67]. A dice loss has been proposed as a solution to address the class imbalance due to the limitations of the cross entropy loss in cases with uneven label distributions.

Dice Loss/Sorensen–Dice Coefficient

The dice loss is based on a statistic that gauges the similarity between two samples or an overlap between two sets called the Sorensen–Dice coefficient (DSC) [68]. For a single example x_i, the dice coefficient is given as

$$DSC(x_i) = \frac{2p_{i1}y_{i1}}{p_{i1} + y_{i1}} \tag{3}$$

The nominator and denominator of the above equation are smoothed by adding a γ factor, resulting in the following equation:

$$DSC(x_i) = \frac{2p_{i1}y_{i1} + \gamma}{p_{i1} + y_{i1} + \gamma} \tag{4}$$

Changing the denominator to the square form for faster convergence gives the dice loss

$$\mathrm{DL} = \frac{1}{N}\sum_{i}[1 - \frac{2p_{i1}y_{i1} + \gamma}{p_{i1}^2 + y_{i1}^2 + \gamma}] \tag{5}$$

or

$$\mathrm{DL} = 1 - \frac{2\sum_i p_i y_i + \gamma}{\sum_i {p_i}^2 + \sum_i {y_i}^2 + \gamma} \tag{6}$$

The DSC gets its maximum value of 1 when two sets, A and B, perfectly overlap. If the two sets do not intersect in any way, DSC starts to fall and eventually reaches zero. As a result, the DSC's range is 0 to 1, with bigger being better. From this, we can utilize 1-DSC as the dice loss to optimize overlapping between two sets.

Although the dice loss views false positives and false negatives as equally important, using the dice loss alone has been shown to be insufficient as it cannot address the prevailing influence of easy-negative examples on the training [67]. Xiaoya et al. [67] demonstrate that while easy negative examples can be easily pushed to a probability of 0, the model fails to

differentiate between positive and hard-negative examples. The authors suggested a weight-adjustment technique that assigns each training example a weight proportional to $(1 - p)$, that changes as training continues and makes the model sensitive to hard negative cases.

This makes Eq. 5 be

$$DSC(x_i) = \frac{2(1 - p_{i1})^{\alpha} p_{i1} y_{i1} + \gamma}{(1 - p_{i1})^{\alpha} p_{i1} + y_{i1} + \gamma} \tag{7}$$

where $(1 - p_{i1})^{\alpha}$ represents the weight assigned to each case, which shifts as training goes on, pushing the weight of easy examples.

Word-level information based language models, particularly those trained on specific-domain corpora like English, are prone to miss important information from VA text because of the nature of its nature [58]. We assert that this is also true for clinical and biomedical dataset text dictionaries and word distributions, and that SOTA text representation algorithms trained on these datasets will not perform as well on the VA data. We present Multi-Step transfer learning to mitigate this challenge by including character-level word representations from the ELMo language model and word-level embeddings from the BERT language model.

In the initial step of our framework, the language modeling pretraining objective primarily concerns the English language. This objective is achieved through unsupervised learning using the ELMo model trained on a combination of English Wikipedia and monolingual news crawl data. In the subsequent step, we leverage the VA embeddings acquired from the English domain to initialize the learning of VA text representations in the biomedical domain. This initialization is done by using an additional embedding layer before the embedding layers of the BERT model trained on PubMed abstracts. The resulting embeddings, which combine knowledge from both the English and biomedical domains, are then utilized for the final task of classifying the COD due to uncontrolled hyperglycemia.

This framework hypothesizes that the ELMo language model, when used to learn VA embeddings in the English domain can reduce the distribution divergence between the English language and the language in the VA corpora. It also assumes that biomedical and clinical knowledge from data transcribed by medical professionals can be used to improve the learning of VA representations via the shared mappings of the BERT model trained in the biomedical domain. The empirical evaluation of the framework involved its implementation on three open-source text classification datasets of English, biomedical and VA domains.

This paper's contributions are the following:

1. A Multi-Step Transfer Learning approach that makes the most use of the domain adaptation processes to improve the COD classification task of VA text.
2. A VA text representation framework with proven transferability to other medical conditions in cardiovascular, pulmonary, gastroenterology, neurology, orthopedics and radiology categories, which are leading CODs globally.
3. An empirical evaluation of the Multi-Step Transfer Learning model on the publicly available VA dataset collected by the Population Health Metrics Research Consortium (PHMRC).

We believe that combining knowledge from the two representation learning approaches will result in VA narrative representations better suited for the target task of COD classification. This is significant because improved VA text representations will accurately convey information about uncontrolled hyperglycemia as a mortality factor, which, when identified and diagnosed in a timely manner, can prevent further complications of type-1 diabetes and death.

3 Methods

This section presents the experimental setup of the Multi-Step Transfer Learning framework applied to a VA dataset. There are three parts to the experiment; The first part focuses on the selection of the best hyper-parameters for the BERT model which formed the second step of the transfer learning framework. The best approach for dealing with class imbalances in text classification is studied in the second part and the third part focuses on the validation of the framework on publicly available datasets of Population Health Metrics Research Consortium (PHMRC) VA Corpus, IMDb movie reviews and a clinical dataset of medical transcriptions.

3.1 Algorithms

BERT [6], BERT Experts-PubMed [69], BERT Experts-Wikibooks [69], ELMo [50] and BioELMo [34].

3.2 Datasets

English Language Corpus
For ELMo, we used the English Wikipedia and the monolingual news crawl data from WMT 2008–2012.[3] For BERT we used the expert version pretrained on combined Wikipedia and BooksCorpus.

General Medical Corpus
A vocabulary drawn from a database of 15,000 clinical research articles from PubMed Central (PMC)[4] that cover a wide spectrum of medical areas was used as the medical domain corpus.

Agincourt Verbal Autopsy Corpus
The verbal autopsy (VA) dataset used in this study is from Agincourt, ethics clearance number:M110138. It is a population health and demographic surveillance system operating in rural South Africa and aids research on causes and impacts of social transitions and populations. The data consists of 8698 VA records collected from 1992 to 2015.

