Konputazio Ingeniaritza eta
Sistema Adimentsuak Unibertsitate Masterra

Máster Universitario en Ingeniería Computacional
y Sistemas Inteligentes

Konputazio Zientziak eta Adimen Artifiziala Saila

Departamento de Ciencias de la Computación e Inteligencia Artificial

Master Tesia
Tesis de Máster
Master's Thesis

Extracción de información de las autopsias verbales / Informazio erauzketa ahozko autopsietatik / Information extraction from verbal autopsies

Ander Cejudo Taramona

Zuzendaritza
Dirección

Alicia Pérez Ramírez

IXA taldea, Euskal Herriko Unibertsitatea (EHU)

Grupo IXA, Universidad del País Vasco (UPV)

Arantza Casillas Rubio

IXA taldea, Euskal Herriko Unibertsitatea (EHU)

Grupo IXA, Universidad del País Vasco (UPV)

Universidad del País Vasco · Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA · FACULTAD DE INFORMÁTICA
Konputazio Ingeniaritza eta Sistema Adimentsuak Unibertsitate Masterra
Máster Universitario en Ingeniería Computacional y Sistemas Inteligentes

Acknowledgements

I want to acknowledge the contributions made to this work to physician and PhD Daniel Cobos for the orientation provided in the most theoretical parts as well as the revision of my work. In addition, I want to thank PhD Maite Oronoz for her contributions in the correctness of this study.

Summary

Civil registration and vital statistics registers births and deaths and compiles statistics. These statistics are a key factor to promote public health policies, register longevity and the health of the population. Death certificates issued in health institutions are the main source to collect the cause of death (CoD). Nevertheless, such counts are not straightforward, indeed, it is estimated that 65% of deaths in the world remain uncounted [D'Ambruoso, 2013]. In places where there is no access to health facilities and, hence, to death certificates, the World Health Organization (WHO) designed the Verbal Autopsy as an instrument to collect evidences about the CoD statistics.

A Verbal Autopsy (VA) consists of an interview to the relative or the caregiver of the deceased. The VA conveys both an open response (OR) and the closed questions (CQs). On the one hand, the OR consists of a free narrative of the events expressed in natural language and without any pre-determined structure. On the other hand, the CQs are a set of a few hundreds controlled questions each with a small number of permitted answers (e.g. yes/no).

InterVA is a suite of computer models and it is included in the WHO 2016 instrument, which gathers several algorithms chosen by the WHO for the analysis of verbal autopsies. InterVA estimates the CoD, based, merely, upon the CQs while the OR is disregarded. We hypothesize that the incorporation of the text provided by the OR might convey relevant information to discern the CoD and, accordingly, InterVA could be benefited from Natural Language Processing approaches. Empirical results corroborated that the CoD prediction capability of the InterVA algorithm is outperformed taking into account the valuable information conveyed by the OR. The experimental layout compares InterVA with other approaches well suited to the processing of structured inputs as is the case of the CQs. Next, alternative approaches based on language models are employed to analyze the OR. Finally, the best approach for each facet (CQs and OR) was combined leading to a multi-modal approach.

Contents

List of Figures

List of Tables

1. Introduction

This chapter introduces the project through a description of the task together with a description of the main challenges, a description of the context in which it is framed and the proposed objectives.

1.1 Project description

The main source of information for mortality rates are medical reports of causes of death certified by a qualified physician. However, in many parts of the world there is no access to hospitals and health professionals, so many diseases are not captured in mortality statistics, causing the health priorities of these countries to be overlooked in the decision-making process of international health policies. In addition, increased registration of causes of death can lead to the detection of disease outbreaks that also need to be taken into account in international decision-making.

In order to collect mortality data in places without access to health facilities, the WHO (World Health Organization) endorses verbal autopsies (VAs) as the second best option in the absence of death certificates. A verbal autopsy is an instrument used to collect causes of death in places where there is not a health cover that allows to certify deaths. A verbal autopsy consists of an interview with the caregiver of the deceased. The interviewer must be familiar with the language of the interviewee and does not need to hold medical expertise. Verbal autopsies include, on the one hand, an open response in which the interviewee is given the opportunity to openly describe the events preceding the death, and on the other hand, closed questions about symptoms and signs preceding the illness. A sample of the data set for the adult age group is depicted in Figure 1.1.1. The responses

from the questionnaire are analyzed by a clinician or by an algorithm to determine the cause of death. That is, the input to the system or any algorithm are the open response and the closed questions and the output is the cause of death. In addition, some complete open responses are shown in Table in 2.3.2 with the corresponding cause of death. In figure 1.1.2 the cause of death distribution is shown for all the verbal autopsies in the dataset along with the list of all the possible causes of death in the x-axis and the module (i.e. age group) where each of them have been assigned.

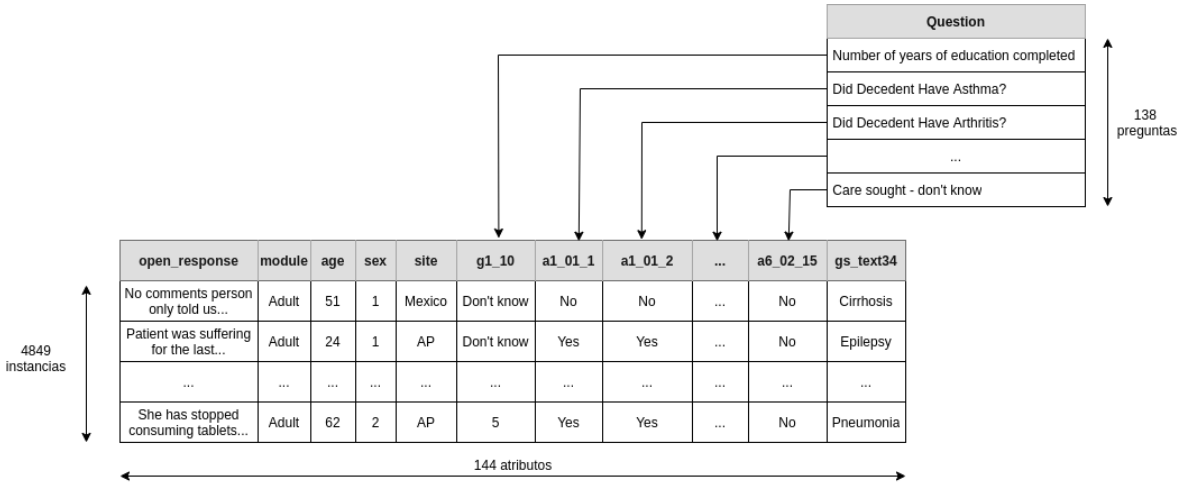

Figure 1.1.1: Sample verbal autopsy data set for the adult age group along with the relationship between the question codes and the statements used by the interviewer.

Open Response (OR)	Cause of Death (CoD)
father set on fire both mother and baby. baby died in the afternoon. father was in love with some other lady and to get married to her he killed his wife and baby.	Fires
respondent said that the deceased died due to high fever, diarrhea and vomiting	Diarrhea/Dysentery
the deceased had deliver twins which died in 3 days. after that the deceased suffered excessive bleeding.	Maternal

Table 1.1.1: Examples of three open responses (ORs) with the corresponding cause of death (CoD).

Causes of death are given in ICD-10 (International Classification of Diseases) format, a standard for documenting diseases in clinical cases. Each disease is assigned an ICD-10 from a a total of 48 codes for verbal autopsies, divided into three age segments (adult, child and neonate).

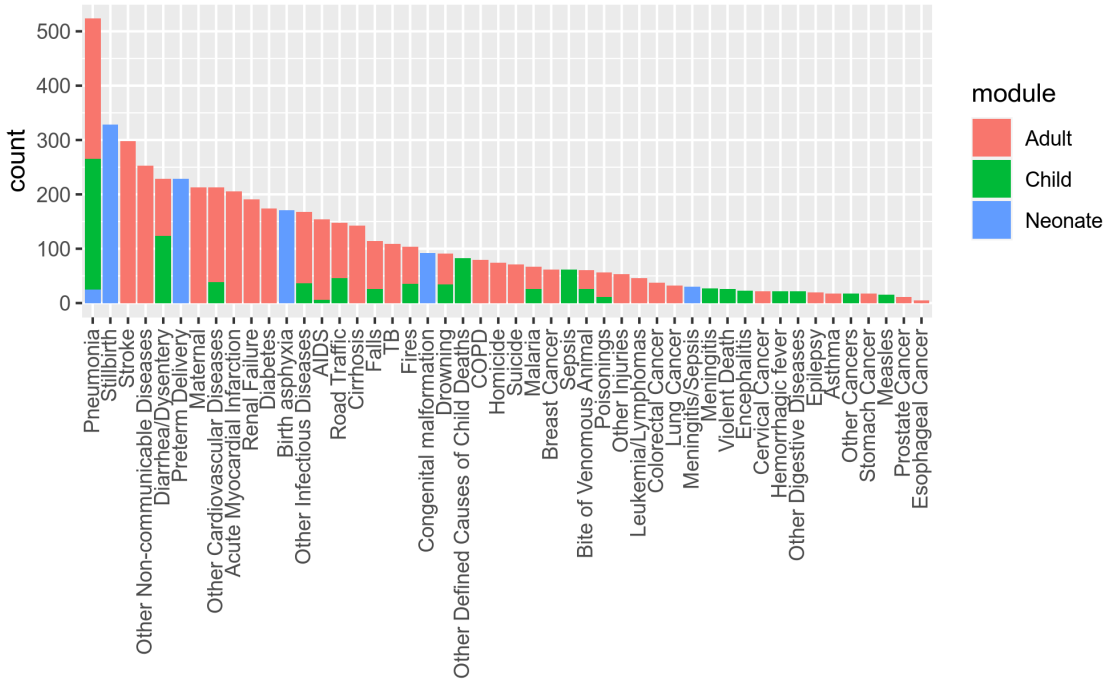

Figure 1.1.2: Cause of death distribution where the x-axis represents the list of all the causes of death present in the verbal autopsis and the y-axis the number of verbal autopsies with that cause of death assigned. In addition, each cause of death count is divided per each of the modules (i.e. age groups): adult, child and neonate.

1.2 Challenges

The topic addressed in this TFM presents several research questions. On the one hand, there is international interest in reducing the 299 closed questions used in the questionnaire to the minimum necessary to know the cause of death from a limited list of causes of death. In this way, it will be possible to remove or reformulate those questions that are not relevant in the identification of the cause of death and thus, maximize the performance of the models generated to perform this task. In addition, reducing the questionnaire would have several benefits since the questions have to be translated and asked to a person whose relative has just died. Therefore, shortening the questionnaire would speed up the whole process and improve sensitivity with the relatives. The WHO has developed several versions of the questionnaire in order to optimize it as much as possible, putting as few closed questions as possible and helping as much as possible in the prediction task of the cause of death. This year, the new WHO 2022 questionnaire has been published [WHO, 2022] [Chandramohan et al., 2021]. This challenge has been tackled and further analyzed in my publication [Cejudo et al., 2021].

Another challenge of this task is to improve the predictive ability of the cause of death from questionnaire responses. Currently, there are several methods endorsed by the WHO: InterVA [Fantahun et al., 2006], InSilicoVA [Clark et al., 2013] and Tariff [Serina et al., 2015b]. All these algorithms are collected in The WHO 2016 tool [Nichols et al., 2018a] and implemented in the R package OpenVA [Li et al., 2017]. This package offers

several functionalities such as loading data from a public repository of verbal autopsy data, adapting them to different questionnaires and applying the aforementioned algorithms. However, studies such as Li et al. [2018] have reported a performance for InterVA of 21.77% by means of accuracy in the cause of death prediction task. In the results reported by Flaxman et al. [2018a], InSilico achieves an accuracy of 37.6% while according to McCormick et al. [2016] using all the algorithms available in The WHO 2016 and top 3 accuracy, achieves results that are below 60%.

The verbal autopsy questionnaire does not only include closed questions, it also offers for each verbal autopsy an open response, in which the interviewee comments on everything he or she considers relevant prior to death. However, none of the methods within The WHO 2016 use this narrative for the analysis and the prediction of cause of death. Nevertheless, there are studies that have analyzed the predictive ability of the cause of death by using only the information provided by the open response. Danso et al. [2014] used simple methods to represent the text of the open response and by using a SVM classifier [Noble, 2006] and a TF-IDF representation [Ramos and Juan, 2003] on a different data set they reach a 41.9% of accuracy [Powers, 2020]. Yan et al. [2019] explored distinct techniques similar to *word-embeddings* to represent text, employing the *Million Death Study* dataset reporting a 75.1% on the F-measure metric [Powers, 2020].

