Extracting Cause of Death from Verbal Autopsy with Deep Learning interpretable methods

Alberto Blanco Alicia Pérez, Arantza Casillas, Daniel Cobos

***Abstract*—The international standard to ascertain the cause of death is medical certification. However, in many low and middle-income countries, the majority of deaths occur outside of health facilities. In these cases, Verbal Autopsy (VA), the narrative provided by a family member or friend together with a questionnaire is designed by the World Health Organization as the main information source. Until now technology allowed us to automatically analyze the responses of the VA questionnaire with the narrative captured by the interviewer excluded. Our work addresses this gap by developing a set of models for automatic Cause of Death (CoD) ascertainment in VAs with a focus on the textual information. Empirical results show that the open response conveys valuable information towards the ascertainment of the Cause of Death, and the combination of the closed-ended questions and the open response lead to the best results. Model interpretation capabilities position the Deep Learning models as the most encouraging choice.**

***Index Terms*—Verbal Autopsy, Decision support system, Recurrent Neural Network, Word Attention**

I. Introduction

RELIABLE information on the number and causes of death is the cornerstone for health planning to inform resource allocation decision-making processes and for policy development. Civil registration and vital statistics (CRVS) in well functioning systems convey valuable information in a country [1], [2] and provide a reliable, continuous and universal flow of information about vital events that can be disaggregated to produce estimates at the local level. According to the World Health Organization (WHO), two-thirds (38 million) of 56 million annual deaths are still not registered [3], [4] and most of them occur in the poorest parts of the World. Where they are available, Causes of Death (CoD) are often biased since they derive, mainly, from hospital statistics in non-rural settings and, therefore, are frequently unrepresentative of the experience of the general population. This situation yields mislead understanding of health priorities in the country. This is particularly troublesome now as the demographic and health transitions are rapidly changing the cause-specific mortality fractions at population level in such countries [5].

Verbal Autopsy (VA) is an indirect method of ascertaining causes of death from information collected by interviewing families and caretakers. It is subscribed by the World Health Organization [6] and gathers both a free narrative and a standard questionnaire with closed-ended questions that would convey signs, symptoms and circumstances preceding death. Traditionally, the information collected in the VA has been interpreted by health experts to generate a probable underlying cause of death (CoD).

A number of advances in the VA questionnaire design [7], [8], in data capture on mobile devices [6], and in the use of computer algorithms for determining and coding the CoD [9], [10], [11], [12], [13] now make mobile, community-based VA an increasingly available and effective proxy means to emulate physician-certified cause of death. Countries using the WHO 2016 VA questionnaire are now able to analyse the responses using multiple computer algorithms as well as physician coding [6].

Currently, the computer algorithms only use part of the information contained in the questionnaire. First, the algorithms only use "Yes" answers usually referring to the presence of a sign or symptom. However, they do not include in the analysis "No" answers. This information could be useful to identify or exclude causes of death and to refine the final CoD. Also, the algorithms do not use the narrative about the events occuring prior to death in their analysis. There is an unused wealth of information in the narrative not only about the CoD but also about health-seeking behaviour and health system responsiveness.

We hypothesized that the performance of existing VA analytical methods can be improved by combining the information in the open narrative with that from the closed questions. Indeed, the contribution of this work is twofold: i) we provide experimental evidences of the synergistic information contained in both the open narrative and questionnaires in the ascertainment of the CoD, paying care to interpretability; ii) we made available the tools developed to promote reproducibility.

II. Materials

In this work we are dealing with the so-called "Gold Standard VA database" [14]. The VA data contain verbal autopsies conducted in six sites from four countries (India, Mexico, Tanzania, and the Philippines) using a standardized VA questionnaire. The data were split in two randomly selected stratified sub-sets, respectively, for training (70%) and testing (30%). Table I shows the main characteristics of the dataset.