The data were examined for indicators of uncontrolled hyperglycemia by a doctor with paediatric training and experience managing type-1 diabetes in high-income, low-income, and middle-income countries as well as paediatric training. In 3708 cases, uncontrolled hyperglycemia symptoms were present; and 77 cases were identified as deaths due to uncontrolled hyperglycemia. The data includes answers to both open and closed-ended questions and free text describing circumstances leading up to a death. We utilized the free text for this study.

Population Health Metrics Research Consortium (PHMRC) Verbal Autopsy Corpus
We validated our framework on the VA dataset collected by the Population Health Metrics Research Consortium (PHMRC). This data was gathered to make it possible to create and test methods for measuring cause-specific mortality in areas where there is limited and inaccurate COD coding [70]. It comprises of 11,979 VA records covering three age groups; neonate, child, and adult. The VA gathered data on potential risk factors, demographics, and other relevant information. The data is compiled in Tanzania, India, the Philippines and Mexico. Of these cases, 7580 are adult cases, and only 6896 of these had the narrative text feature.

IMDb Movie Reviews

[3] WMT is a collection of datasets used in shared tasks of the Third Conference on Machine Translation.

[4] PubMed is an online medical publication repository and contains published medical research across a very wide spectrum of clinical subjects. https://www.ncbi.nlm.nih.gov/pmc/.

We also validated our framework on two classification datasets in the English and clinical domains. For the English dataset, we utilized the informal movie reviews from Internet Movie Database (IMDb) [71] dataset provided by Keras. This dataset comprises 50,000 reviews, equally split between 25,000 negative and 25,000 positive cases. We selected this dataset due to its use as a benchmark for the Paragraph Vector [72] on sentiment analysis and information retrieval tasks. We therefore used it evaluate the Multi-Step Transfer Learning framework for the text classification task of sentiment analysis. We chose this dataset as it is larger than the VA dataset and it boasts an even class distribution.

Medical Transcriptions

The dataset consists of 2324 transcribed medical transcription sample reports across 21 categories of medical conditions including cardiovascular, pulmonary, gastroenterology, neurology, orthopedics and radiology [73]. Although the medical transcriptions are almost similar to VA narratives and have an imbalanced category distribution, this dataset is fairly smaller than the VA dataset. The Multi-Step Transfer Learning framework was evaluated on the multi-class text classification using this dataset.

3.3 Experiments

Similar to the work by Manaka et al. [49], the initial step of the Multi-Step Transfer Learning framework involves an exploratory search for models across a number of domains to identify the one that best represents the VA corpus. Three sets of ELMo language models were trained in three different domains, with the objective of identifying the optimal text representations for the VA language modeling task. Two BERT Experts language models were initialized with random weights for training. The set of embeddings from the ELMo model with the lowest perplexity scores was then transferred to the BERT models.

3.3.1 ELMo

The tensorflow implementation of ELMo from github repository [5] was cloned to train and evaluate the ELMo embeddings in the English, medical and public health domains. Input data was prepared by randomly splitting the training data into many training files, each containing pre-tokenized and white space-separated text, one sentence per line. The three ELMo models were trained using the same hyperparameters as the original ELMo model: one on a vocabulary derived from the English Wikipedia and monolingual news crawl data from WMT 2008–2012, and one on vocabulary derived from the VA corpus. The third ELMo language model was trained on a vocabulary from 10 M PubMed abstracts.

The datasets were preprocessed by removing punctuation as well as lower casing the text. When creating a vocabulary using VA data, we did a comparison of when stop words were removed and when they were not and used a vocabulary that gave a less perplexity score. Following the paper implementation [74], the language models were trained via the multi-task learning of next word prediction and natural language inference. The trained language model embeddings were used as feature extractors to initialize random word vectors of VA language corpora when using the pretrained BERT models in the second transfer learning step. All ELMo layers were combined into a single vector in order to be used in this target task. The three language models were evaluated on the perplexity score, results of which are depicted in Table 2.

[5] https://github.com/allenai/bilm-tf.

3.3.2 BERT

We built a basic fine-tuned model, which involved creating a preprocessing model, utilizing a pretrained model from BERT Experts, implementing an embedding layer with ELMo embeddings for initializing the BERT embeddings, incorporating a fully connected layer, and adding a dropout layer for COD classification. To facilitate comparison, we employed two BERT models from TensorFlow Hub,[6] which were pretrained on medical and English domains, specifically Wikibooks and PubMed abstracts.

In addition to masked language modeling, the other training object of BERT is next sentence prediction, which can be thought of as a form of sentence modeling and it incorporates both tasks via multitask learning. A map with three key values was created by the BERT models: pooled output which represented each input sequence as a whole, i.e. the embedding for all VA data, sequence output which represented each input token in context, i.e. the contextual embedding for every token in the VA corpus, and encoder outputs which represented intermediate activations in the transformer blocks. For extracted BERT embeddings we used the 768-element pooled output array.

Devlin et al. [6] observed that BERT models were sensitive to the choice of hyperparameters for smaller datasets compared to larger ones. The authors recommended the following hyperparameters for fine-tuning the model: a batch size of 16 or 32, learning rates (Adam) of: $5e^{-5}$, $3e^{-5}$ and $2e^{-5}$ when the number of training epochs ranges between 2 and 4. We conducted experiments using various combinations of these hyperparameters to identify those giving the optimal performance. With these combinations, we also compared the model performances using both the dice loss and the weighted binary cross-entropy loss functions. The corresponding results are given in Table 3.

We conducted experiments using various combinations of these hyperparameters to identify the configurations yielding optimal performance. Within these configurations, we also compared model performance using both the dice loss and the weighted binary cross-entropy loss functions. The corresponding results are presented in Table 3.