Current data mining algorithms are effective when large volumes of data are available. However, the work at hand is challenging due to the limited amount of verbal autopsies available, namely 5,509 verbal autopsies. Proposing an algorithm that is able to generalize and correctly classify new verbal autopsies from these data is one of the biggest challenges for many of the algorithms used in the field of artificial intelligence. On the other hand, algorithms tend to be designed to discriminate between two values (e.g. yes/no). However, as mentioned above, there are 48 different causes of death which hinders the predictive ability of any model created to predict these diseases. Therefore, finding a method that automatically reduces the questionnaire and can extrapolate it to all the new records obtained while ensuring a similar efficacy as in the original questionnaire is a difficult task. In addition, it is necessary to find the most appropriate way of working with the text, given that the algorithms used only understand numbers.

Taking these challenges into account, it is of great interest to combine the information provided by both the closed questions and the open response in order to increase the gain of information and, therefore, obtain a better result when predicting the causes of death. In addition, it is of interest using the algorithms endorsed by the WHO and collected in The WHO 2016 tool, adding the text to these algorithms and evaluating the results with respect to those obtained using only the closed questions. Another option is to test other classifiers, thus comparing the performance with the WHO-endorsed algorithms and see if it is more beneficial to add the text to the proposed algorithm.

1.3 Objectives

Taking into account the different challenges presented by this task and the work already done on questionnaire reduction, the **main objective** of this work is to add the information provided by the open response to the analysis. The aim is to extract relevant elements provided by the verbal autopsies and to study the predictive capacity that can be achieved by integrating the textual information.

For this purpose, we propose to study and search for the best algorithm to classify using the closed questions of the questionnaire and, on the other hand, to obtain the best model to classify employing the open response. Subsequently, a model known as *ensemble* will be proposed, which is responsible for combining the models generated for the prediction of the cause of death using separately the open response and the closed questions.

Another objective is to use one of the algorithms endorsed by the WHO and, since none of them use the open response for the analysis, to add textual information in this work and to study whether the performance of these algorithms can be improved.

Finally, the aim is to apply what has been learned during the master's degree and use state-of-the-art models to predict the cause of death from the open response. To this end, pre-trained and transformer-based models will be explored, which have proven to be very useful in small datasets as is the case when learning from datasets with a much larger size.

To sum up, a brief list of the different objectives is provided:

1. Include both the closed questions and the open response.
2. Include one of the algorithms of The WHO 2016 tool to the analysis.
3. Use transformer-based models and apply some of the techniques learnt in the master's degree to make the most of the open response.

1.4 Contextualization

This TFM is a research study. It is confidential and arises from the collaboration of the IXA research group with Swiss TPH, an international public health institute. This topic is proposed after a meeting between several political leaders from different countries in which they expressed their interest in modifying the questionnaire.

This work has a strong artificial intelligence component, not only because of the learning task, but also because of the application of that knowledge in some of the experiments. As a result of my work on the verbal autopsies and the collaboration with the IXA research group of the UPV/EHU, this academic year 2022/23 I have managed to publish two papers. From the verbal autopsies we will try, using artificial intelligence and specifically data mining, to extract the cause of death by processing these clinical texts. For the application of these two areas, this project will also have a software component that will aim at

modularization and reusability.

2. Approach

This chapter includes the paper-format study created as the result of this work, since I have chosen to present the TFM in this format. The contribution of this work consists in experimentation, on the one hand, with closed questions including InterVA and, on the other hand, with open response through a series of models based on transformers.

In this way, the results obtained will be used to measure the predictive capacity using only each of the verbal autopsy features separately. Subsequently, these results will be used to make a comparison with the proposed *ensemble* model. This model uses both the open response and the closed questions as input in order to predict the cause of death.

The way in which the experiments have been designed allows us to answer a series of **research questions**:

- **RQ1:** What predictive ability is obtained with closed questions (CQ)?
- **RQ2:** What predictive capacity is obtained with the open response (OR)?
- **RQ3:** Does combining both CQs and OR allow for improved predictive ability on cause of death?
- **RQ4:** Is it possible to improve InterVA performance by adding the open response to the analysis?
- **RQ5:** Which of the input models within the ensemble model has more impact in the final output?

All of these questions are discussed and explained in the study added below.

Civil registration and vital statistics registers births and deaths and compiles statistics. These statistics are a key factor to promote public health policies, register longevity and the health of the population. Death certificates issued in health institutions are the main source to collect the cause of death (CoD). Nevertheless, such counts are not straightforward, indeed, it is estimated that 65% of deaths in the world remain uncounted [D'Ambruoso, 2013]. In places where there is no access to health facilities and, hence, to death certificates, the World Health Organization (WHO) designed the Verbal Autopsy as an instrument to collect evidences about the CoD statistics.

A Verbal Autopsy (VA) consists of an interview to the relative or the caregiver of the deceased. The VA conveys both an open response (OR) and the closed questions (CQs). On the one hand, the OR consists of a free narrative of the events expressed in natural language and without any pre-determined structure. On the other hand, the CQs are a set of a few hundreds controlled questions each with a small number of permitted answers (e.g. yes/no).

InterVA is a suite of computer models and it is included in the WHO 2016 instrument, which gathers several algorithms chosen by the WHO for the analysis of verbal autopsies. InterVA estimates the CoD, based, merely, upon the CQs while the OR is disregarded. We hypothesize that the incorporation of the text provided by the OR might convey relevant information to discern the CoD and, accordingly, InterVA could be benefited from Natural Language Processing approaches. Empirical results corroborated that the CoD prediction capability of the InterVA algorithm is outperformed taking into account the valuable information conveyed by the OR. The experimental layout compares InterVA with other approaches well suited to the processing of structured inputs as is the case of the CQs. Next, alternative approaches based on language models are employed to analyze the OR. Finally, the best approach for each facet (CQs and OR) was combined leading to a multi-modal approach.

2.1 Introduction

A Verbal Autopsy (VA) consists in a series of questions about the signs, symptoms, demographic characteristics and the condition that led to death answered by the relatives or the caregiver of the deceased. The VA instrument includes closed questions (CQ), and an open response (OR) where the interviewees can talk freely about how the death occurred. Expert clinicians are able to discern the most probable CoD from the answers conveyed in the VA. Needless to say, manual inspection of VA is time consuming for expert clinicians.

Since the verbal autopsies were proposed, many studies such as Yang et al. [2006] in China, Setel et al. [2006] in Tanzania and Baqui et al. [1998] in Bangladesh, have validated the procedure and the effectiveness of the VA. However, many other studies have pointed out the limitations. For example, Todd et al. [1994] found out that the effects of Malaria are often mistaken, besides, Baqui et al. [1998] and Garenne and Fauveau [2006] summarise

further limitations like that the usefulness of the verbal autopsy is closely related with the quality of the conducted interview.

In order to extract the CoD, VAs can be analyzed either by physicians or with automated algorithms. Given the scale of some VA implementations, manual coding of VAs is becoming unrealistic for many countries and the use of automated methods is getting traction. To delve into the quantitative assessment of the VA, the Institute for Health Metrics and Evaluation (IHME) along with other organizations conducted the PHMRC study [Murray et al., 2011] where they tested the suitability of VAs in hospitals where the CoD was known. As a result of this study, a public verbal autopsy gold-standard was released. Since then, some tools and algorithms have been implemented to automatically estimate the CoD given the CQs [Flaxman et al., 2018a] [Serina et al., 2015a]. Automatic classification of VA into CoD offers a mean to alleviate the burden of manual analysis and speed up the collection of CoDs to aid in-time decision making of public health policies.

One of the algorithms tested in the PHMRC data was InterVA [Fantahun et al., 2006], indeed, one of the WHO standards. The core idea of this algorithm is to mimic the behaviour and decisions of a physician in terms of a heuristic algorithm based on hand-crafted rules and it is being improved over the years [Byass et al., 2012, 2019]. Later, as an alternative to heuristics, machine learning strategies were incorporated as is the case of Tariff 2.0 [Serina et al., 2015a] and InSilicoVA [Flaxman et al., 2018a]. InSilicoVA identifies the most likely joint probability distribution of cause-specific mortality fractions. Eventually, the WHO 2016 VA instrument [Nichols et al., 2018b] was released adopting the InterVA and InSilicoVA methods as the most effective means to analyze verbal autopsy data. All these methods have in common that they estimate the CoD taking into account just the CQs while the OR gets disregarded. Nonetheless, several works, such as Yan et al. [2019] have employed the OR for VA analysis.

In this work we explore the potential of the OR to contribute in the assignment of the CoD and our first **goal** is to quantitatively assess the predictive capabilities of the OR in comparison to the CQs. We hypothesise that the OR brings valuable information that should not be disregarded by automated methods. Besides, we wonder if CQ and OR are redundant or, instead, are synergistic. Given that the OR was as useful as the CQ to get the CoD, the implication would be that we could remove the complex hierarchical questionnaire and simply record the OR. If both CQs and OR would complement one the other, this would imply that we could gain accuracy in the prediction of the CoD. To cope with the aforementioned research questions, first, we compared InterVA with alternative approaches well suited to analyze structured information, as is the case of the CQs e.g. XGBoost algorithm [Chen et al., 2015]. After having been able to estimate the CoD given, merely, the CQs, next, we made the prediction given just the OR. In order to analyze the OR we resorted to state-of-the-art techniques in Natural Language Processing based on language models such as pre-trained transformer based models. With the best approach for each of the two input facets (CQ and OR) a combined model was built resulting in noticeable improvements with respect to InterVA.

2.2 Related work

The PHMRC data-set is not the only known VA data-set, the Million Death Study (MDS) [Jha et al., 2006] collected the verbal autopsy of thousands of deaths that occurred in India. There is also a data-set collected in Ghana [Danso et al., 2013] and another in Malasya [Ganapathy et al., 2017], among others. However, as far as we know, the only one publicly available is the PHMRC data-set, due to the fact that verbal autopsies might convey sensitive information and these data-sets are often private.

There are multiple studies that have tried to assign and measure the most likely CoD to each verbal autopsy. Some of these studies like Li et al. [2018], using the PHMRC data-set and the OpenVA [Li et al., 2017] toolkit, which includes algorithms like InterVA and InSilicoVA, have reported a poor performance. That is, a 21.24% of accuracy for InterVA and being the highest performance reached a 37.77% for the NBC [Rish and Irina, 2001] algorithm. They also provide results at population level and not only at verbal autopsy individual level. Flaxman et al. [2018b] have carried out a very similar study, but only with InSilicoVA, reporting a maximum accuracy of 37.6% for CoD assignment. McCormick et al. [2016] have compared also the algorithms inside the WHO 2016 instrument having as input the PHMRC data-set and top 3 CoD accuracy evaluation getting results under the 60%.

As the standard methods have achieved a low CoD prediction capability, a variety of studies have handled this task employing different machine learning and deep learning methods for CoD assignment having as input either the OR or the CQs. Danso et al. [2014] have used simple text representation methods in a different data-set of VAs collected in Ghana and the Support Vector Machine (SVM) [Noble, 2006] classifier reported the highest macro F1-score score [Hripcsak and Rothschild, 2005] of 41.9% with TF-IDF [Ramos and Juan, 2003] representation. Yan et al. [2019], instead, used word and character embeddings with the MDS data-set and achieves a 75.1% in F1-score metric for the adult age group.

For CQs, Moran et al. [2021], splitting and testing into different subsets of the PHMRC data-set, have made a comparison between some of the WHO 2016 instrument algorithms, such as the NBC, Tariff and InSilicoVA and their own Bayesian-based classifier named BF, achieving better results. Li et al. [2020] did the same comparison but using a Gaussian mixture [Reynolds, 2009] that seems to also improve the WHO 2016 algorithms.

2.3 Materials

This work makes use of the Verbal Autopsy Golden Standard (VA-GS) generated as the result of the PHMRC study [Murray et al., 2011] and made publicly available. In order to learn from VAs and evaluate the performance of the generated models, the data is splitted into train and test sets with stratified subsets with the 70% and 30% of the VAs, respectively. In Table 2.3.1 a description of the whole data-set is shown. All together there are above 7,400 samples randomly divided into train and test subsets with stratification. Each VA is

described with two types of features (OR and CQs) and has a unique CoD assigned out of a total of 48 possible CoDs. The CQs are mainly categorical. Regarding the OR, the average length is 90, 75 and 87 for adult, child and neonate, respectively. Nonetheless, ORs with less than 10 words and over the 200 words can be found in the data-set. Naturally, some CoDs are more frequent than others. For instance, 524 VAs were cases of pneumonia while only 20 VAs had epilepsy as the CoD. On average, there are 109 VA per CoD and, per age group, 100 for adult, 45 for child and 146 for neonate. Nevertheless the deviation from the average is high as the CoD class is unbalanced. That is, there are different CoD and the number of VAs per CoD is significantly different. Nevertheless, it is important to keep the same proportion of the CoD in both train and test sets, to this end the train and test subsets were randomly selected with stratification. The split is done according to Blanco et al. [2020] so the results can be compared in the experimental results.