A. Blanco, A. Casillas and A.Perez are with the HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Manuel Lardizabal 1, 20080 Donostia, Spain e-mail: alberto.blanco@ehu.eus, arantza.casillas@ehu.eus, alicia.perez@ehu.eus and D. Cobos is with the Swiss Tropical and Public Health Institute, Basel, Switzerland and University of Basel, Basel, Switzerland, e-mail: daniel.cobos@swisstph.ch, (Corresponding author: Alberto Blanco.)

TABLE I
DESCRIPTION OF THE VA DATASET. THE INPUT INFORMATION SHOWS A DUAL NATURE WITH 1) OPEN-RESPONSE IN FREE TEXT; 2) CLOSED-ENDED QUESTIONNAIRE. THE OUTPUT IS THE SET OF CODS.

	Information	Age-range			Total
inp	**Open response**	**Adult**	**Child**	**Neonate**	
train	Total VAs	5,489	1,444	1,837	8,770
	Informative VAs	3,394	938	876	5,209
	Vocabulary	8,140	3,598	3,327	9,644
	N. Words $\pm$ SD	92 $\pm$ 75	78 $\pm$ 76	88 $\pm$ 74	89 $\pm$ 75
test	Total VAs	2,352	620	788	3,760
	Informative VAs	1,455	403	376	2,233
	OOV tokens	1,503	569	550	1,699
	N. Words $\pm$ SD	93 $\pm$ 76	73 $\pm$ 68	85 $\pm$ 72	87 $\pm$ 75
inp	**Closed questions**	**Adult**	**Child**	**Neonate**	**Total**
tr/te	Total Features	142	87	110	303
	Numeric Feat.	3	4	4	4
	Categoric Feat.	139	83	104	299
out	**CoD**	**Adult**	**Child**	**Neonate**	**Total**
train	Total Classes	5,489	1,444	1,837	8,770
	Distinct. Classes	34	21	6	48
test	Total Classes	2,352	620	788	3,760
	Distinct. Classes	34	21	6	48

There are 8.770 samples in the training set. The main limitation of the data is that the textual **input** was not always correctly completed. 40% of samples are considered non-informative (here referred to as stop-texts), meaning that the registered open response does not refer to the CoD e.g. "*Respondent has nothing to add*". The standard deviation (SD) of the length (N. Words) of the open response (i.e. the length of the narratives as the count of words) is high and one of the main reasons is, again, the presence of non-informative samples, like "*no comment*". However, some short narratives are really meaningful e.g. "*baby died because of fever*". There are also long open responses exceeding a thousand words. The number of Out-of-vocabulary (OOV) words in the test set are high due to the lexical variability in the clinical field, where the vocabularies (i.e. distinct words) tend to be broad. The open responses are, in general, short texts (with a mean length of ≈ 87 words per response, as shown in Table I) and, in conjunction with the high OOV words, lead to inputs with few known words. This is an inherent challenge of the task.

So far we have focused on the quantitative and qualitative description of the open response (the input to our system). In what follows we describe the **output** which, within the supervised machine learning framework, is the variable that we want to predict, often referred to as 'class' or 'label' and, in this case, the CoD. The set of classes comprise 48 different CoDs (6 for neonates, 21 for children and 34 for adults) as is in the original dataset [15], shown in Table I and detailed in Appendix A. Overall, the dominant CoD is *Pneumonia* (10.0% of all the CoDs), followed by *Stillbirth* $(6,3\%)$ and *Stroke* $(5,7\%)$.

The presence of some diseases is very high, while other diseases are infrequent. This leads to skewed class distribution (measured in Figures 8, 9 and 10). While humans might be well trained on both frequent and infrequent CoDs, inference algorithms tend not to learn what has not been seen frequently in the training data and often fail to predict infrequent classes [16].