The overall framework of the Multi-Step Transfer Learning framework is the following:

1. *Pretraining: ELMo Pretraining for Spelling Variations and Language*: The ELMo language model pretrained in the English domain (English Wikipedia and the monolingual news crawl data) was used in the initial step of Multi-Step Transfer Learning. The assumption is that ELMo will learn the generic linguistic features of a VA text report like syntax, semantics and context. By leveraging character-level information, ELMo embeddings can potentially help in handling spelling errors, variations in language, and even rare or out-of-vocabulary words in VA texts. The domain adaptation technique from this task to the one in the next step is feature extraction, as the source task (language modeling) and the target task (language modeling) are similar. This learning paradigm is classified as cross-domain learning because it entails learning embeddings in one domain and transferring them to another.
2. *Intermediate Training: BERT Pretraining for Medical Information*: ELMo embeddings from the first step were used to initialize the embedding layers of the BERT model trained in the biomedical domain (PubMed abstracts). This intermediate training step aimed to find the optimal embeddings for the final COD classification task. This initialization helps BERT start from a more informative point enabling it to achieve better performance.

[6] https://www.tensorflow.org/hub.

BERT's biomedical domain embeddings will help capture medical terminology, concepts, and domain-specific information present in VA texts. This is crucial for accurately extracting medical information and understanding the context of symptoms and COD.

3. *Fine-tuning*: After initializing the embedding layers, the entire BERT model was fine-tuned on the target COD classification task using the labeled VA data. The assumption is that the BERT language models will deeply understand and learn the relationships between words and contexts, resulting in a highly contextualized embeddings. This setting where the source task and target task differ is called cross-task learning and the domain adaptation technique of fine-tuning was used in this step.
4. *Text Classification*: To ensure comparability, all models were trained and evaluated using the same split. For the task of sentence classification of COD due to uncontrolled hyperglycemia, a fully connected layer was added above the BERT self-attention layers.

For the classification task, the cost-sensitive classification [65] technique was used to enhance the weight of mislabeling a VA case by altering the fully connected layer's cost function during training by multiplying each example's loss by a factor. The computed class weights ratio is 0.50494305 : 51.07608696. To ensure comparability, all models were trained and evaluated using the same split and for the sentence classification task, a fully connected layer was added above the BERT self-attention layers.

In addition to Multi-Step Transfer Learning, which uses ELMo embeddings to initialize the learning of BERT embeddings, we also experimented with the concatenation of ELMo and BERT embeddings. We used the resultant embeddings for the classification of COD due to uncontrolled hyperglycaemia from VA reports with a feed-forward neural network. Our comparison included evaluating these embeddings as features alongside the binary features extracted from a VA report and when both binary and VA text features were used in combination. This would ultimately provide insights into the effectiveness of VA narrative embeddings in the classification task of COD due to uncontrolled hyperglycaemia from VA reports (Fig. 1).

3.3.3 Validation

To assess the framework, we conducted tests comparing BERT's performance when employing both feature extraction and fine-tuning domain adaptation methods. The testing involved evaluating the BERT model embeddings in isolation and when combined with ELMo embeddings. In the case of using BERT and ELMo embeddings together, the first scenario, Multi-Step Transfer Learning utilized fine-tuning adaptation, while the second scenario involved feature extraction, where ELMo embeddings were concatenated with BERT embeddings.

While only accuracy was employed for multi-class classification, the performance of the framework for sentiment analysis, similar to binary classification was assessed using recall, precision, F1-score, the area under the ROC curve (AUC–ROC), and accuracy. For multi-class classification, we also studied the effect of reducing the dimension of the text by extracting clinical domain entities. We used the ScispaCy[7] package to detect medical entities in the medical transcriptions.

[7] https://allenai.github.io/scispacy/.