In addition, it should be taken into account that the questionnaire differs per age group. For the adult segment, there are 142 CQs, 86 for child and 109 for neonate, including sex, site, age and age group as CQs. A big number of CQs lack of value due to two phenomena: the presence of values such as ”Don’t know” on the one hand and the so-called skip-patterns e.g. questions that were not asked as are considered unnecessary given the previous answers or contexts. This skip patterns are applied in some cases when, for a determinate CQ, a ”Don’t know” is answered and some of the subsequent CQs that follow up in the questionnaire are not asked, getting also a ”Don’t know” value. Another case is when, for example, the deceased is male and the questions are designed to be answered by a woman. For instance, one of the CQs of the questionnaire is ”Was the deceased a singleton or a multiple bird?” and if the answer is ”Dont’ know”, it makes sense to put the same value in the next question of the questionnaire: ”Was this the first, second or later in the birth order?”. In addition, the CQs have on average 3 to 4 possible answers.

VA-GS data-set				Adult	Child	Neonate	Total
train	Sample instances			3.389	945	875	5.209
	Input: Features	OR	|Vocabulary|	8.056	3.059	3.284	9.540
			# Words	306.189	71.759	76.293	454.238
		CQ	Categorical	140	84	108	303
			Numerical	2	2	1	3
	Output: Class	CoD	Count	34	21	6	48
test	Sample instances			1.460	396	377	2.233
	Input: Features	OR	|Vocabulary|	5.407	2.358	2.189	6.379
			# Words	131.068	29.264	29.893	190.225
			OOV	1.384	612	549	1.641
		CQ	Categorical	140	84	108	303
			Numerical	2	2	1	3
	Output: Class	CoD	Count	34	20	6	48

Table 2.3.1: Description of the VA-GS data-set. Each verbal autopsy is described in terms of bi-modal information, i.e. an open response (OR) in terms of free text and a set closed questions (CQs), and has annotated the cause of death (CoD). The VA data-set was divided into train and test. Note that the set of CQs formulated to ascertain the CoD is dependant on the each age-segment (adult, child and neonate).

In Table 2.3.2, two examples of the OR and the corresponding cause of death are shown. As it can be seen, the type of language used is non-technical and from the first example the corresponding CoD can be easily extracted. The CoD corresponds to an ICD-9 code from a finite list of possible ICD-9 codes for this task, which has a name associated like, for example, "Pneumonia". In the second example, there is a grammatical mistake and the first two words do not add any useful information. This happens throughout some of the ORs, many have errors and others do not give any information while some ORs have the CoD explicitly.

Open Response (OR)	Cause of Death (CoD)
father set on fire both mother and baby. baby died in the afternoon. father was in love with some other lady and to get married to her he killed his wife and baby.	Fires
no comments they were twins. the boy died because his lungs were not developed, and had respiratory problems.	Preterm delivery

Table 2.3.2: Examples of two ORs with the corresponding CoD.

2.4 Methods

This section presents the methodological approach employed to estimate the CoD. As shown in Table 2.3.1 each VA is characterized by means of two facets or a bi-modal input $(CQ, OR) = (\mathbf{x_1}, \mathbf{x_2})$ and the aim is to estimate a CoD as the output of the approach. In

this section we present different approaches to exploit the dual input. Each modality of the input shall be characterized by a feature-vector, respectively, of size m and n as in (2.4.1).

$$\begin{aligned} \mathbf{x} &= (\mathbf{x_1}, \mathbf{x_2}) \in \mathbb{F}^m \times \mathbb{R}^n \text{ with} && (2.4.1)\\ \mathbf{x_1} &= (x_{11}, \cdots, x_{1m}) \in \mathbb{F}^m \text{ to represent the CQs} && (2.4.2)\\ \mathbf{x_2} &= (x_{21}, \cdots, x_{2n}) \in \mathbb{R}^n \text{ to represent the OR} && (2.4.3) \end{aligned}$$

In section 2.4.1, the inference of the CoD given the CQs is presented while in 2.4.2, the results given the OR are shown. Finally, in 2.4.3, the proposed approach to cope with the dual OR and CQs is presented. The novelty rests on the development of models capable of learning from VAs, not only from the CQs but also from the OR and, through this, the fact of having improved InterVA for CoD estimation.

2.4.1 Models based on closed questions

In this section we explore approaches well suited for the CQs and able to infer classifiers from either categorical or numerical feature-vectors (denoted, for simplicity, as in an m-sized space $\mathbb{F}^m$). In general, we shall refer to these methods as f_{CQ2CoD}, as in (2.4.4), due to the fact that they are able to compute the likelihood of the i-th CoD (y_{1i}) given the responses to the set of CQs ($\mathbf{x_1} \in \mathbb{F}^m$). Thus, each y_{1i} is bound to the interval $[0,1]$.

$$\begin{aligned} f_{CQ2CoD} : \mathbb{F}^m &\longrightarrow [0,1]^{|CoD|} \\ \mathbf{x_1} &\longrightarrow f_{CQ2CoD}(\mathbf{x_1}) = (y_{11}, \ldots, y_{1|CoD|}) = \mathbf{y_1} \end{aligned} \tag{2.4.4}$$

Typically, the estimated CoD, formally $\hat{c}_{CQ2CoD}$, is the most likely CoD, as in (2.4.5).

$$\hat{c}_{CQ2CoD} = \arg \max_{i=1}^{|CoD|} y_{1i} \tag{2.4.5}$$

Among the CQ2CoD approaches we count, on the one hand, with the standard InterVA-4 included in the WHO 2016 instrument [Nichols et al., 2018b]. The InterVA-4 algorithm is available in the OpenVA package [Li et al., 2017]. This package is implemented in R and it offers functionalities such as downloading the VA data, parsing the information into different formats and also functions to train and assess the performance for CoD estimation

in VAs. InterVA implements a hand-made decision tree implementing the heuristic that an expert clinician would follow based on the standard way of discarding diseases. In brief, it implements a series of and-or rules e.g. if the response to question i was 'negative' and the response of question j is 'high' or 'medium', then, the CoD determined is 'Pneumonia'.

On the other hand, beyond handcrafted methods, we could rely on data-driven supervised inference approaches. Among them, we considered XGBoost [Chen et al., 2015], a gradient boosting algorithm. This algorithm is based on an ensemble of decision trees in a sequence, where each decision tree is adapted to minimize the errors made by the previous tree as it has as the input the output provided by the previous tree. The sequential adding of decision trees is done until the error can not be further improved and this is called gradient descent. A variety of parameters can be set in order to maximize the performance for the specific task that it is used for: number of iterations (i.e. the number of trees to use), the maximum depth of the decision trees that are inferred during the training process and η as the learning rate. The XGBoost classifier was shown to be one of the best methods with the highest performance not only in our preliminary experiments but also in a related task by Blanco et al. [2020].

Both methods included in the CQ2CoD approach output a probability for each CoD being the output bound to [0,1] , that is, how likely is the input VA to belong to each of the possible CoD.

Note that the CQs are suited for each age-segment (Adult, Child, Neonate) and, hence, a different set of CQs (i.e. 142 for Adults, 86 Children and 109 for Neonates) are involved in each age-segment (see Table 2.3.1). This implies that the value of m, the size of the input in (2.4.4) varies by each age-segment and, accordingly, the function f_{CQ2CoD} is suited by age-segment. This is the case of both InterVA (incluiding handcrafted rules by segment) and also inferred estimators (a specific model has to be inferred by segment).

2.4.2 Models based on open response

As a second approach, we explored the OR as valuable information to ascertain the CoD as in (2.4.6).

$$\begin{aligned} f_{OR2CoD} : \mathbb{R}^n &\longrightarrow \mathbb{R}^{|CoD|} \\ \mathbf{x_2} &\longrightarrow f_{OR2CoD}(\mathbf{x_2}) = (y_{21}, \ldots, y_{2|CoD|}) = \mathbf{y_2} \end{aligned} \tag{2.4.6}$$

Again, the estimated CoD, $\hat{c}_{OR2CoD}$, is the most reliable CoD, as in (2.4.7).

$$\hat{c}_{OR2CoD} = \arg \max_{i=1}^{|CoD|} y_{2i} \tag{2.4.7}$$

Figure 2.4.1 shows an example of the output of this approach. The input to the model is the OR (a short text) and the output is an array ($\mathbf{y_2}$) with weights related to the reliability of each code as denoted in (2.4.6). For the input string given in the example the system estimates that the CoD 'Asthma' shows a reliability of $y_{2Asthma} = -11.0$ and, thus, is less reliable than 'Suicide' (with $y_{2Suicide} = 4.4$) but 'Homicide' has the highest reliability (with $y_{2Homicide} = 58.6$). Then, as in (2.4.7) the model provides the most reliable code ($\hat{c}_{OR2CoD}$ is 'Homicide' in the example). Note that, in the example, the expected or gold CoD was, by contrast, 'Suicide', meaning that the system failed the estimation.

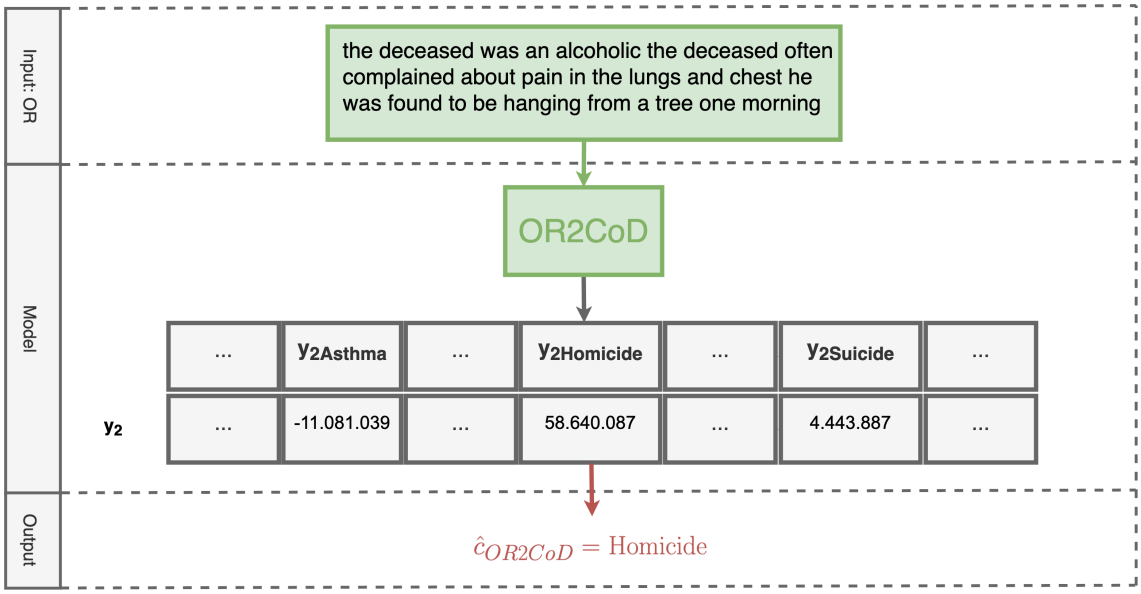

Figure 2.4.1: Example of an incorrect estimation. On the top of the figure we have the Input (OR), and the expected or actual CoD (Suicide). For that input in the test set BERT predicted the output $\widehat{CoD}$ = Homicide as it was the CoD with the highest output value among the all possibles outcomes.

A key issue of this approach is the conversion of the free narrative into a meaningful numeric feature-vector ($\mathbf{x_2}$). Classical techniques like Bag of Words (BoW) [D'Souza, 2018] and TF-IDF [Ramos and Juan, 2003] make possible such a vectorization. Nevertheless, these simple approaches do not convey contextual information as the strings are represented by mere count of the presence of words in the narrative. Besides, these representations suffer of high dimension (large n) due to the fact that the vocabulary involved in free narratives tends to be large (see the vocabulary involved in our task in Table 2.3.1). Instead, word embedding [Mikolov et al., 2010] is one of the emerging successful methods being currently employed to represent the text with numeric vectors. The added value of these methods rests on the fact that they encompass the contextual information of symbols (words) into dense numeric vectors of small dimension (small n).

Admittedly, having represented the OR into a vector, the information provided by the OR facet can be explored with classical data mining approaches such as Support Vector Machines [Noble, 2006], Logistic Regression [Kleinbaum et al., 2002], XGBoost itself and RNNs [Mikolov et al., 2010] among others. Beyond the classical approaches, recent models like RNNs [Mikolov et al., 2010] and transformers [Vaswani et al., 2017] are suited

to learn the features i.e. the word embeddings to effectively represent the input corpus. Moreover, the transformers count on attention mechanisms that learn to focus in different parts of the input data in order to accurately interpret the contextual information. Current trends in natural language understanding rest on language modelling as a mean of keeping contextual information rather than mere symbolic representations. Transformer based models like BERT [Devlin et al., 2018] are gaining importance to solve difficult tasks that require from looking at the contextual information implicit in the language. BERT is composed of stacked encoder layers which have, mainly, two functions: several heads and a feed forward layer. A head is how an attention mechanism is called and the one that focuses more in some words rather than orders to maximize the final prediction accuracy. The information conveyed in the language model can be extrapolated in down-stream tasks involving language understanding. In this case, pre-trained transformers seem appropriate to cope with language modeling while fine-tuning the approach to the estimation of the CoD.