We have mentioned several aspects of this dataset that are challenging for machine learning algorithms (e.g. uninformative inputs, OOVs, class imbalance). Next, we assessed, quantitatively, the **difficulty** of the open response in the framework of text classification. To this end, we turned to standard NLP metrics [17] related to frequencies of words and n-grams (sequences of n words together with class distribution and imbalance, as shown in Table II. In [18] the standard metrics were combined heuristically and the term "Difficulty" was proposed to provide an idea of the task's difficulty. A brief description of each measure contextualized for our use-case is given here: i) **Dataset size:** the number of available VAs from the 'Gold standard VA database' discarding those VAs with non-informative narratives. ii) **Vocab Size:** the number of unique words through all the narratives. iii) **Number of Classes:** the number of CoDs through all the VAs. iv) **Mean Items Per Class:** the mean number of VAs associated to each CoD. v) **Min. Items in a Class:** the number of VAs associated to the least frequent CoD (i.e. Esophageal Cancer). vi) **Average Sentence Length:** the average sentence length through all the narratives. vii) **Distinct Words / Total Words:** the ratio between the unique words and the total number of words as a measure of the diversity of the vocabulary. viii) **Class Diversity:** the relative frequency (support) of different CoDs. ix) **Max. Hellinger Similarity:** the Hellinger Distance is used to quantify the similarity between two probability distributions, in this case, the distribution of each pair of CoDs. x) **Mutual Information:** [19] relates two random variables (i.e. the 10 most frequent n-grams for each class and the class itself) and measures their dependence (roughly, how easy it is to discern which CoD belongs to the narrative given only the most frequent n-grams). xi) **Difficulty:** a heuristic combination of the aforementioned metrics.

Altogether, the quantitative assessment, given in Table II, regards as "very high" the difficulty of the open response classification task.

TABLE II
QUANTITATIVE ASSESSMENT OF THE DIFFICULTY OF THE TEXT CLASSIFICATION TASK.

Indicator	Value	Judgement
Dataset Size	7.442	-
Vocab Size	11.024	-
Number of Classes	48	-
Mean Items Per Class	155	-
Min. Items in a Class	7	EXTREMELY LOW
Average Sentence Length	470	-
Distinct Words / Total Words	0.0171	GOOD
Class Diversity	3.48	VERY HIGH
Max. Hellinger Similarity	0.801	HIGH
Mutual Information	0.992	GOOD
Difficulty	6.028	VERY HIGH

III. METHODS

A. Exploiting textual information

The Recurrent Neural Network [20] based model architecture consists of an embedding layer, a bidirectional recurrent layer [21] with Gated Recurrent Units (GRU) [22], max-pooling and average-pooling layers and a fully-connected layer with Softmax activation [23].

The embedding layer is a lookup table between word-tokens $(w_1, \ldots, w_T)$ and the dense vector representation $(\mathbf{v}_1, \ldots, \mathbf{v}_T)$ with $\mathbf{v}_t \in \mathbb{R}^d$, i.e. T is the length of the input sequence and d the dimension of the embedding vectors. Then, the dense representation of the sequence is fed to the bidirectional recurrent layer, from left to right (forward, $\rightarrow$) and from right to left (backward, $\leftarrow$), producing, at each time step (t), a pair of hidden states, $\overleftarrow{\mathbf{h}}_t \in \mathbb{R}^m$ and $\overrightarrow{\mathbf{h}}_t \in \mathbb{R}^m$, with m being the number of hidden units of the recurrent layer. These states will be concatenated into a single hidden state, $\mathbf{h}_t \in \mathbb{R}^{2m}$, as follows:

$$\mathbf{h}_t = [\overleftarrow{\mathbf{h}}_t \| \overrightarrow{\mathbf{h}}_t] \quad \text{with} \quad 1 \leq t \leq T$$

The hidden state computation is defined by (1), the representations of the GRU units where $\mathbf{x}_t$ is the input at time-step t and the bias terms are omitted.

$$\begin{aligned} \mathbf{z}_t &= \sigma(\mathbf{U}_z \mathbf{x}_t + \mathbf{W}_z \mathbf{h}_{t-1}) \\ \mathbf{r}_t &= \sigma(\mathbf{U}_r \mathbf{x}_t + \mathbf{W}_r \mathbf{h}_{t-1}) \\ \mathbf{c}_t &= \mathbf{tanh}(\mathbf{U}_z \mathbf{x}_t + \mathbf{W}_c(\mathbf{h}_{t-1} \odot \mathbf{r}_t) \\ \mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{c}_t + \mathbf{z}_t \odot \mathbf{h}_{t-1} \end{aligned} \tag{1}$$