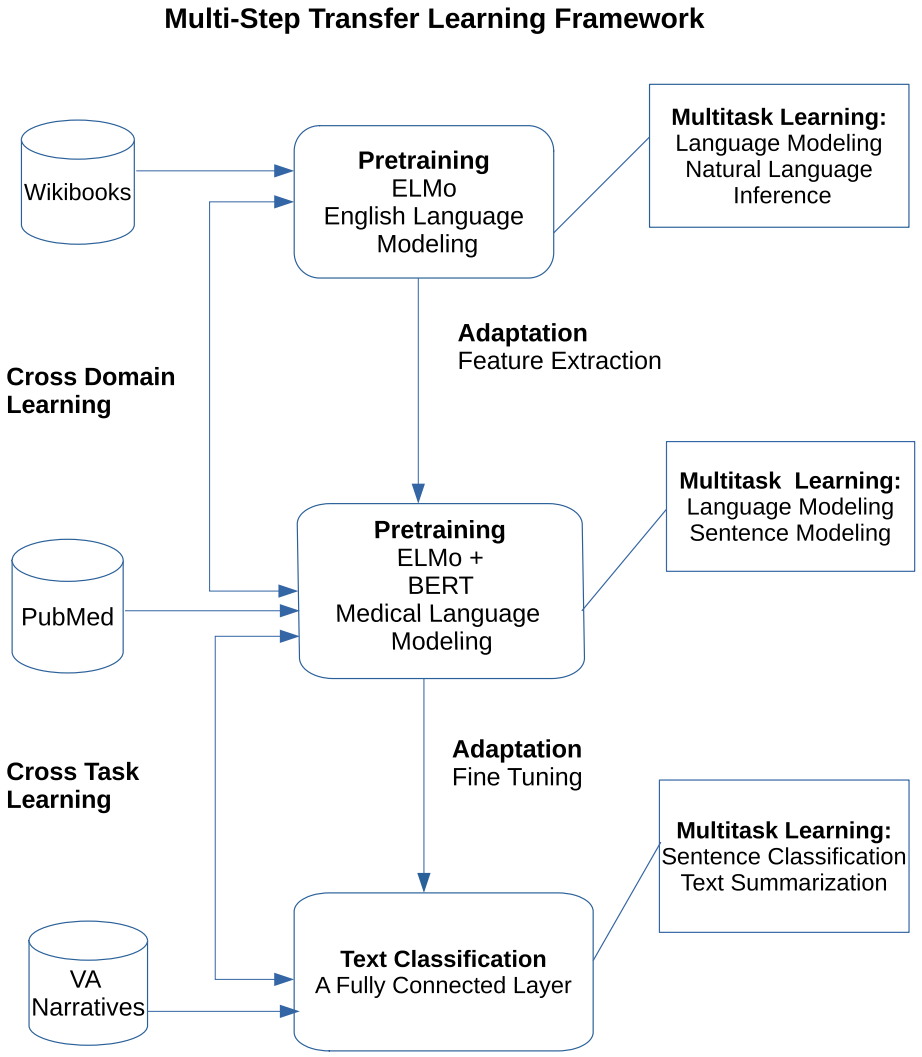

Fig. 1 Multi-step transfer learning framework

4 Results and Discussion

The ELMo model trained on Wikibooks and Book Corpus vocabulary exhibits better perplexity scores on the training and testing datasets than the ELMo models pretrained on PubMed abstracts and the VA vocabulary (Table 2). This is due to the sizes of the vocabulary sets from these datasets. This also means that the distribution divergence gap of features between the VA corpus and the English domain is less than that between the VA and the health domain. We believe that ELMo was able to extract a lot of linguistic knowledge, including spelling, syntax, and grammar in the English domain. Generally, the dice loss function has better results across all the metrics than the weighted binary cross entropy loss function (Table 3). These findings are consistent with the results of a study by Xiaoya et al. [67], which showed

Table 2 Evaluation of ELMo language models

Technique	Vocabulary	Tokens	Train perplexity	Test perplexity
ELMo	English Wikipedia	5.5B	43.23	31.32
ELMo	Agincourt Verbal Autopsy	982 495	71.55	50.01
BioELMo	PubMed Abstracts	2.46B	47.44	33.01
ELMo	PHMRC Verbal Autopsy	475 005	67.22	51.42
ELMo	IMDb Reviews	11.7M	52.47	44.56
BioELMo	Medical transcriptions	909 830	72.12	52.32

Table 3 BERT Fine-Tuning Hyperparameter Search

Epochs	L-Rate	Recall	Precision	F1-Score	AUC-ROC
Dice Loss					
2	$2e^{-5}$	0.7843	0.7312	0.7568	0.8514
	$3e^{-5}$	**0.8000**	0.7656	0.7824	0.8433
	$5e^{-5}$	0.7254	0.7789	0.7512	0.8087
3	$2e^{-5}$	**0.8000**	0.7811	0.7904	0.8293
	$3e^{-5}$	0.7541	**0.8475**	**0.7981**	**0.8881**
	$5e^{-5}$	0.7461	0.7255	0.7356	0.8591
4	$2e^{-5}$	0.7111	0.7000	0.7056	0.8532
	$3e^{-5}$	0.7661	0.6286	0.6906	0.8497
	$5e^{-5}$	0.7100	0.6375	0.6718	0.8245
Weighted Cross-Entropy					
2	$2e^{-5}$	0.5822	0.5900	0.5861	0.6472
	$3e^{-5}$	0.5344	0.6415	0.5830	0.6166
	$5e^{-5}$	0.6300	0.5574	0.5914	0.6881
3	$2e^{-5}$	0.6500	0.6100	0.6294	0.7105
	$3e^{-5}$	0.5333	0.4621	0.4952	0.5629
	$5e^{-5}$	0.4700	0.5248	0.4959	0.5568
4	$2e^{-5}$	0.6344	0.6000	0.6167	0.6562
	$3e^{-5}$	0.5866	0.5223	0.5525	0.6178
	$5e^{-5}$	0.5167	0.5469	0.5314	0.6381

that when the imbalance between classes is extreme, the weighted binary cross entropy loss is unable to alleviate the imbalance in datasets. The authors show how the impact of easy negative examples causes this as the only thing that weighting the classes does is balance the labels such that the training and test times are equal.

Madabushi et al. [65] showed that although BERT can handle imbalanced datasets without the requirement for further data augmentation, evaluation findings of the weighted cross entropy loss function demonstrate that it fails to generalize when the train and test sets differ. Our results show that this is the case with VA reports as well which consist of differing narrations.

Table 4 Multi-step transfer learning on sets of BERT and ELMo Embeddings of the Agincourt VA Dataset

BERT	ELMo	Recall	Precision	F1-Score	AUC-ROC
Wiki Books	None	0.6141	0.5594	0.5855	0.7001
	Verbal Autopsy	0.6581	0.6147	0.6357	0.7101
	Wikipedia	0.7687	0.7787	0.7734	0.8507
	PubMed Abstracts	0.7581	0.6991	0.7324	0.8111
PubMed Abstracts	None	0.7141	0.6011	0.6528	0.7399
	Verbal Autopsy	0.8065	0.6011	0.6888	0.7981
	Wikipedia	**0.8171**	**0.8644**	**0.8401**	**0.9144**
	PubMed Abstracts	0.7496	0.6987	0.7233	0.8149

Table 5 Multi-step transfer learning on sets of BERT and ELMo Embeddings of the PHMRC VA Dataset

BERT	ELMo	Recall	Precision	F1-Score	AUC-ROC
Wiki Books	None	0.7044	0.6741	0.6889	0.7447
	Verbal Autopsy	0.6743	0.5561	0.6095	0.7253
	Wikipedia	0.7848	0.7548	0.7695	0.8465
	PubMed Abstracts	0.7341	0.6791	0.7056	0.8112
PubMed Abstracts	None	0.6946	0.6681	0.6811	0.7422
	Verbal Autopsy	0.7215	0.7046	0.7129	0.7956
	Wikipedia	**0.8363**	**0.7941**	**0.8147**	**0.9017**
	PubMed Abstracts	0.7561	0.7148	0.7349	0.8764

Table 4 gives results of a comparison of sets of VA ELMo embeddings trained on different domain vocabularies. BERT pretrained on Wikibooks and BERT pretrained on PubMed abstracts perform best with VA corpus embeddings pretrained on English Wikipedia in all settings. This can be explained by the fact that PubMed abstracts and the VA corpus contain numerous mentions of words in English Wikipedia and Books Corpus and that both domains form subsets of the English domain.

The benefits of adding ELMo embeddings are evident from the performance scores that are greater than for a BERT model without ELMo embeddings, except for the setting where ELMo embeddings are trained on VA corpus vocabulary, where the model performs worse than the setting without ELMo embeddings. We believe this to be due to the VA corpus vocabulary being smaller in size, containing misspelled words and grammatical errors. We argue that this limits linguistic knowledge and that the total capacity of the ELMo model was not leveraged because the model is data-intensive.