In this work a variety of transformer based models were considered and included among the OR2CoD approaches: Small BERT, BERT [Devlin et al., 2018] and BioClinical BERT [Alsentzer et al., 2019]. These transformer based models provide weights for each CoD, not a probability, that is, y_{2i} elements in (2.4.6) are not bound to $[0, 1]$ and could be negative as in the example provided in Figure 2.4.1. Small BERT and BERT are trained from a general knowledge corpus (i.e. Wikipedia) and they differ in the size of the architecture. Small BERT has 4 encoder layers and 512 heads while BERT has 12 encoder layers and 768 heads. BioClinical BERT is the same as BERT but it was fine-tuned in corpus from the medical domain including articles from PubMed. In practice, smaller models are more suitable for simpler tasks in order to avoid overfitting (i.e. no capacity to generalize for new data) such as Small BERT while BERT is expected to have a better performance is more complex tasks. As BioClinical BERT has the same architecture as BERT but is trained with different data, it is expected to have a soft improvement with respect to BERT in those tasks enclosed in a technical and medical scope.

While the models based on CQs had to be suited for each age-segment (training a particular f_{CQ2CoD} for each age-segment), the models based on ORs are transparent in the sense that a single approach can cope with all the age segments making the f_{OR2CoD} versatile. Nevertheless, as the final dual-input model is not only trained and tested with the whole VA data-set that leads to an additional set of experiments with the input data for OR analysis divided into the different age groups (i.e. adult, child and neonate). Consequently, the amount of data provided for the OR2CoD approaches is significantly reduced for the fine-tuning process of these approaches but also the target CoDs to predict. We also wonder if the proposed pre-trained models in OR2CoD will suffer more from a data amount decrement rather than from having less CoD to ascertain.

2.4.3 Models based on dual input

In order to combine both the OR and the CQs, we propose a dual-input approach. A model based on CQs (e.g. InterVA or XGBoost) and a model based on the OR (e.g. XGBoost or

BERT) mentioned, respectively, in sections 2.4.1 and 2.4.2, are assembled to cope with the insights extracted from both the CQs and the OR. Figure 2.4.2 depicts the architecture proposed.

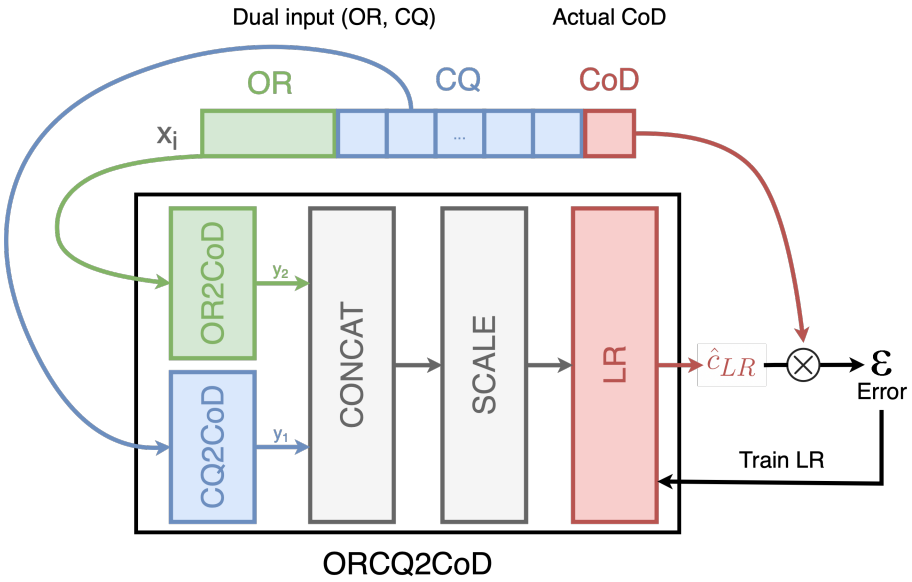

Figure 2.4.2: Proposed ensemble model architecture having either XGBoost or InterVA for closed questions (CQ) treatment and BERT model for the open response (OR). The output of the logistic regression (LR) $\hat{c_{LR}}$ is the final prediction and it is compared with the actual cause of death (CoD). Finally the error (ε) is measured.

The dual input approach relies upon both f_{CQ2CoD} and f_{OR2CoD} to get, respectively, $\mathbf{y_1} \in [0,1]^{|CoD|}$ and $\mathbf{y_2} \in \mathbb{R}^{|CoD|}$, each of which determines the reliability of the available CoDs. Next, both weight-vectors are concatenated, leading to $\mathbf{y} = (\mathbf{y_1}, \mathbf{y_2}) \in [0,1]^{|CoD|} \times \mathbb{R}^{|CoD|}$. The output of each model ($\mathbf{y_1}$ and $\mathbf{y_2}$) has a slightly different meaning. The output given by the models based on CQs ($\mathbf{y_1}$) is probabilistic, values bound to $[0,1]$. By contrast, in the case of the models based on ORs the output, $\mathbf{y_2}$, can entail real values either positive or negative. In order to combine both outputs, an scaling operation is applied in order to adequate the meaning of each output. Note that, the concatenation operation is computed for each input instance. By contrast, the standardization is computed for each input-attribute.

The result of the transformation is, now, the input feature-vector of a simple logistic regression (LR) approach, and is denoted as $\mathbf{x'} = standardize(concat(\mathbf{y_1}, \mathbf{y_2}))$.

$$\begin{aligned} f_{CQOR2CoD} : \mathbb{R}^{2|CoD|} &\longrightarrow \mathbb{R}^{|CoD|} \\ \mathbf{x'} &\longrightarrow f_{CQOR2CoD}(\mathbf{x'}) = (y'_1, \ldots, y'_{|CoD|}) = \mathbf{y'} \end{aligned} \tag{2.4.8}$$

The logistic regression is a classifier that learns a hyperplane for each of the CoD that tries to minimize the error separating as much as possible the VAs with the same CoD as

possible. This hyperplane is inferred with a linear combination between the input and a set of weights (these are learnt in the training process). The capability of this classifier to combine the input is beneficial as it will automatically find the best combination of the input that maximizes the performance of CoD prediction during the training. The set of weights that are learnt is defined in (2.4.10) and the output will be a probability for each CoD computed from a linear combination between the input and the set of weights (i.e. α_i and β_i for the i-th CoD). The main objective is to maximize the prediction accuracy and to validate this method, the logistic regression will at least have the accuracy of the input model with the highest accuracy and it will be expected to take into account the output of both input models to increase the performance of using both of them separately.

In (2.4.9), the input is the concatenation and the standardize of the outputs of the transformer and the XGBoost model. This time, the final output will be $\hat{c}_{LR}$, that is, the predicted cause of death for the i-th verbal autopsy. Then, this prediction is compared with the actual CoD and the error is measured, so the logistic regression is the layer that makes the prediction which is compared with the actual CoD for the ensemble performance measure.

Some preliminar experiments have shown the simple approach of estimating the CoD by choosing the CoD with the highest value in $\mathbf{x}'$ (similar as done in (2.4.7)) to achieve lower evaluation scores. Thus, the need of a more complex way of combining the input is required and that is the reason why a logistic regression approach is employed.

$$\begin{aligned} \mathbf{x}' &= standardize(concat(\mathbf{y_1}, \mathbf{y_2})) \\ y_{LR}(\mathbf{x}') &= (z_1, \cdots, z_{|CoD|}) \in \mathbb{R}^{|CoD|} \quad \text{with} \quad z_i \quad \text{as in (2.4.10)} \\ \hat{c}_{Ensemble} &= \arg\max_{i=1}^{|CoD|} z_i \end{aligned} \tag{2.4.9}$$

The output of the logistic regression, $\mathbf{y}_{LR} = (y_1, \cdots, y_{|CoD|})$, is computed as in (2.4.10), where z_i represents a weight of the input VA being associated with the i-th CoD. In the computation, the parameters α_i are inferred in the training process of the logistic regression approach, which has a vector of weights (α_i) associated and learnt during the training process of the logistic regression. If the learnt weight for the k input, that is α_{ik}, is large, the output given for the i-th cause of death (i.e z_i) will be influenced by the value given by the k input. Extracting this weights open rooms for interpretability as shown in Figures 2.5.2 and 2.5.3, where each row corresponds to the learnt α_i vector and each column is the k input of the model.

$$\begin{aligned} z_i &= \vec{\alpha}_i \times \mathbf{x}' + \beta_i \quad \text{with} \\ \alpha_i &= (\alpha_{i1}, \ldots, \alpha_{i|\mathbf{x}'|}) \in \mathbb{R}^{2|CoD|} \end{aligned} \tag{2.4.10}$$

2.5 Experimental results

In this section we offer and discuss the results given by the aforementioned methods. In subsection 2.5.1, a comparison between XGBoost and InterVA is done having as input the CQs. In subsection 2.5.2, instead, a XGBoost and a variety of transformer based models are compared for the treatment of the OR. Finally, in subsection 2.5.3, the results of the proposed ensemble model are shown.

2.5.1 Assessment of models based on closed questions

In this first experiment, the objective is to measure the performance of the WHO standard InterVA for cause of death prediction and compare it with the proposed XGBoost model. These models only use the CQs divided by age group, as the questionnaire is different for each group. Thus, a model suited for each age segment is assessed.

We have followed the steps shown by Li et al. [2017] and we have made available an R notebook where the XGBoost and the InterVA are compared for the adult group in two widely employed evaluation approaches: 10-fold cross validation and hold-out evaluation. The comparison between both models with hold-out evaluation (i.e. splitting the data-set into train and test sets) is done with the same data partitions and the description of the train and test sets is in Table 2.3.1.

In the training stage, several parameters were adjusted to optimize the performance of the resulting XGBoost model leading to the following values: maximum depth to 2, η value to 0.25, sample type equal to 'weighted' and a grow policy as 'lossguide'. We have found that η has a great impact in the performance and that the optimal number of rounds (i.e. training iterations) is 75 for adult, 15 for child and 33 for neonate.

The results of this experiment can be seen in Table 2.5.1 where the XGBoost outperforms the InterVA in each of the different age groups. For instance, the accuracy for the adult group for InterVA is 27.73% while XGBoost attains a 50.61% of accuracy, that is more than a 20% of improvement in terms of accuracy and the same happens for the F1-score metric. The difference in accuracy between the InterVA and XGBoost is even more dramatic in case of the child and neonate groups, with an increase of almost a 30% in both accuracy and F1-score.

In addition, the only remarkable result is not only the improvement that the XGBoost provides, it is also the performance that the InterVA achieves, as it only classifies correctly less than the 30% of the verbal autopsies for the adult age group. That value is not feasible and even if the data-set used is not as large as we would want to, the InterVA could be expected to get a better result regarding the assignment of the corresponding cause of death to each verbal autopsy.

Model based on CQs	Age group	Accuracy	Precision	Recall	F1-score
InterVA	Adult	27.73	43.57	27.73	29.59
	Child	28.78	34.75	39.96	32.36
	Neonate	46.41	64.06	46.41	46.88
XGBoost	Adult	**50.61**	**49.13**	**50.61**	**50.22**
	Child	**57.57**	**54.17**	**57.57**	**58.14**
	Neonate	**75.06**	**71.86**	**75.06**	**74.38**

Table 2.5.1: Assessment of the models based on CQs (denoted as CQ2CoD in Figure 2.4.2): comparison between InterVA and XGBoost for each age group by means of accuracy, precision, recall and F1-score. The average method is 'weighted'.

2.5.2 Assessment of models based on open response

These set of experiments are related with the use of the OR to extract the cause of death along with a comparison between different approaches with most of them based in transformers architecture: XGBoost, Small BERT, BioClinical BERT and BERT. The goal with these set of experiments is to see whether valuable information can be extracted from the OR and which model performs better.

These 3 BERT-based models were fine-tuned with the training corpus by means of the transformers python package [Wolf et al., 2020] for sequence classification. The computation was carried out in a Google Colaboratory notebook for 10 epochs in a single GPU. In addition, the input tokens are padded until a maximum sequence size fixed in 175.

In table 2.5.1, the performance of different approaches can be seen. In general, BERT seems to get the best results except for the adult set, where the accuracy and the recall is higher with the XGBoost model, and for neonate, where BioClinical BERT gets the best score for precision and F1-score. Taking into account that this corpus is in a medical scope, BioClinical BERT could be expected to get a higher score than the other two transformers-based models. Nonetheless, BERT surpasses BioClinical BERT in most of the cases and it can be due to the corpus used by BioClinical BERT, which is far more technical than the OR collected in the verbal autopsies. At the end of the day, even if our corpus is in a medical scope, it is just what the interviewees say and it can be considered text of general knowledge and, consequently, non-technical.

In Figure 2.5.1, a comparison on the performance of the test set between the three transformer based models can be seen per each of the training epochs. The BERT model has a higher accuracy in all the epochs than the other two approaches, getting the highest score at the end of the training with an accuracy of a 47.55% for 48 different causes of death prediction. Furthermore, the Bioclinical BERT seems to behave like BERT with a lower performance but close until the fifth epoch where it begins to behave more like the small BERT in terms of accuracy.