The GRU hidden state computation relies on several vectors also known as controlling gates: update gate ($\mathbf{z}$) and reset gate ($\mathbf{r}$) along with an intermediate cell state ($\mathbf{c}$). The concatenated hidden state ($\mathbf{h}_t$) is next fed to the max-pooling and average-pooling layers. Each of them computes, respectively, the maximum and average value over time in order to get a compressed representation of the input sequence. These two compressed representations will be concatenated as $\mathbf{h} = [\mathbf{h}_{max} \| \mathbf{h}_{avg}] \in \mathbb{R}^{2T}$ and, next, fed into the final fully-connected layer to obtain the logits ($\mathbf{o} \in \mathbb{R}^K$) as in (2), where K is the number of classes, $\mathbf{W}$ is the weight matrix, and b is the bias term.

$$\mathbf{o} = \mathbf{W}\mathbf{h} + b \tag{2}$$

Finally, the logits are turned into K probabilities, $\mathbf{p} = (p_1, \cdots, p_K)$, by means of the $\mathrm{softmax}$ function shown in (3).

$$p_k = \frac{\exp(o_k)}{\sum_{j=1}^{K} \exp(o_j)} \quad \text{with} \quad 1 \leq k \leq K \tag{3}$$

In addition to the aforementioned Bi-GRU model, we built a variation in an attempt to reach the high-level information learned by the model and make it interpretable. To this end, we included an **attention layer** [24]. This layer captures the focus of the model and measures the responsibility of each token (within the sequence of tokens) to the probability estimated for each CoD. We shall refer to this variation as BiGru|Att. The attention mechanism not only enables us to interpret the Deep Learning process but also has an additional advantage: it provides an extra latent representation, an alternative to which pooling techniques provide; this is also fed into the final dense layer. Figure 1 shows the architecture of the model.

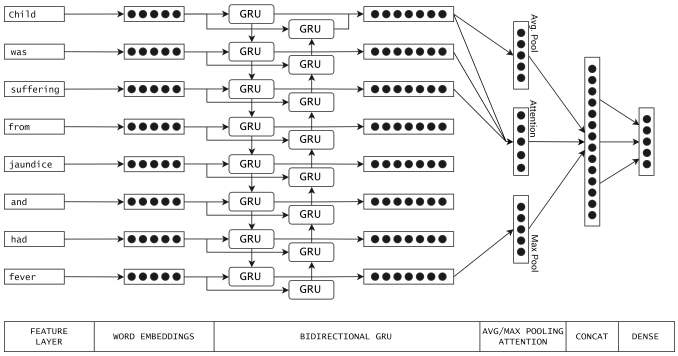

Fig. 1. Architecture of the BiGru|Att model: extends the architecture of the plain BiGru model by incorporating the attention layer.

B. Exploiting categorical information

So far, we have focused on recurrent neural networks (as BiGRU), well suited for sequential inputs, as is the case of the open response. Nevertheless, the input available in our task is dual (closed-ended questionnaire and open-text response). It is well-known that deep neural mechanisms struggle to learn effectively, from sparse categorical data. Hence, to handle the closed questions we chose an alternative to the raw feeding of categorical variables. The solution adopted was to embed these parameters (as we did for text). The embeddings enable us to transfer large sparse vectors into a low-dimensional dense space that preserves the relationships among categorical values i.e. for a categorical variable holding the 'site', the embedding vectors of similar places (such as geographically close places) would lie close in the embedding space.

The **categorical embeddings** were inferred together with the predictive model in an attempt to generate synergies. As a result of the training process, a d_i-dimensional embedding unit is created for each categorical variable. The dimension of the embedding layers e_i are hyperparameters, with theoretical bounds at $[1, m_i - 1)$. In practice, we chose the dimensions empirically, setting $\mathbf{e} = (e_1, \ldots, e_i, \ldots, e_n)$ with the following heuristic rule: $e_i = \frac{m_i}{2}$ where m_i is the number of values for the i-th categorical variable and n is the number of categorical variables. Setting the appropriate dimension of categorical embeddings is still an open question with active research onegoing [25].