Other authors have reported higher F1-scores (95.57%) when using a combination of off-the-shelf ELMo and BERT embeddings as an initial step to a combination of a bidirectional LSTM and conditional random fields (BiLSTM-CRF) module on CoNLL-2003 and OntoNotes 5.0 datasets for English named entity recognition (NER) task [59]. Nonetheless, our results are comparable with those of Boukkouri et al. [51] who reported F1-scores ranging from 70 to 89% on CharacterBERT, a variant of BERT that drops the word piece system altogether in favor of a character-CNN on a series of NER tasks on the medical corpus. However, with F1-scores around 0.8633–0.8907, we are convinced character information can improve the classification of cause of death (COD) from VA reports.

Table 6 Multi-step transfer learning framework on IMDb reviews for BERT pretrained on English and medical domains

BERT	Adaptation	F1-Score	AUC-ROC	Accurary
Wiki Books	Fine tuning	**0.9007**	**0.9576**	**0.8895**
	Feature extraction	0.8356	0.8181	0.8518
PubMed Abstracts	Fine tuning	0.8382	0.9571	0.8862
	Feature extraction	0.6695	0.7005	0.7062

Table 7 Multi-step transfer learning on medical transcriptions detected and undetected for clinical entities accuracy scores

Model	Adaptation	Detected	Undetected
BERT	Fine tuning	0.6872	0.7629
	Feature extraction	0.5841	0.6577
Multi-Step (BERT)	Fine tuning	0.7148	0.8244
	Feature extraction	0.6824	0.7479
BERT-PubMed	Fine tuning	0.7525	0.8336
	Feature extraction	0.5844	0.5952
Multi-Step (BERT-PubMed)	Fine tuning	**0.8421**	**0.8946**
	Feature extraction	0.7669	0.7769

Danso et al. [75] showed that there is no corpus similar to VA and that the only dataset available to provide a gold standard for the evaluation of computational approaches to VA analysis, the PHMRC corpus was not suitable for linguistic research. Evaluation of our approach on the PHMRC VA dataset gave similar results, (shown in Table 5) to those of the Agincourt VA corpus in terms of performance across the different embeddings combinations. The authors show that the preprocessing steps involved in its annotation like the removal of syntax rules and linguistic information, removal of words that do not occur frequently, and only taking into account medically relevant concept-terms have resulted in the loss of important information. We found these to not have an affected our approach as some of these steps were used in preprocessing the Agincourt VA data for the ELMo models.

Evaluation of our framework on the IMDb movie reviews dataset (English domain) and the medical transcriptions corpus (medical domain), where the former is larger than the VA corpus and the latter is smaller, shows that the approach is capable of handling datasets of different sizes from multiple domains (Tables 6 and 7). The Multi-Step transfer learning framework can also generalize to both balanced and unbalanced datasets as the movie reviews are equally balanced while the medical transcriptions dataset is not.

Works by Beltagy et al. [31], See et al. [76], Jin et al. [34] and Boukkouri et al. [51] have shown that the general English-domain word-piece vocabularies are not suitable for specialized domain applications like clinical and biomedical domains. Evaluation of the Multi-Step transfer learning framework results are in agreement with this (Tables 6 and 7), where BERT trained in the English domain gets higher F1-scores than BERT pretrained on PubMed abstracts on the IMDb reviews. The PubMed pre-trained BERT model also outperforms the general English domain pre-trained BERT model on the medical transcriptions. This is because even though BERT can achieve the right balance between the flexibility

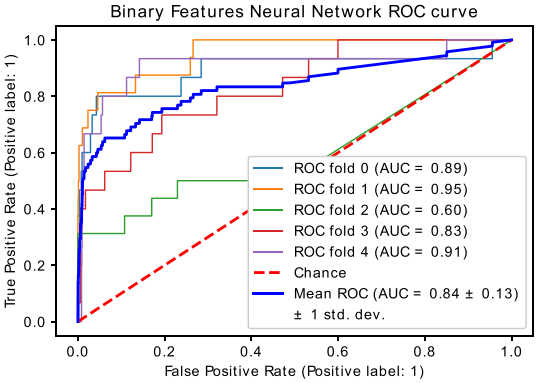

Fig. 2 Receiver Operating Characteristic (ROC) curves and the Area Under the ROC Curve (AUC-ROC) generated by the Neural Network Classifier using Binary Features from a VA Report

of characters and the utility of full words, employing predefined word piece vocabularies from the general domain is not always appropriate, especially when building specialized domain models. Further evaluation shows that generally, the fine-tuning adaptation technique achieves better results than the feature extraction adaption across all settings (Tables 6 and 7). These results are in line with Peters et al. [33] in the English domain and Jin et al. [34] in the biomedical domain where both works compared the two transfer learning adaptation techniques.

In comparison to our framework's sensitivity (recall) of 0.8171 and 0.8363 on the Agincourt and PHMRC datasets for adult deaths, Jeblee et al. [77] reported a mean sensitivity of 0.7700 and in another work where they combined word2vec embeddings and key phrases they achieved a recall score of 0.7780 [78]. These works however used word frequency counts as features and both methods don't take word order and context into account. On more recent works that incorporate context and character information, Yan et al. [58] achieved recall scores of 0.6990 while Manaka et al. [79] gave 0.6000. Considering the improvement added by character information on improving COD classification, Manaka et al. [49] added features from multiple domains and reported a score of 0.8755. These findings suggest that our proposed transfer learning methodology can adapt to VA datasets across various demographics as the works compared against used VA datasets collected in other developing countries, including Ghana, India and Tanzania.

Figures 2, 3, and 4 illustrate the receiver operating characteristic (ROC) curves for the neural network classifier across three different sets of VA features. These features are a concatenation of VA embeddings learned with character-level information using ELMo in the English domain, and those learned with word-level information using BERT in the biomedical domain. Both sets of embeddings were extracted through the feature extraction domain adaptation techniques. In all three scenarios, the ROC curves demonstrate an upward rise towards the upper-left corner, signifying the accurate prediction of positive and negative cases. Notably, when comparing the individual text and binary features with the combined text and binary features setting, the latter exhibits the highest AUC-ROC score (93%). This highlights the significance of text features in the classification of COD by uncontrolled hyperglycemia.