As the conclusion of this experiment, we think that BERT is the best model to ascertain

the cause of death with the OR except for adult and neonate, but the difference with the XGBoost model for adults is small and the F1-score attained by BERT is higher. In addition, it seems that valuable information can be extracted from the OR and we hypothesize that adding this information to these CQs could be useful to improve the performance when predicting the cause of death.

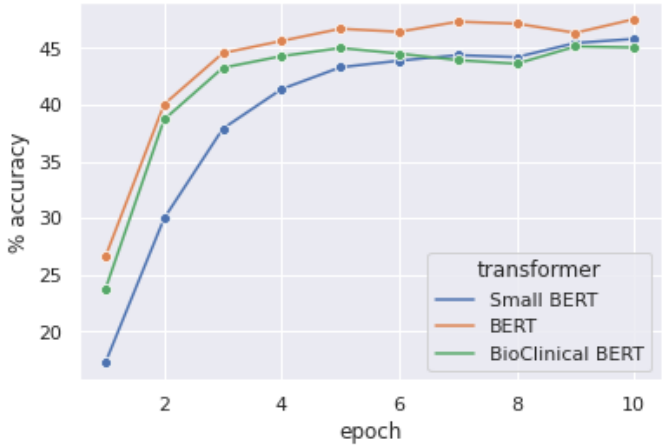

Figure 2.5.1: Assessment of the models based on OR (denoted as OR2CoD in Figure 2.4.2): comparison between three transformer-based models by means of accuracy on the test set for each of the training epochs for all the age groups.

Models based on OR	Age group	Accuracy	Precision	Recall	F1-score
XGBoost [Blanco et al., 2020]	Adult	**45.60**	46.00	**45.60**	44.70
	Child	46.90	44.50	46.90	43.70
	Neonate	59.30	54.20	59.30	55.30
Small BERT	Adult	43.97	48.25	43.97	45.16
	Child	51.01	69.07	51.01	56.63
	Neonate	58.35	66.97	58.35	61.95
	All	45.81	51.04	45.81	47.42
BERT	Adult	45.48	**48.57**	45.48	**46.28**
	Child	**53.78**	**69.21**	**53.78**	**58.93**
	Neonate	**61.80**	71.80	**61.80**	64.79
	All	**47.55**	**51.66**	**47.55**	**48.85**
BioClinical BERT	Adult	43.63	47.18	43.63	44.62
	Child	51.26	60.35	51.26	54.18
	Neonate	60.74	**75.81**	60.74	**66.49**
	All	45.05	47.42	45.05	45.74

Table 2.5.2: Assessment of the models based on OR (denoted as OR2CoD in Figure 2.4.2): BERT-based approaches compared to the XGBoost in the antecedents [Blanco et al., 2020]. The results are presented per age group and measured with accuracy, precision, recall and F1-score. The average method is 'weighted'.

2.5.3 Assessment of models based on dual input

In this final set of experiments we aim to combine the best two models obtained in Sections 2.5.1 and 2.5.2 in order to make use of both the OR and the CQs to predict the cause of death. For the OR, the BERT model will be used while for the CQs, even though the XGBoost would be most suitable model for CoD prediction, testing the performance with the InterVA and determining whether the OR adds valuable information or not is one of the key points of this work.

The results of these set of experiments are shown in Table 2.5.3, given that the proposed ensemble models are based in the architecture represented in Figure 2.4.2 and varying the model that handles the CQs.

Models based on dual input	Age group	Accuracy	Precision	Recall	F1-score
XGBoost+BERT+LR	Adult	**51.57**	**50.89**	**51.57**	**50.89**
	Child	**54.29**	**53.38**	**54.29**	**52.77**
	Neonate	**70.82**	**68.87**	**70.82**	**69.17**
	All	**56.20**	**56.17**	**53.87**	**53.87**
InterVA+BERT+LR	Adult	46.57	46.27	46.57	46.13
	Child	51.01	52.07	51.01	50.23
	Neonate	65.51	61.70	65.51	63.38
	All	46.17	46.62	46.17	45.78

Table 2.5.3: Assessment of the models based on dual input (i.e. ORCQ2CoD in Figure 2.4.2): with the CQs handled by either InterVA or XGBoost and the OR handled by BERT. The final output is given by the logistic regression and the performance is measured by means of accuracy, precision, recall and F1-score. The average method is 'weighted'.

In Table 2.5.3, as expected, the performance of the ensemble model that uses the XGBoost for the CQs is better than the ensemble with the InterVA for all the age groups and also using the whole data-set. However, comparing with the results obtained in Table 2.5.1, the gap between the performance of both models is lower. For instance, now the ensemble with XGBoost gets an accuracy of 51.57% compared to the 46.57% attained by the ensemble with InterVA, that is a difference of the 4% and in Table 2.5.1, the difference between XGBoost and InterVA was higher, more than 20%.

It is also remarkable that with the ensemble model with XGBoost trained to classify per age group the performance is almost the same or even worse than using just the CQs with the XGBoost model. For example, the accuracy for adult using the ensemble with XGBoost is 51.57% while only the XGBoost with the CQs gets 50.61% in Table 2.5.1. For child and neonate the performance decreases when adding the OR. We hypothesize that this is due to the BERT model, that benefits from more data rather than from less classes to predict. This can be seen reflected when using the ensemble for all the age groups, that attains a 56.20% of accuracy, that means that taking into account that BERT gets a 47.55%, adding the CQs leverages that accuracy to a 56.20%, and an increase in performance of almost a 8%.

On the other hand, adding the OR to the InterVA model has led to an accuracy of the 46.56% in case of the adult group, but it has to be taken into account that BERT had an accuracy of 45.48% so it was foreseeable that the ensemble model with InterVA would end up with a similar performance as BERT with a low relative increase respect to BERT, of the 2.3%.

Finally, if both ensemble models are compared, in accuracy, the ensemble with XGBoost for all the age groups is better, with a 10% more in accuracy and for the F1-score measure, an 8% of absolute improvement. In addition, it is important to know from where does that performance of the ensemble with InteVA come.

As pointed out in the research questions RQ5, we want to study the liability of each input in the decision of the ensemble model. That is, we wondered to which extent was relying the LR on the OR and on the CQ. Given that the LR conveys a simple linear combination of both inputs, the answer to this question can be attained by a simple inspection to the parameters involved in the combination. The parameters α_i inferred during the training process of the LR, mentioned in expression 2.4.10 were graphically depicted in Figures 2.5.2 and 2.5.3 in an attempt to sed light to the interpretability of the model. In Figures 2.5.2 and 2.5.3 two heatmaps of the weights learned by the logistic regression layer of the ensemble model are shown. The bigger the weight, the more relevant the input is for the LR to make the predicted CoD. That is, we are comparing which of the elements (the outcome given by either the BERT or the InterVA) are gaining importance when it comes to making the decision through the ensemble approach. The row i and column j of the heatmap in Figure 2.5.2 shows the relevance that the logistic regression has assigned to the credibility given by InterVA to the j-th CoD when it comes to predicting the i-th CoD. Indeed, the LR opts for CoD i combining all the CoDs (all the j columns) or the credibility given by InterVA to all the CoD. Thus, each row conveys the set of weights that the logistic regression has assigned to each of the CoD proposed by InterVA. Implicitly, what the logistic regression learns is, for each cause of death, how reliable is the cause of death value given by the InterVA model. Likewise, Figure 2.5.3 depicts how reliable to the LR is each of the outcomes given by the BERT model.

Taking both Figure 2.5.2 and 2.5.3 into account, it is derived that the ensemble relies much more on the BERT model, as all the BERT inputs in the main diagonal have a large weight, while the logistic regression does not trust in the majority of the InterVA inputs, as they have much smaller weights. The same exploration of the weights have been done for the ensemble model trained with the XGBoost for the CQs and, instead of having just the diagonal with large values as in 2.5.3, there were other relevant values as well. For example, for the output corresponding to 'Suicide' the logistic regression was relying in both BERT and XGBoost. However, for BERT, it was not only relying in the output given for 'Suicide', it was also relying in the output given for 'Homicide'. In Figure 2.4.1, we can see an example of an OR, whose cause of death is 'Suicide', but the BERT model has given the highest value to 'Homicide'. Because this seems to be a common mistake, the logistic regression relies on both values, hence, the LR is also able to reduce the impact of the common mistakes that both input models can make.

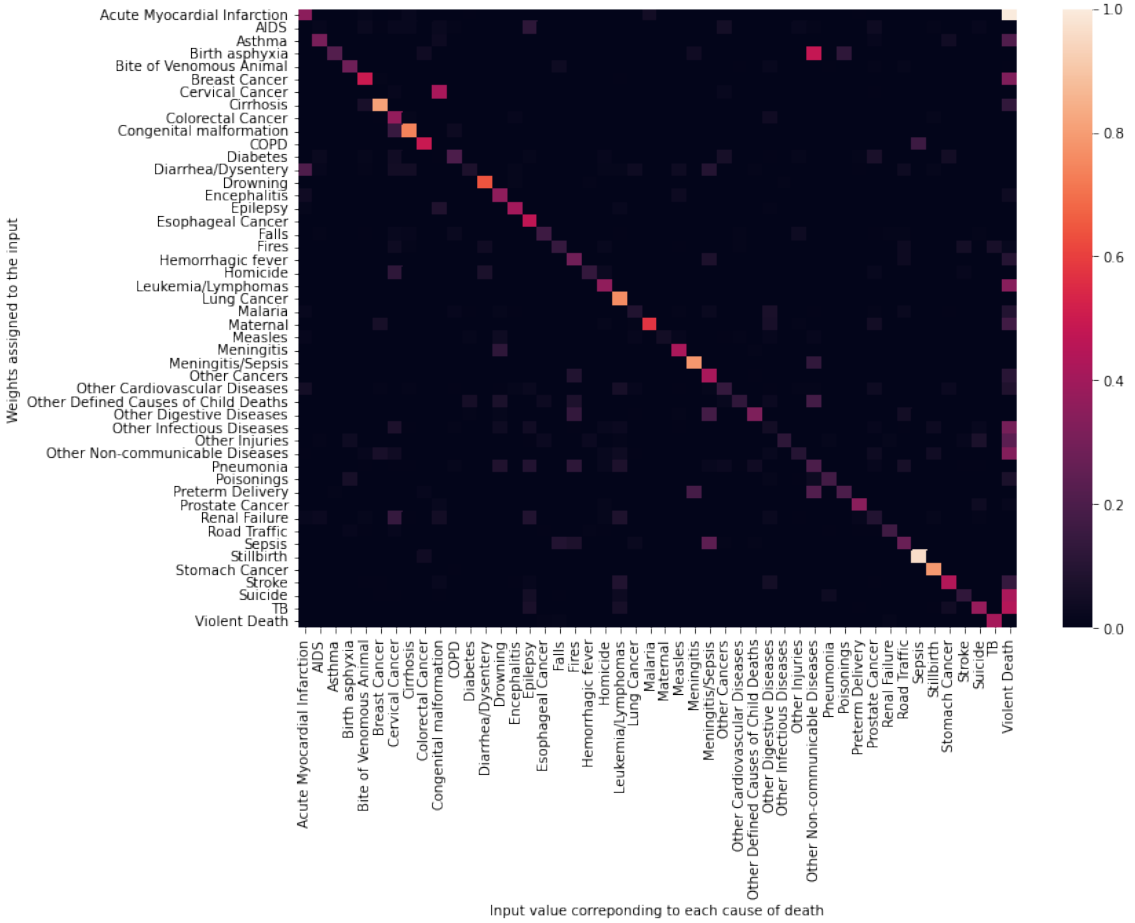

Figure 2.5.2: Heatmap of the weights learned in the last layer of the ensemble model of Figure 2.4.2 by logistic regression for the input given by the InterVA model scaled between 0 and 1. The y axis are the weights learnt for the final output while the y axis indicates to which of the InterVA outputs (i.e. probability given by InterVA for a particular CoD) corresponds that weight.