C. Exploiting both textual and categorical information

Finally, we combined both sources of information (open response and closed-ended questions) into the deep neural network. The text and categorical embedding layers were combined by concatenation. As depicted in Figure 2, we combined the categorical embedding units with the pooled output of the text embeddings and the attention vector, if applicable. The concatenated representations are fed to a regular densely-connected layer with softmax activation, which generates the output and is connected to the loss function.

This implementation is versatile, indeed, a simple selector enables three operation modes that exploit alternative sources of information as input: 1) only open response input; 2) only

closed questions; 3) both text and closed questions. Note that the recurrence only resides in the textual input.

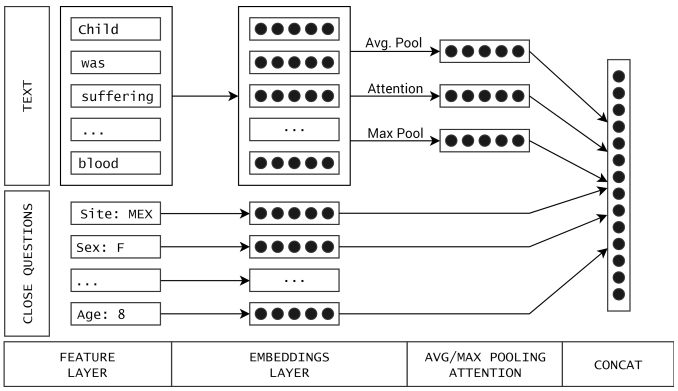

Fig. 2. The dual input combines textual and categorical inputs, from the open responses and closed questions respectively. We compute the latent representations of text from the word embeddings and the Pooling and attention techniques, and the latent representations of the categorical variables with categorical embeddings. Finally, each latent representation is concatenated.

IV. Results

Following the methods presented before, the results section is divided into three main blocks. First, we present the individual contribution of each input type towards the prediction of the CoD; input types were either open response as presented in section III-A or closed questions as in section III-B. Next, we combine both inputs (open response and closed questions following the methodology presented in section III-C) in order to discern if the inputs are redundant or complementary.

In all the cases, apart from the deep neural models (the Recurrent Neural Networks with and without attention, respectively BiGru|Att and BiGru), we also applied alternative models (as **baseline**): a statistical approach, by means of a Naïve Bayes classifier (NB) [26]; a linear approach, by means of a Logistic Regression (LR) classifier [27]; and ensemble models, by means of Random Forest (RF) [28] and Gradient Boosting (XGB) [29] classifiers.

A. Assessment

We evaluated these models by means of Precision, Recall and F-Score [30], which are common **evaluation metrics** apart from NLP [31]. In plain words, "Precision" is defined as the following ratio: of those predicted as belonging to class 'i', how many are indeed from class 'i'; "Recall" is defined as: of those samples whose real class is 'i', how many are correctly predicted; "F-Score" is the harmonic mean of Precision and Recall. To facilitate comparisons with related works, we also provided the "Accuracy" metric. The accuracy is, simply, the correctly predicted instances relative to the total number of predictions made (i.e. hit rate). Note that, given that this is not a binary but a multi-class classification problem (with 48 classes) Tables III, IV and V reported the weighted-average scores [32]. In addition, Appendix A shows per-class detailed results for the best model.

Taking advantage of the fact that the neural network presents the multi-class output as a vector of likelihoods associated to each class (i.e. CoD) as in (3), we can make our system operate as a computer aided VA classification system within the framework of Decision Support Systems. Often, the n-likeliest lists are used to help human coders discriminate from a sub-set rather than from the entire set of classes. For the list to be of any use to a human expert, the correct prediction should be in top positions. To assess the system, in this framework, we provided a commonly used metric to assess ranked lists, that is, Mean Reciprocal Rank (MRR) [33]. The ability of the system to position the true CoD in top positions is what MRR aims to measure. The reciprocal rank of a list of CoDs is the multiplicative inverse of the rank of the correct CoD: 1 if it is predicted in the first place, $\frac{1}{2}$ for second place, etc.