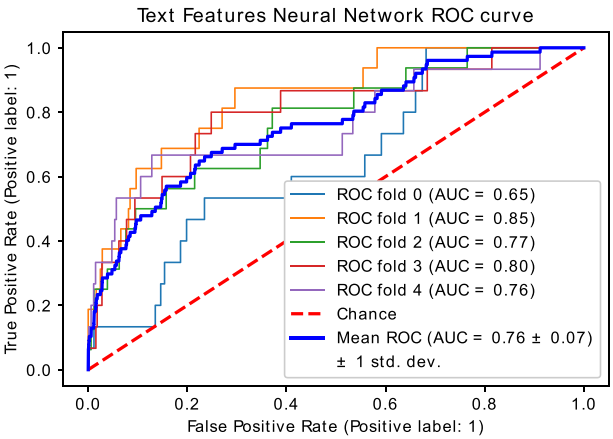

Fig. 3 Receiver Operating Characteristic (ROC) curves and the Area Under the ROC Curve (AUC-ROC) for the Neural Network classifier applied to Text Features from a VA Report

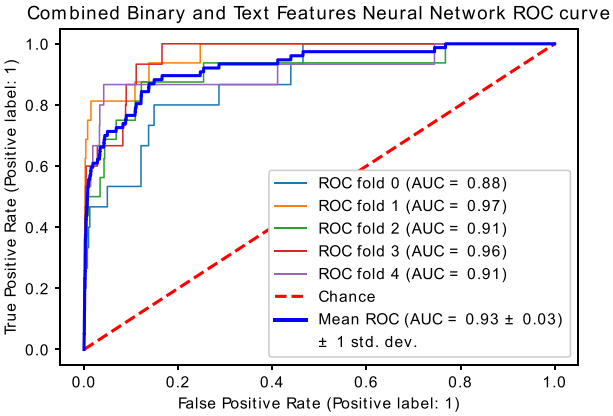

Fig. 4 Receiver Operating Characteristic (ROC) curves and the Area Under the ROC Curve (AUC-ROC) for the Neural Network classifier using both Binary and Text Features from a VA Report

5 Limitations of the Study

This research is restricted to the text classification task. More research can be conducted on named entity recognition (NER) and relation extraction tasks, both critical in NLP applications in the health domain. Furthermore, because the classification framework significantly impacts results, additional experimentation could reveal whether or not a similar behaviour occurs for other subsets of the English domain. It would also be interesting to look into the same architecture using different character and word embedding models.

Due to limited computational resources, the data were split into smaller batches, and embeddings had to be computed one batch at a time. This may have impacted the computation of the embeddings because the experiments were carried out on Google Colaboratory, which allocated different GPUs for each batch run.

6 Conclusion

We have demonstrated through experimentation that the Multi-Step transfer learning framework can enhance the representations of text from VA reports. Consequently, it leads to an improved COD classification due to uncontrolled hyperglycemia derived from VA reports. As part of our future work, we intend to explore additional NLP techniques that incorporate Transformer and CNN architectures. This incorporation within a similar hybrid framework will allow us to assess their performance. Furthermore, we plan to investigate the impact of VA narrative embeddings from this framework when combined with the binary features of VA reports. Exploring the potential application of this framework to other CODs mentioned in the VA reports is also of interest. We will additionally explore using the ChatGPT language model for COD classification from VA reports.

Acknowledgements We thank the MRC/Wits-Agincourt Unit for providing us with the dataset and for assistance in understanding its history. We are thankful to the United Nations' Organization of Women in Science for the Developing World (OWSD) for the support granted to carry out this study.

Author Contributions The authors contributed equally to this work. We confirm that all named authors have read, reviewed, and approved the manuscript and that no other individuals who meet the requirements for authorship but are not listed have contributed to the work. We also confirm that we all approved of the order in which the authors are listed in the manuscript.

Funding Open access funding provided by University of the Witwatersrand. The Organization for Women in Science for the Developing World (OWSD) funded this research.

Availability of Data and Materials Due to patient privacy and confidentiality policies, the Agincourt dataset analysed during this study is not publicly available. Still, it may be obtained with Data Use Agreements with the MRC/Wits-Agincourt Unit. Researchers interested in access to the data may contact Dalby Dawn at Dawn.Dalby@wits.ac.za.The PHMRC verbal autopsy dataset used to validate this study is publicly available at https://osf.io/xuk5q/, the IMDB dataset also used in the validation of the study is publicly available at https://datasets.imdbws.com/ while the medical transcriptions dataset is publicly available at https://www.mtsamples.com/. The three data sets are available under the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Code Availability All code for data cleaning and analysis associated with the current submission is available upon request from the corresponding author at [email address masked for blind review].

Declarations

Conflict of interest We wish to reaffirm that no known conflicts of interest related to this publication or substantial financial support might have impacted the research's findings.

Ethics Approval We further confirm that any aspect of the work covered in this manuscript that has involved either experimental animals or human patients has been conducted with the ethical approval of all relevant bodies and that such approvals are acknowledged within the manuscript (ethics clearance number: M110138).

Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

1. United Nations (2013) Department of economic and social affairs, population division, united nations. World Population Prospects: The 2012 revision
2. World Health Organisation (2007) Verbal autopsy standards: ascertaining and attributing cause of death, Geneva. Switzerland, World Health Organisation
3. Hirschman L, Chapman WW, D'Avolio LW, Savova GK, Uzuner O (2011) Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 18(5):450–453
4. Ohno-Machado L, Nadkarni P, Chapman W (2011) Natural language processing: an introduction. J Am Med Inform Assoc 18:544–51
5. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
6. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
7. Kooverjee N, James S, Van Zyl T (2022) Investigating transfer learning in graph neural networks. Electronics 11(8):1202
8. Bhana N, van Zyl TL (2022) Knowledge graph fusion for language model fine-tuning. In: 2022 9th international conference on soft computing and machine intelligence (ISCMI)
9. Kim Y (2014) Convolutional neural networks for sentence classification. In: Proceedings of the conference on empirical methods in natural language processing, pp 1746–1751
10. Ramachandran P, Liu PJ, Le QV (2016) Unsupervised pretraining for sequence to sequence learning. arXiv:1611.02683
11. Delrue, L., Gosselin, R., Ilsen, B., Landeghem, A.V., de Mey, J., Duyck, P.: Difficulties in the interpretation of chest radiography. Comparative Interpretation of CT and Standard Radiography of the Chest, 27–49 (2011)
12. Goergen SK, Pool FJ, Turner TJ, Grimm JE, Appleyard MN, Crock C, Fahey MC, Fay MF, Ferris NJ, Liew SM, Perry RD, Revell A, Russell GM, Wang SC, Wriedt C (2013) Evidence-based guideline for the written radiology report: methods, recommendations and implementation challenges. J Med Imaging Radiat Oncol 57(1):1–7
13. Brady A, Laoide R, Mccarthy P, Mcdermott R (2012) Discrepancy and error in radiology: concepts, causes and consequences. Ulster Med J 81:3–9
14. Liu F, You C, Wu X, Ge S, Sun X (2021) Auto-encoding knowledge graph for unsupervised medical report generation. CoRR abs/2111.04318
15. Liu F, Yang B, You C, Wu X, Ge S, Liu Z, Sun X, Yang Y, Clifton D (2022) Retrieve, reason, and refine: generating accurate and faithful patient instructions. NeurIPS 35:18864–18877
16. Li J, Wang X, Wu X, Zhang Z, Xu X, Fu J, Tiwari P, Wan X, Wang B (2023) Huatuo-26m, a large-scale chinese medical qa dataset. CoRR abs/2305.01526
17. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J (2020) Measuring massive multitask language understanding. CoRR abs/2009.03300
18. Abacha AB, Shivade C, Demner-Fushman D (2019) Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp 370–379
19. Zhou P, Wang Z, Chong D, Guo Z, Hua Y, Su Z, Teng Z, Wu J, Yang J (2022) Mets-cov: A dataset of medical entity and targeted sentiment on covid-19 related tweets. NeurIPS 35:21916–21932
20. Nori H, King N, McKinney SM, Carignan D, Horvitz E (2023) Capabilities of gpt-4 on medical challenge problems. CoRR abs/2303.13375
21. Fang C, Ling J, Zhou J, Wang Y, Liu X, Jiang Y, Wu Y, Chen Y, Zhu Z, Ma J, Yan Z (2023) How does chatgpt4 preform on non-english national medical licensing examination? an evaluation in chinese language. medRxiv 35
22. Zeng Q, Garay L, Zhou P, Chong D, Hua Y, Wu J, Pan Y, Zhou H, Voigt R, Yang J (2022) Greenplm: Cross-lingual transfer of monolingual pre-trained language models at almost no cost. The 32nd International Joint Conference on Artificial Intelligence
23. Liu J, Zhou P, Hua Y, Chong D, Tian Z, Liu A, Wang H, You C, Guo Z, Zhu L, Li M (2023) Benchmarking large language models on cmexam - a comprehensive chinese medical exam dataset. CoRR abs/2306.03030

24. Liu F, Zhu T, Wu X, Yang B, You C, Wang C, Lu L, Liu Z, Zheng Y, Sun X, Yang Y, Clifton L, Clifton DA (2023) A medical multimodal large language model for future pandemics. npj Digit. Med 6:226
25. Baxter J (2000) A model of inductive bias learning. J Artific Intell Res 12:149–198
26. Huang Z, Zweig G, Dmoulin B (2014) Cache based recurrent neural network language model inference for first pass speech recognition. IEEE ICASSP, pp 6354–6358
27. Wen Z, Lu X, Reddy S (2020) Medal: Medical abbreviation disambiguation dataset for natural language understanding pretraining. Proceedings of the 3rd clinical natural language processing workshop, pp 130–135
28. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott MBA (2019) Publicly available clinical bert embeddings. arXiv:1904.03323
29. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240
30. Qiao J, Bhuwan D, William C, Xinghua L (2019) Probing biomedical embeddings from language models. In: Proceedings of the 3rd workshop on evaluating vector space representations for NLP, pp 82–89
31. Beltagy I, Cohan A, Lo K (2019) Scibert: pretrained contextualized embeddings for scientific text. arXiv:1903.10676
32. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG (2016) Mimic-III, a freely accessible critical care database. Sci Data3
33. Peters M, Ruder S, Smith N (2019) To tune or not to tune? adapting pretrained representations to diverse tasks. arXiv:1903.05987
34. Jin Q, Dhingra B, Cohen W, Lu X (2019) Probing biomedical embeddings from language models. arXiv:1904.02181
35. Zhao S, Li B, Reed C, Xu P, Keutzer K (2020) Multi-source domain adaptation in the deep learning era: a systematic survey. arXiv:2002.12169
36. Torralba A, Efros AA (2011) Unbiased look at dataset bias. In CVPR
37. Zhao S, Zhao X, Ding G, Keutzer K (2018) Emotiongan: Un-supervised domain adaptation for learning discrete probability distributions of image emotions. In ACM MM
38. III HD (2007) Frustratingly easy domain adaptation. Association for Computational Linguistic (ACL), pp 256–263
39. Sun S, Shi H, Wu Y (2015) A survey of multi-source domain adaptation. Inf Fusion 24:84–92
40. Riemer M, Cases I, Ajemian R, Liu M, Rish I, Tu Y, Tesauro G (2019) Learning to learn without forgetting by maximizing transfer and minimizing interference. In ICLR
41. Sun Q, Chattopadhyay R, Panchanathan S, Ye J (2011) A two-stage weighting framework for multi-source domain adaptation. Adv Neural Inform Process Syst 24:505–513
42. Schweikert G, Rätsch G, Widmer C, Schölkopf B (2009) An empirical analysis of domain adaptation algorithms for genomic sequence analysis. Adv Neural Inform Process Syst 21:1433–1440
43. Guo H, Pasunuru R, Bansal M (2020) Multi-source domain adaptation for text classification via distancenet-bandits. In AAAI
44. Zhao S, Li B, Yue X, Gu Y, Xu P, Hu R, Chai H, Keutzer K (2019) Multi-source domain adaptation for semantic segmentation. NeurIPS
45. Li X, Lv S, Li M, Jiang Y, Qin Y, Luo H, Yin S (2023) SDMT: spatial dependence multi-task transformer network for 3d knee MRI segmentation and landmark localization. IEEE Trans Med Imaging 42(8):2274–2285. https://doi.org/10.1109/TMI.2023.3247543
46. Li X, Jiang Y, Li M, Yin S (2020) Lightweight attention convolutional neural network for retinal vessel image segmentation. IEEE Trans Ind Inf 17(3):1958–1967
47. Hu K, Wu W, Li W, Simic M, Zomaya A, Wang Z (2022) Adversarial evolving neural network for longitudinal knee osteoarthritis prediction. IEEE Trans Med Imaging 41(11):3207–3217
48. Wan Y, Jiang Z (2023) Transcrispr: transformer based hybrid model for predicting CRISPR/cas9 single guide RNA cleavage efficiency. IEEE Trans Med Imaging 20(2):1518–1528
49. Manaka T, Van Zyl TL, Kar D (2022) Improving cause-of-death classification from verbal autopsy reports. arXiv:2210.17161
50. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. NAACL
51. Boukkouri HE, Ferret O, Lavergne T, Noji H, Zweigenbaum P, Tsujii J (2020) Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters
52. Vaswani A, Shazeer N, Parmar N, Uszkoreita J, Jones L, Gomez AN (2017) Attention is all you need. NIPS, pp 6000–6010
53. He K, Zhang X, Ren S, Jian S (2016) Deep residual learning for image recognition. AI Open 3:770–778. https://doi.org/10.1109/CVPR.2016.90
54. Ba LJ, Kiros JR, Hinton GE (2016) Layer normalization. arXiv:1607.06450

55. Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. Adv Neural Inf Process Syst, pp 649–657
56. Verwimp L, Pelemans J, hamme HV, Wambacq P (2017) Character-word lstm language models. Proceedings of the 15th conference of the European chapter of the association for computational linguistics vol 1, pp 417–427
57. Si Y, Roberts K (2018) A frame-based nlp system for cancer-related information extraction. AMIA Ann Symp Proc, pp 1524–1533
58. Yan Z, Jeblee S, Hirst G (2019) Can character embeddings improve cause-of-death classification for verbal autopsy narratives? BioNLP@ACL
59. Affi M, Latiri C (2021) Be-blc: Bert-elmo-based deep neural network architecture for English named entity recognition task. Proc Comput Sci 192
60. Lin T, Wang Y, Liu X, Qiu X (2022) A survey of transformers. AI Open 3:111–132. https://doi.org/10.1016/j.aiopen.2022.10.001
61. Guo M, Zhang Y, Liu T (2019) Gaussian transformer: a lightweight approach for natural language inference. In: Proceedings of AAAI, pp 6489–6496. https://doi.org/10.1609/aaai.v33i01.33016489.
62. Yang B, Tu Z, Wong DF, Meng F, Chao LS, Zhang T (2018) Modeling localness for self-attention networks. In: Proceedings of EMNLP. Brussels, Belgium, pp 4449–4458. https://doi.org/10.1109/CVPR.2016.90.
63. Wang W, Li X, Ren H, Gao D, Fang A (2023) Chinese clinical named entity recognition from electronic medical records based on multisemantic features by using robustly optimized bidirectional encoder representation from transformers pretraining approach whole word masking and convolutional neural networks: model development and validation. JMIR Med Inform 11(e44597)
64. Kong J, Zhang L, Jiang M, Liu T (2021) Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition. J Biomed Inform 116:103737. https://doi.org/10.1016/j.jbi.2021.103737
65. Madabushi HT, Kochkina E, Castelle M (2020) Cost-sensitive BERT for generalisable sentence classification with imbalanced data. arXiv:2003.11563
66. Wei JW, Zou K (2019) Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv:arXiv:1901.11196
67. Xiaoya L, Xiaofei S, Yuxian M, Junjun L, Fei W, Jiwei L (2020) Dice loss for data-imbalanced NLP tasks. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 465–476
68. Sorensen TA (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Kong Dan Vidensk Selsk Biol Skr 5:1–34
69. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado G.S, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jozefowicz R, Jia Y, Kaiser L, Kudlur M, Levenberg J, Mané D, Schuster M, Monga R, Moore S, Murray D, Olah C, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. https://www.tensorflow.org/
70. Flaxman AD, Harman L, Joseph J, Brown J, Murray CJ (2018) A de-identified database of 11,979 verbal autopsy open-ended responses. Gates Open Res 2:18
71. Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics
72. Le QV, Mikolov T (2014) Distributed representations of sentences and documents. In Proceedings of the 31st international conference on machine learning (ICML 2014), pp 1188–1196
73. Mtsamples (2022) Transcribed medical transcription sample reports and examples. Great collection of transcription samples. https://www.mtsamples.com/
74. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. arXiv:1802.05365
75. Danso S, Johnson O, Ten Asbroek A, Soromekun S, Edmond K, Hurt C, Hurt L, Zandoh C, Tawiah C, Fenty J, Etego SA, Aygei SO, Kirkwood B (2013) A semantically annotated verbal autopsy corpus for automatic analysis of cause of death. ICAME J Int Comput Arch Modern Mediev English 37:37–69
76. See A, Liu PJ, Manning CD (2017) Get to the point: Summarization with pointer-generator networks. arXiv:1704.04368
77. Jeblee S, Gomes M, Jha P, Rudzicz F, Hirst G (2019) Automatically determining cause of death from verbal autopsy narratives. BMC Med Inf Decis Mak 19(127)
78. Jeblee S, Gomes M, Hirst G (2018) Multi-task learning for interpretable cause of death classification using key phrase predictions. In Proceedings of the BioNLP 2018 Workshop vol 34, no 19, pp 12–27

79. Manaka T, Van Zyl TL, Wade AN, Kar D (2022) Using machine learning to fuse verbal autopsy narratives and binary features in the analysis of deaths from hyperglycaemia. arXiv:2204.12169