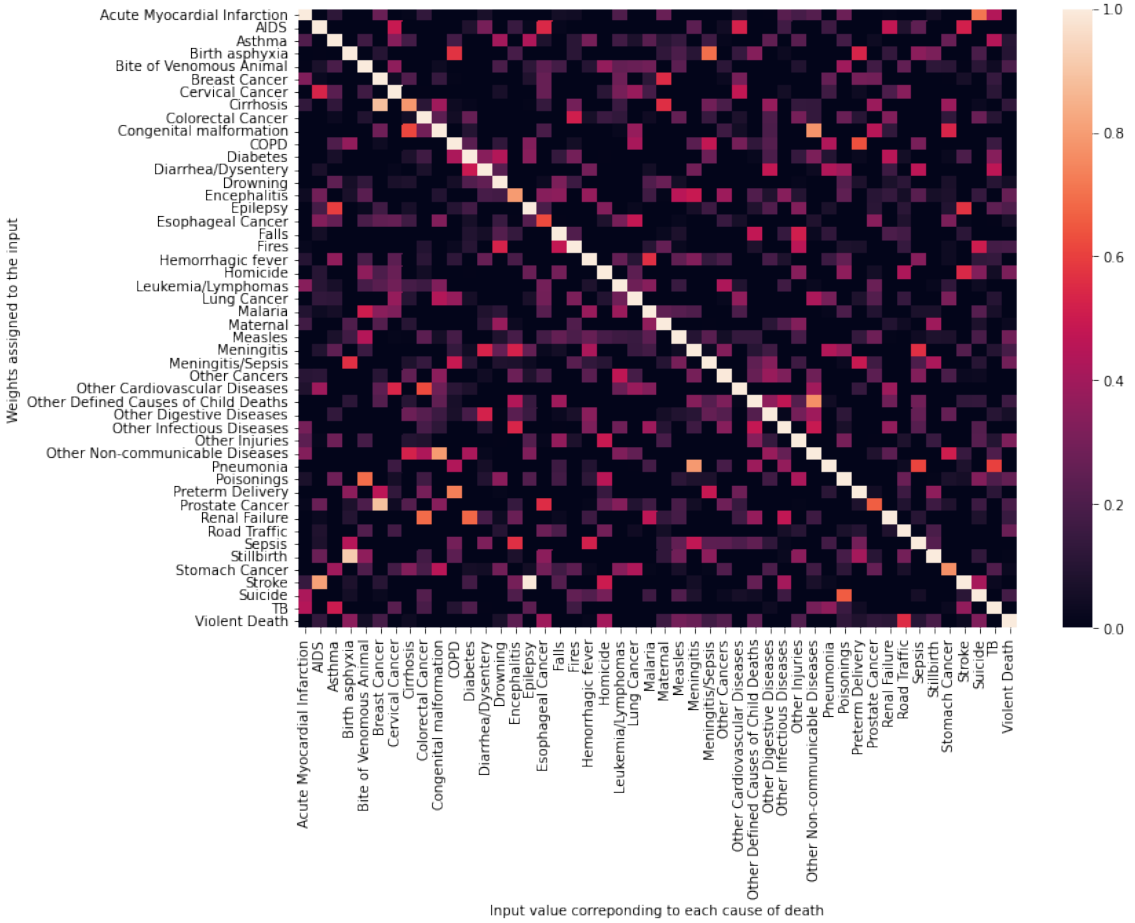

Figure 2.5.3: Heatmap of the weights learned in the last layer of the ensemble model of Figure 2.4.2 by logistic regression for the input given by the BERT model scaled between 0 and 1. The y axis are the weights learnt for the final output while the y axis indicates to which of the BERT outputs (i.e. value given by BERT for a particular CoD) corresponds that weight.

As a conclusion, first, if we compare the predictive capability of the CQ, we found that XGBoost is superior to InterVA for every age group. On its side, simple and imperfect though the OR might seem, the OR has proven to convey competitive predictive capabilities and is superior to InterVA. Finally, when the dual input is available, marginal improvements are attained over the XGBoost, though significant improvements with respect to InterVA. In any case, the OR is gaining importance over the CQ in the ensemble approach.

2.6 Discussion

Through this work we have applied some of the state-of-the-art NLP techniques to extract valuable information from the OR in order to improve the performance of the InterVA model from the WHO 2016 instrument, which only uses the CQs. We have compared the InterVA performance with a XGBoost model per each age group and we have also fine-tuned three different transformer-based models for the OR treatment. In addition. we have proposed an ensemble model to take both the OR and the CQs as input and to ascertain the cause of death using the InterVA and the XGBoost models for the CQs, a BERT-based transformer for the OR and a logistic regression for the final output.

As for the RQ1 and the CQ model, we have compared InterVA with XGBoost and the latter provided a general increment in accuracy of around the 20-30%. For the RQ2 and thus, the OR, the transformer-based and pre-trained models have been fine-tuned and compared with an XGBoost. The results show a marginal increment in performance in favor to the BERT model compared with other two transformer-based models except for the adult age group in which XGBoost resulted to have similar performance. A well known drawback of the CQ is the complexity of the interview since there are hundreds of questions arranged in a decision tree structure and also the lengthy duration of the interview. The OR is shorter in duration and results less complex for the interviewer and the interviewee and what is more, without being detrimental to the prediction accuracy compared with the widely accepted InterVA approach. Besides, as further version of the questionnaires arise, the approaches based on CQs must be updated, while ascertaining the CoD from the OR would have as an added advantage, the versatility.

Regarding the RQ3, the best ensemble model has achieved an accuracy of 56.20% for all the age groups. That result implies an improvement of the 9% in accuracy compared with just having the open response in the input, proving that when both inputs are combined the performance of the system can be improved. The same happens for the RQ4, when adding the open response to the InterVA method, the improvement is of almost a 20% compared to just using the CQs.

Finally, as stated in RQ5, the interpretation of the weights of the resulting ensemble model with InterVA for the CQs shows that the ensemble model does not rely too much in the InterVA while it fully relies in the BERT model. This conclusion could be expected, as the BERT model has a much larger accuracy.

3. Conclusions

This chapter presents the final conclusions of this research, as well as the scientific contributions and the future work.

3.1 Concluding remarks

This work has focused on verbal autopsies, specifically, on predicting the cause of death assigned by experts taking into account the answers collected from the questionnaire. For this purpose, a wide variety of methods have been employed for both closed questions and the open response. For the closed questions, a series of experiments have been carried out using the InterVA algorithm, which is endorsed by WHO in The WHO 2016 tool, and the XGBoost classifier. For the open response, on the other hand, several tests have been performed with different models based on BERT. Finally, an *ensemble* model has been proposed that combines the models mentioned above to make a final prediction.

Regarding the proposed **objectives** in Section 1.3, I consider that all the objectives proposed initially have been fulfilled:

1. An ensemble model has been created in order to combine the information provided by both the open response and the closed questions. This ensemble model has been able to improve the performance of using just both inputs independently.
2. The InterVA algorithm has been included in the analysis and extended with the integration of the open response, resulting in a remarkable increase in performance.
3. Many teachniques learnt in the master's degree have been applied to the analaysis, specially the transformer-based models such as BERT that have enabled a great

performance in cause of death predicition.

I **conclude** that my approach was able to improve the performance of the InterVA model adding the processing of the OR, resulting in a remarkable increase in performance and, moreover, I have found a better alternative to the InterVA for the prediction of the cause of death with the CQs.

Even though the InterVA model has been improved, the problem is not solved yet, as the performance can be far improved. However, it is hard to achieve an acceptable performance for this task as we believe that more data should be released. This is a complex multi-class classification task and the models developed in this work, specially BERT, would benefit from a data size increase. In addition, as seen not only in this work but also in previous works, there is evidence of the poor performance achieved by the WHO 2016 algorithms in the cause of death prediction task and we think that the WHO should reconsider the current approaches and take others into account in future versions.

Taking into account the reduced number of instances provided by the dataset and the high number of classes to predict, the final *ensemble* model achieves a good result with an accuracy (i.e. hit rate on the test set) of 56.20%. This result implies an increase of 6.0% in accuracy compared to using only the closed questions or a 9.0% using only the open response. In addition, all the proposed models achieve a much better performance than InterVA, which obtains an accuracy of 27.73% for adults. Note that the accuracy reported in this work is at the individual verbal autopsy level and not at the population level as presented in the articles of the WHO-endorsed algorithms, which is why they report a higher performance.

Therefore, this work has proposed a new line of work showing the benefit of extending the analysis of verbal autopsies by integrating the open response and exploring state-of-the-art models. I want to point out the importance of using pre-trained models to assign cause of death using text due to the limitations of this dataset. As an added conclusion, i was able to improve the performance of InterVA by adding the information provided by the open response.

This work has meant the immersion into a world within natural language processing that is gaining a lot of strength, such as pre-trained and transformer-based models, making me much more qualified than before starting it. It has also contributed to the development of a web prototype for the analysis of the causes of death, being able to test a wide variety of methods and thus allowing the analysis of the results obtained. This contribution is not part of the initially proposed objectives but it has been developed as a side contribution for the ease of the experimentation part. This prototype is explained in Appendix A.

The master's degree in Computational Engineering and Intelligent Systems of the UPV/EHU has played a relevant role in obtaining the necessary knowledge to carry out this work. Throughout this master's degree I have learned to program in the R programming language and to use R Studio, which has been vital to carry out this work as InterVA is included in an R package. In the subjects of Data Mining and Data Exploration and Analysis I have

learned exploratory techniques that have influenced the paths I have followed throughout this study. Finally, the subject *Deep Learning* taught by the IXA research group, mainly, by Eneko Agirre, its principal researcher, has been of much importance given that the subject contained practical laboratories that I have used to train models based on *transformers* as can be seen in the results provided in this work.

One of the most positive outcomes of this work has been my collaboration with the WHO. During this study, we collaborated with a very influential physician on the challenge of verbal autopsies. In the summer I worked with him on the study of public opinion on public health issues and during this year he has been supporting and guiding my work. In addition, the physician has confidence in the methods and conclusions drawn from my work and he has offered me the opportunity to continue contributing to this work.

3.2 Scientific contributions

1. Related to the closed questions we carried out a work that has resulted in a publication in an international double-blind peer-reviewed conference [Cejudo et al., 2021]:

   Authors: Ander Cejudo, Owen Trigueros, Alicia Pérez, Arantza Casillas, Daniel Cobos Title: Verbal Autopsy: first steps towards questionnaire reduction
   Ref. Book: Lecture Notes in Computer Science. Computational Linguistics and Intelligent Text Processing
   Fecha: 2021
   Publisher: Spring-Verlag

2. Additionally, connected with the classification methods learned in the master and applied in this document arose the collaboration between Ixa and UNED-LSI research groups contribute at the CLPsych2022 *shared task* on suicide risk detection and has resulted in a publication that is currently accepted [Fabregat et al., 2022]:

   Authors: Gildo Fabregat, Ander Cejudo, Juan Martinez-Romo, Alicia Pérez, Lourdes Araujo, Nuria Lebeña, Maite Oronoz, and Arantza Casillas
   Title: Approximate nearest neighbour extraction techniques and neural networks for suicide risk prediction in the CLPsych 2022 shared task.
   Ref. Book: In Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change.
   Date: 2022
   Publisher: Association for Computational Linguistics

3.3 Future work

This study has left many open lines, such as trying to add the open response analysis to the rest of the algorithms endorsed by the WHO in search of a remarkable performance

increase. It also gives rise to the possibility of combining open response and closed questions in multiple ways, trying to maximize as much as possible the predictive capacity of the cause of death of the final system.

Another contribution to make is to unify the whole ensemble model system, offering it in a single package as a single flow, since currently the software is very segmented as part of the implementation is in R and part in Python.

Finally, different techniques can be applied to the open response, such as summarizing the text, obtaining entities or even creating an information retrieval system to try to associate a new open response to the cause of death with the one that has more open responses in common.

A. Appendix: Web prototype

In this chapter, the created web prototype as a secondary and initially not planned contribution is presented along with some of the web prototype's functionalities, which are described and shown through screenshots. During the development of this TFM, the web prototype has been implemented in parallel as an opportunity to obtain an additional publication. In general terms, this prototype implements many of the conclusions obtained in the first publication [Cejudo et al., 2021] as well as others that have been obtained during the course of this project.

This prototype can be used by two different user profiles: researcher and physician. The researcher profile can use the dataset provided in this prototype and vary the methods and parameters provided and investigate the implementation of new methods from these. The physician profile can upload the collected verbal autopsies, use the methods that are selected by default and rely on the provided results for their final prediction. The results are displayed in a series of tables, showing the probability of belonging to the different causes of death in order to help them in their final decision as shown in Figure A.5.

The table A.1 shows the main functionalities offered by the application for each of the different modules implemented. Then, a series of screenshots of the application are shown.

In Figure A.1, the user interface when entering to the prototype is shown. In this screen, the user can change the input for the analysis: open response, closed questions, closed questions reduced and the whole questionnaire. Depending on the input selected, the user will be given the chance to either upload a file, as shown in the screenshot with verbal autopsies, or fill in some on the questionnaire fields. Finally, the user can press the button of the "Run" section to predict the cause of death for each of the input verbal autopsies.

Module	Features
Input	- Analysis with CQ - Analysis with OR - Analysis with CQr - Analysis with OR+CQ - Upload csv files - OR and CQr manual data entry
Configuration	- Classifier selection: LR, Naive Bayes, XGBoost and InterVA - Text processing selection: Embeddings (50, 100, 200, 300) and TF-IDF - Data source selection: train, test or manual - Questionnaire reduction for CQs
Output	- Precision at k for labeled data (k->1 = accuracy) - Selection of analyzed verbal autopsies - Viewer of selected verbal autopsy responses - Viewer of probabilities of membership for each cause of death - Indicator for eliminated questionnaire questions - Entity recognizer for open response - Entity recognizer for summarized open response
General	- Help boxes in all options - Animation when obtaining results - Saving of models - Loading animation

Table A.1: Table of functionalities of the web prototype combining several inputs: open response (OR), closed response questions (CQ) and a reduced set of closed response questions (CQr).