For a system that would rank the correct CoD either in 2nd or 3rd position the MRR would be between 50 and 33. Note that an MRR above 50 suggests that several true CoDs are ranked in the first position.Showing the likeliest CoD as a ranked list switches the application of the model from a fully-automatic classification into a clinical decision support system (DSS), which can be useful for dead-on-arrival cases or when physicians have not previously been with the patient.

With respect to the **assessment scheme** [34], some antecedents [35] applied a repeated hold-out scheme, while in other works it was not clearly stated if the authors used either k-fold-cross-validation or a simple hold-out. In our case, we employed hold-out with the partition shown in Table I. Together with this work, we released the preprocessing software that we implemented[1].

In brief, we assessed the performance of the system for both automatic classification (through Precision, Recall, F-score, Accuracy) and computer-aided classification tasks (through MRR) employing a hold-out scheme. All the metrics are given in a 0-100 scale and with the highest being the best.

B. Performance with only open response

The first set of experiments compare the performance of every model trained only with the open-response. That is, the first operating model explored in the methodology (subsection III-A). Table III shows a summary of the results.

[1]The software is available at http://ixa2.si.ehu.es/prosamed/CorpusVA and can be downloaded with user `CorpusVA` and password `IXACorpusVA`. If the the software is used in any way, this article should be cited.

TABLE III
RESULTS FOR EACH MODEL AND AGE TRAINED WITH **OPEN-RESPONSE**. PRECISION, RECALL AND F-SCORE ARE WEIGHTED AVERAGED.

OR						
Age	Model	Acc.	Prec.	Recall	F-Score	MRR
Adult	NB - Statistical	36.4	37.7	36.4	35.8	51.9
	LR - Linear	44.5	44.9	44.5	44.1	59.2
	RF - Ensemble	44.9	45.4	44.9	44.4	**59.5**
	XGB - Ensemble	**45.6**	**46.0**	**45.6**	**44.7**	59.2
	BiGru - DL	44.7	45.7	44.7	43.5	55.9
	BiGru|Att - DL	43.7	43.8	43.7	42.8	55.8
Child	NB - Statistical	50.1	48.2	50.1	47.3	65.2
	LR	49.4	49.8	49.4	**49.1**	64.4
	RF - Ensemble	50.9	50.8	50.9	47.9	**66.3**
	XGB - Ensemble	46.9	44.5	46.9	43.7	62.7
	BiGru - DL	52.1	**51.9**	52.1	48.5	61.7
	BiGru|Att - DL	**53.3**	47.7	**53.3**	48.8	63.1
Neo	NB - Statistical	61.7	57.9	61.7	59.0	76.4
	LR	62.0	58.7	62.0	59.8	**76.5**
	RF - Ensemble	60.4	54.3	60.4	56.6	75.5
	XGB - Ensemble	59.3	54.2	59.3	55.3	74.6
	BiGru - DL	60.4	57.0	60.4	58.1	74.9
	BiGru|Att - DL	**63.0**	**63.2**	**63.0**	**61.3**	76.4

The experimental evidence reveals several outcomes. First we compared the models by age-module. Regarding the adult age-range, the linear, ensemble and Deep Learning models perform similarly, with the Naïve Bayes falling behind. With respect to the child age-range, the Boosting Gradient model weakens, while the rest of the models, including the Naïve Bayes, present a slight improvement with respect to the results of the adult age-range. Finally, concerning the neonate age-module, the deep models are superior, with the BiGru with attention being the best model overall, only ceding the MRR in favour of the Logistic Regression. The highest MRRs per age-module are 59.5, 66.3 and 76.5, which means that most codes are found between 1st and 2nd position.