If the user selects closed questions as the only or as one of the inputs, two more options are displayed as in A.2: select the questionnaire reduction method and the number of questions to drop. According to Cejudo et al. [2021], some questionnaire reduction techniques were implemented and thus, added to this web prototype.

For the closed questions input, as shown in A.3, if ”advanced mode” is selected, the user will be given the chance to select between several classifiers: XGBoost, logistic regression, naive bayes and InterVA. If the input included the open response, additionally it would be possible to choose between the text representation method (i.e. TF-IDF and word-embeddings).

After pressing the button and when the loading step has finished, the screen is shifted and the output section is displayed as in A.4. In this screen, if in the input the actual causes of death are provided, the precision at k evaluation metric is provided for each of the entry age groups varying the k value. Precision at k is similar to the accuracy score metric but in this case the prediction is considered correct when the most likely k causes of death include the actual cause of death. For k equal to 1 the metric is the same as accuracy. Moreover, a verbal autopsy picker is offered where the user can choose between one of the verbal autopsies provided in the input in order to analyze it.

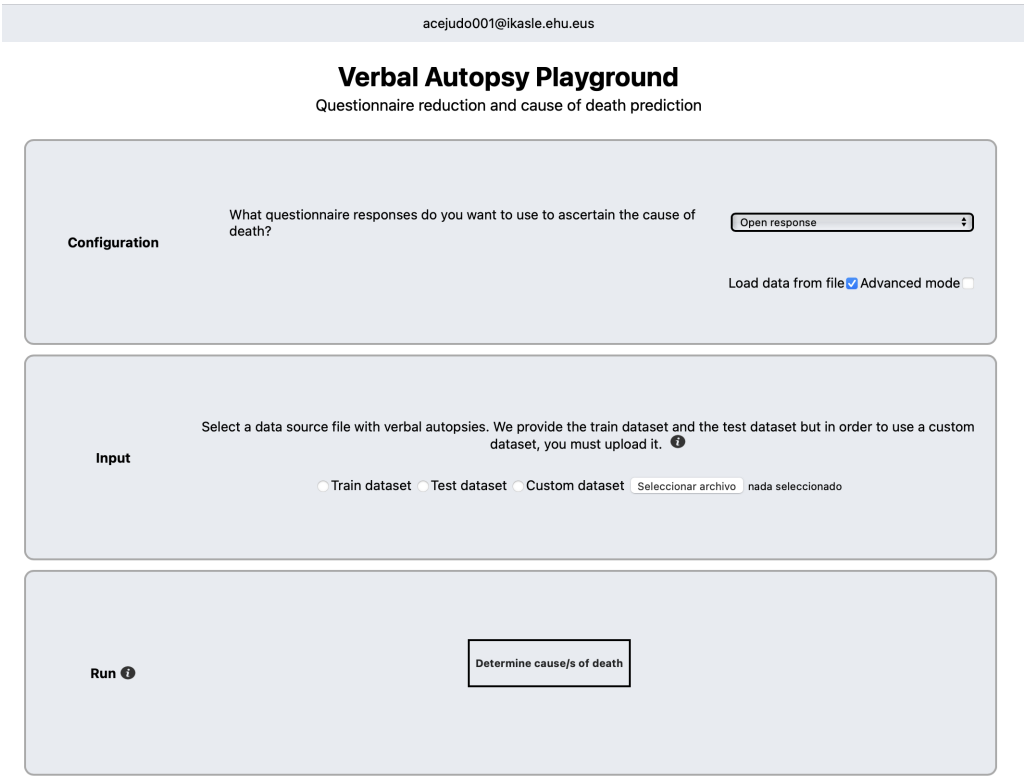

Figure A.1: Screenshot of the application at login.

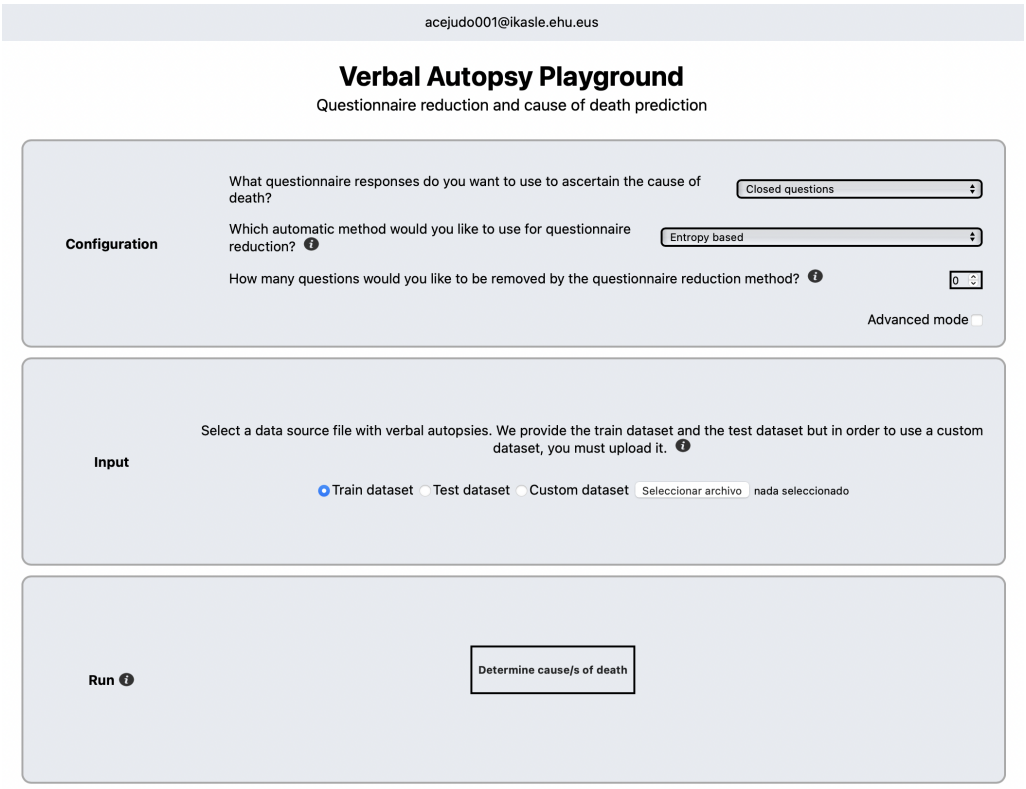

Figure A.2: Screenshot of the application when selecting closed questions for analysis.

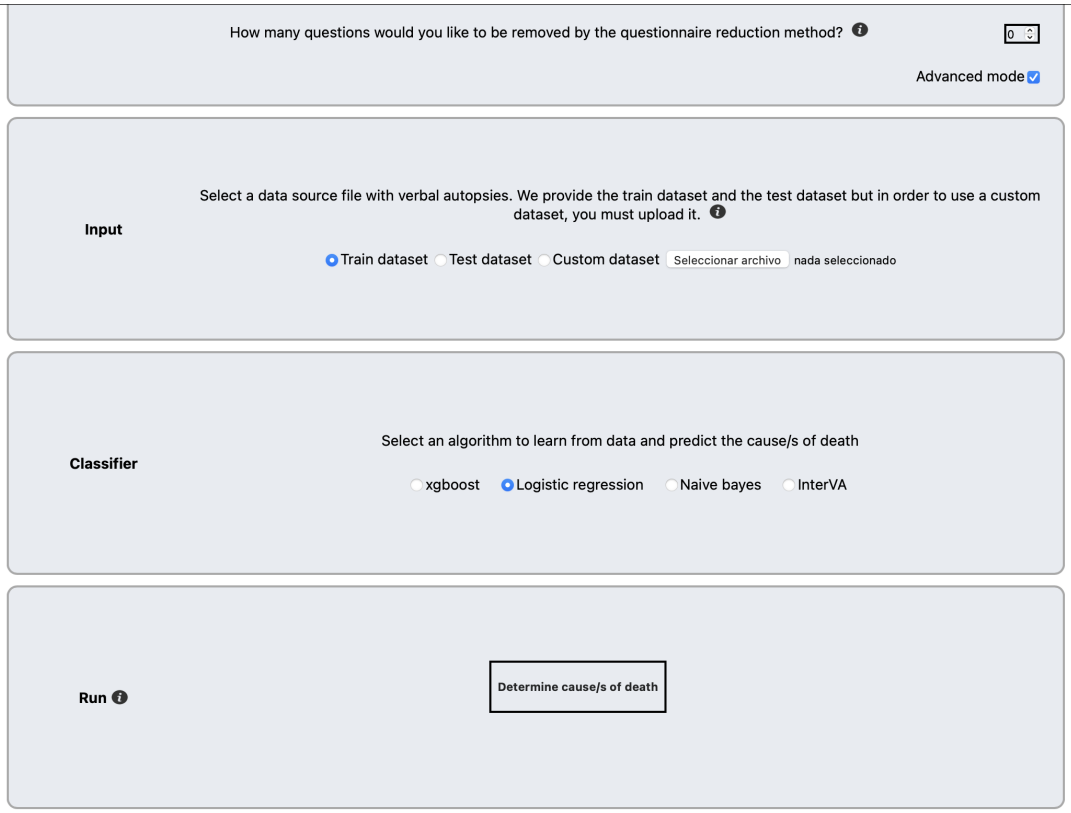

Figure A.3: Screenshot of the application with the classifiers to choose after selecting the advanced mode.

After choosing a verbal autopsy, the user is provided with a cause of death probability ranking as in Figure A.5. In this table, the user can see which cause of death is the most probable and the probability assigned by the chosen classifier. Additionally, as the closed questions are present in the input, when selecting a verbal autopsy all the data related to this verbal autopsy is displayed in the last table and the column that where not removed by the chosen questionnaire reduction method are colored in green while the removed ones are colored in red.

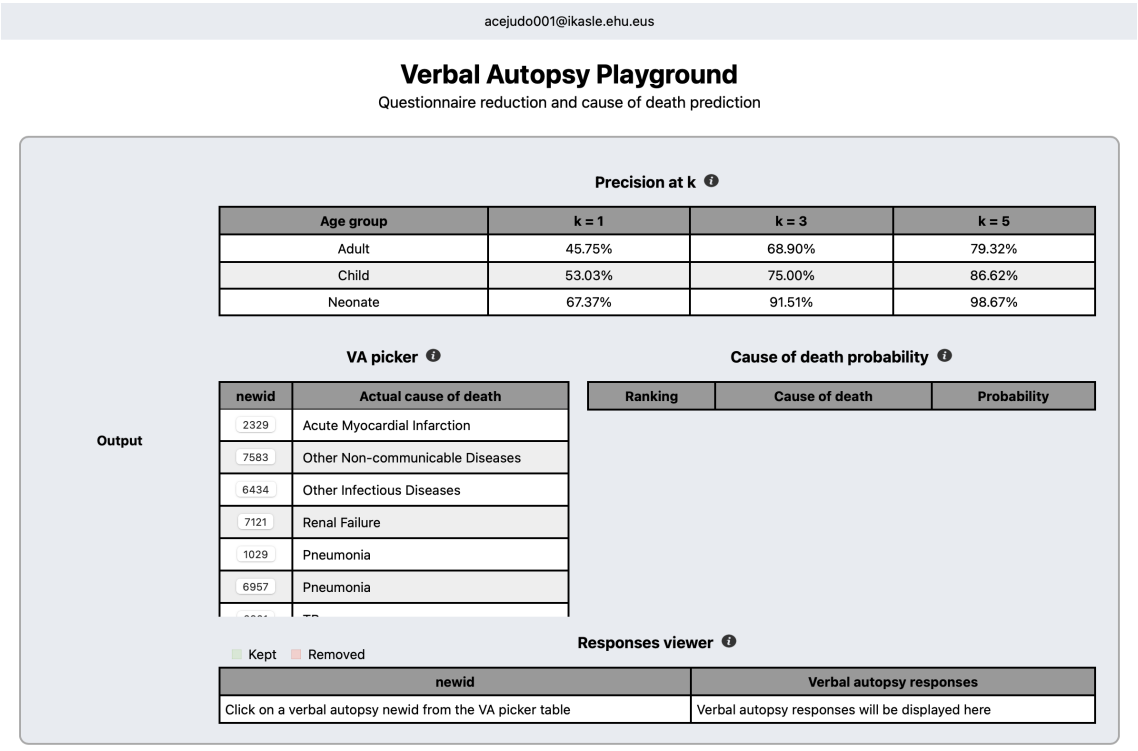

Figure A.4: Screenshot of the application providing the results.

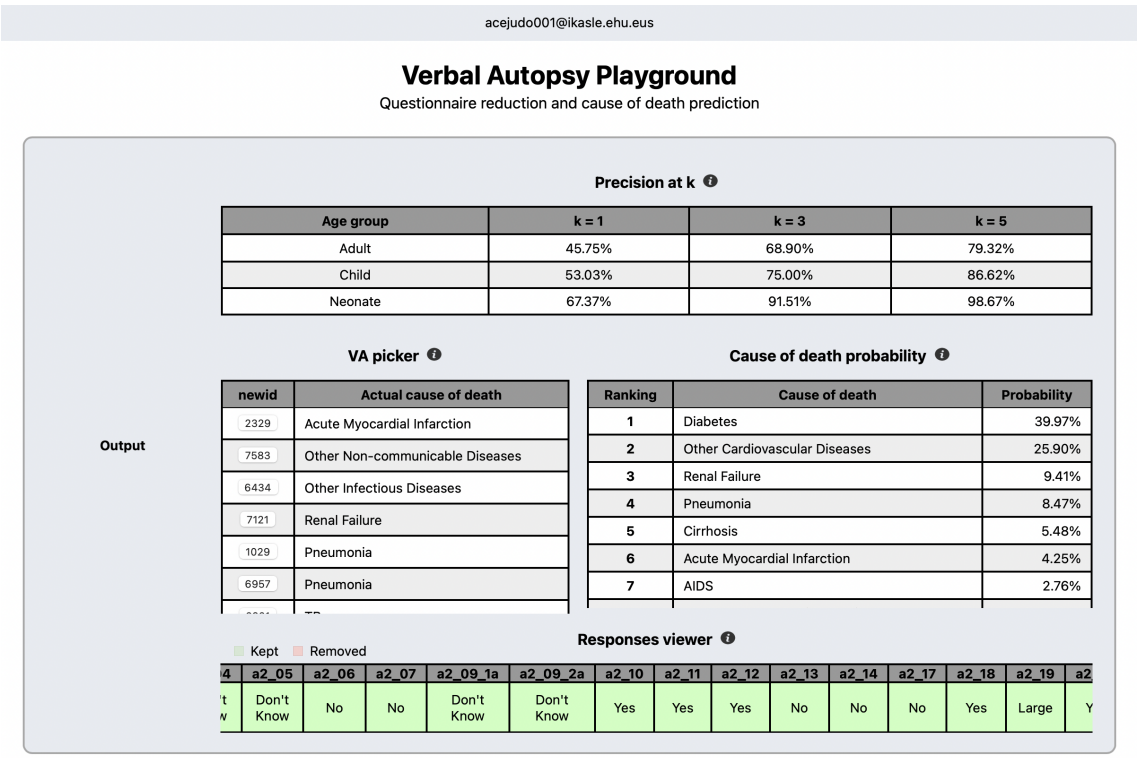

Figure A.5: Screenshot of the application when selecting a verbal autopsy in the results.

Bibliography

Alsentzer, E., J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott
2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Baqui, A. H., R. E. Black, S. Arifeen, K. Hill, S. Mitra, and A. Al Sabir
1998. Causes of childhood deaths in Bangladesh: results of a nationwide verbal autopsy study. *Bulletin of the World Health Organization*, 76(2):161.

Blanco, A., A. Pérez, A. Casillas, and D. Cobos
2020. Extracting cause of death from verbal autopsy with deep learning interpretable methods. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1315–1325.

Byass, P., D. Chandramohan, S. J. Clark, L. D'ambruoso, E. Fottrell, W. J. Graham, A. J. Herbst, A. Hodgson, S. Hounton, K. Kahn, et al.
2012. Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool. *Global health action*, 5(1):19281.

Byass, P., L. Hussain-Alkhateeb, L. D'Ambruoso, S. Clark, J. Davies, E. Fottrell, J. Bird, C. Kabudula, S. Tollman, K. Kahn, et al.
2019. An integrated approach to processing WHO-2016 verbal autopsy data: the InterVA-5 model. *BMC medicine*, 17(1):1–12.

Cejudo, A., O. Trigueros, A. Pérez, A. Casillas, and D. Cobos
2021. Verbal autopsy: first steps towards questionnaire reduction. *Lecture Notes in Computer Science. Computational Linguistics and Intelligent Text Processing*.

Chandramohan, D., E. Fottrell, J. Leitao, E. Nichols, S. J. Clark, C. Alsokhn, D. Co-

bos Munoz, C. AbouZahr, A. Di Pasquale, R. Mswia, et al.
2021. Estimating causes of death where there is no medical certification: evolution and state of the art of verbal autopsy. *Global Health Action*, 14(sup1):1982486.

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al.
2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Clark, S. J., T. Mccormick, Z. Li, and J. Wakefield
2013. Insilicova: a method to automate cause of death assignment for verbal autopsy.

D'Ambruoso, L.
2013. Worldwide, 65% of deaths go uncounted – here's how to change that. https://theconversation.com/worldwide-65-of-deaths-go-uncounted-heres-how-to-change-that-46644. (Accessed on 06/27/2022).

Danso, S., E. Atwell, and O. Johnson
2014. A comparative study of machine learning methods for verbal autopsy text classification. *arXiv preprint arXiv:1402.4380*.

Danso, S., E. Atwell, O. Johnson, A. H, S. Soremekun, K. Edmond, C. Hurt, L. Hurt, C. Zandoh, C. Tawiah, J. Fenty, S. Amenga-Etego, S. Owusu-Agyei, and B. Kirkwood
2013. A semantically annotated verbal autopsy corpus for automatic analysis of cause of death. *ICAME Journal of the International Computer Archive of Modern and Medieval English*, 37:37–69.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova
2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

D'Souza, J.
2018. An introduction to bag-of-words in nlp | medium. https://medium.com/greyatom/an-introduction-to-bag-of-words-in-nlp-ac967d43b428. (Accessed on 04/30/2022).

Fabregat, G., A. Cejudo, J. Martinez-Romo, A. Pérez, L. Araujo, N. Lebeña, M. Oronoz, and A. Casillas
2022. Verbal autopsy: first steps towards questionnaire reduction. *In Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change.*

Fantahun, M., E. Fottrell, Y. Berhane, S. Wall, U. Högberg, and P. Byass
2006. Assessing a new approach to verbal autopsy interpretation in a rural Ethiopian community: the interva model. *Bulletin of the World Health Organization*, 84(3):204–210.

Flaxman, A. D., J. C. Joseph, C. J. Murray, I. D. Riley, and A. D. Lopez
2018a. Performance of InSilicoVA for assigning causes of death to verbal autopsies: multisite validation study using clinical diagnostic gold standards. *BMC medicine*, 16(1):1–11.

Flaxman, A. D., J. C. Joseph, C. J. L. Murray, I. D. Riley, and A. D. Lopez
2018b. Performance of InSilicoVA for assigning causes of death to verbal autopsies: multisite validation study using clinical diagnostic gold standards. *BMC Medicine*, 16(1):56.

Ganapathy, S. S., K. Y. Yi, M. A. Omar, M. F. M. Anuar, C. Jeevananthan, and C. Rao
2017. Validation of verbal autopsy: determination of cause of deaths in Malaysia 2013. *BMC public health*, 17(1):1–8.

Garenne, M. and V. Fauveau
2006. Potential and limits of verbal autopsies. *Bulletin of the World Health Organization*, 84:164–164.

Hripcsak, G. and A. S. Rothschild
2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Jha, P., V. Gajalakshmi, P. C. Gupta, R. Kumar, P. Mony, N. Dhingra, R. Peto, and R. P. S. Collaborators
2006. Prospective study of one million deaths in India: rationale, design, and validation results. *PLoS medicine*, 3(2):e18.

Kleinbaum, D. G., K. Dietz, M. Gail, M. Klein, and M. Klein
2002. *Logistic regression*. Springer.

Li, Z. R., T. H. McComick, and S. J. Clark
2020. Using bayesian latent Gaussian graphical models to infer symptom associations in verbal autopsies. *Bayesian analysis*, 15(3):781.

Li, Z. R., T. H. McCormick, and S. J. Clark
2017. The openva toolkit for verbal autopsies. https://zehangli.com/openVA/openVA-vignette_2017.pdf (Accedido en 07/06/2020).

Li, Z. R., T. H. McCormick, and S. J. Clark
2018. Verbal autopsy analysis using openva.

McCormick, T. H., Z. R. Li, C. Calvert, A. C. Crampin, K. Kahn, and S. J. Clark
2016. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049.

Mikolov, T., M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur
2010. Recurrent neural network based language model. In *Interspeech*, volume 2, Pp. 1045–1048. Makuhari.

Moran, K. R., E. L. Turner, D. Dunson, and A. H. Herring
2021. Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(3):532–557.

Murray, C. J., A. D. Lopez, R. Black, R. Ahuja, S. M. Ali, A. Baqui, L. Dandona, E. Dantzer, V. Das, U. Dhingra, A. Dutta, W. Fawzi, A. D. Flaxman, S. Gómez, B. Hernández, R. Joshi, H. Kalter, A. Kumar, V. Kumar, R. Lozano, M. Lucero, S. Mehta, B. Neal, S. L. Ohno, R. Prasad, D. Praveen, Z. Premji, D. Ramírez-Villalobos, H. Remolador, I. Riley, M. Romero, M. Said, D. Sanvictores, S. Sazawal, and V. Tallo
2011. Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Population Health Metrics*, 9(1):27.

Nichols, E. K., P. Byass, D. Chandramohan, S. J. Clark, A. D. Flaxman, R. Jakob, J. Leitao, N. Maire, C. Rao, I. Riley, et al.
2018a. The who 2016 verbal autopsy instrument: An international standard suitable for automated analysis by interva, insilicova, and tariff 2.0. *PLoS medicine*, 15(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5761828/ (Accedido en 07/06/2020).

Nichols, E. K., P. Byass, D. Chandramohan, S. J. Clark, A. D. Flaxman, R. Jakob, J. Leitao, N. Maire, C. Rao, I. Riley, et al.
2018b. The who 2016 verbal autopsy instrument: An international standard suitable for automated analysis by interva, insilicova, and tariff 2.0. *PLoS medicine*, 15(1):e1002486.

Noble, W. S.
2006. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567.

Powers, D. M. W.
2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Ramos and Juan
2003. Using TF-IDF to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, Pp. 29–48. Citeseer.

Reynolds, D. A.
2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663).

Rish and Irina
2001. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, Pp. 41–46.

Serina, P., I. Riley, A. Stewart, S. L. James, A. D. Flaxman, R. Lozano, B. Hernandez, M. D. Mooney, R. Luning, R. Black, R. Ahuja, N. Alam, S. S. Alam, S. M. Ali, C. Atkinson, A. H. Baqui, H. R. Chowdhury, L. Dandona, R. Dandona, E. Dantzer, G. L. Darmstadt, V. Das, U. Dhingra, A. Dutta, W. Fawzi, M. Freeman, S. Gomez, H. N. Gouda, R. Joshi, H. D. Kalter, A. Kumar, V. Kumar, M. Lucero, S. Maraga, S. Mehta, B. Neal, S. L. Ohno, D. Phillips, K. Pierce, R. Prasad, D. Praveen, Z. Premji, D. Ramirez-Villalobos, P. Rarau, H. Remolador, M. Romero, M. Said, D. Sanvictores, S. Sazawal, P. K. Streatfield, V. Tallo, A. Vadhatpour, M. Vano, C. J. L. Murray, and A. D. Lopez
2015a. Improving performance of the Tariff Method for assigning causes of death to verbal autopsies. *BMC Medicine*, 13(1):291.

Serina, P., I. Riley, A. Stewart, S. L. James, A. D. Flaxman, R. Lozano, B. Hernandez, M. D. Mooney, R. Luning, R. Black, et al.
2015b. Improving performance of the tariff method for assigning causes of death to verbal autopsies. *BMC medicine*, 13(1):1–13.

Setel, P. W., D. R. Whiting, Y. Hemed, D. Chandramohan, L. J. Wolfson, K. Alberti, and A. D. Lopez
2006. Validity of verbal autopsy procedures for determining cause of death in Tanzania. *Tropical Medicine & International Health*, 11(5):681–696.

Todd, J., A. De Francisco, T. O'dempsey, and B. Greenwood
1994. The limitations of verbal autopsy in a malaria-endemic region. *Annals of tropical paediatrics*, 14(1):31–36.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin
2017. Attention is all you need.

WHO
2022. Webinar upon the release of 2022 who verbal autopsy instrument. https://www.who.int/news-room/events/detail/2022/04/06/default-calendar/webinar-upon-the-release-of-2022-who-verbal-autopsy-instrument. (Accessed on 05/31/2022).

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush
2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Pp. 38–45, Online. Association for Computational Linguistics.

Yan, Z., S. Jeblee, and G. Hirst
2019. Can character embeddings improve cause-of-death classification for verbal autopsy narratives? In *Proceedings of the 18th BioNLP Workshop and Shared Task*, Pp. 234–239.

Yang, G., C. Rao, J. Ma, L. Wang, X. Wan, G. Dubrovsky, and A. D. Lopez
2006. Validation of verbal autopsy procedures for adult deaths in China. *International journal of epidemiology*, 35(3):741–748.