On the one hand, while the Gradient Boosting model performs well on the adult segment, it becomes weaker on the others. On the other hand, the opposite happens to the Deep Learning models, which show better performance in the child segment than in the adult, and the best performance in the neonate segment. It is important to note that the number of classes dramatically changes among age-ranges, as shown in Figures 8, 9 and 10, so the XGB model might do better with a higher number of classes while the BiGru would prevail with fewer classes.

An **error analysis** was carried out to find out which classes are often misclassified and which classes were predicted instead. This was depicted in terms of confusion matrices represented as heat-maps in Figure 3. Note that regarding the color-bars, the brighter, the greater the number of samples found on that specific coordinate of the confusion matrix. External causes of death (e.g., *Fires*, *Road Traffic*, ...) are mostly correctly predicted for every model and in every age-module. *Stroke* is often misclassified with several causes of death (e.g., *Diabetes* or *Acute Myocardial Infarction*). *Pneumonia* is a frequent class usually predicted correctly but is often misclassified with *Sepsis*, *Meningitis*, *Measles* or *Malaria* (infectious diseases). In the adult segment, VAs assigned with *Other Non-communicable Diseases* as the CoD are often misclassified as *Stomach Cancer*, *Prostate Cancer*, *Lung Cancer*, *Colorectal Cancer*, *Breast Cancer* and *Cervical Cancer*. Note that, cancer is, indeed, a sub-set of non-communicable diseases. In the child segment *Diarrhea/Dysentery* is often misclassified with *Other Infectious Diseases* (*Sepsis*, *Meningitis*, *Encephalitis* etc.).

It seems that the evaluation of the classifiers as Decision Support Systems (with MRR metric), as opposed to their use as automatic classifiers, could lead to stronger performances. It is possible to use our classifiers as a DSS because the systems provide the CoDs ranked, which helps to generate a set of the most probable Causes of Death for each Verbal Autopsy.

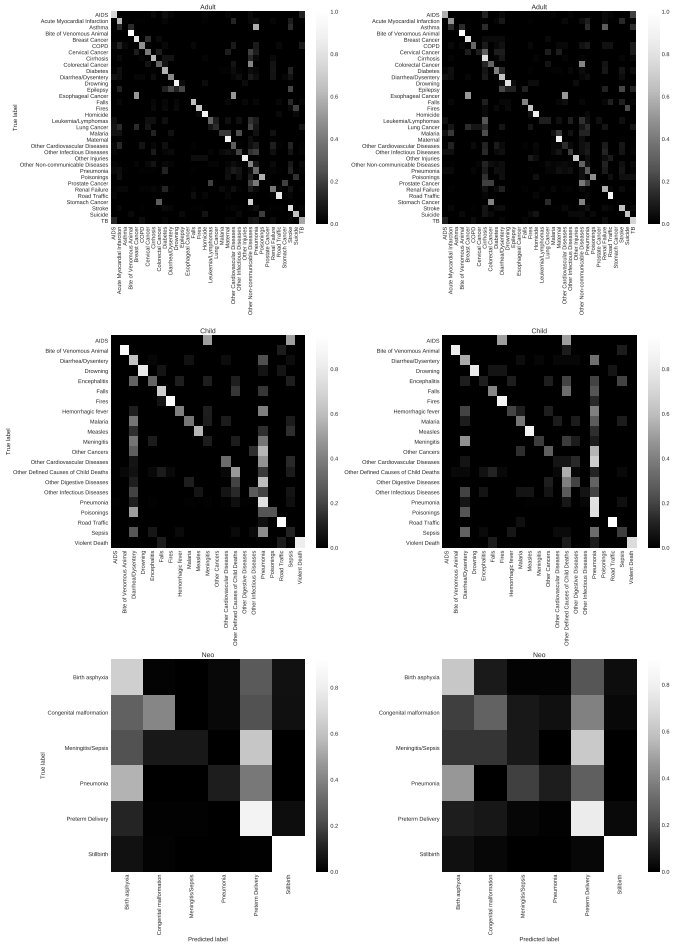

Fig. 3. Head-to-head comparison of normalized confusion matrices showing true label density (y axis) predicted label density (x axis) for each age-range trained only with open-response. Left column: Gradient Boosting. Right column: BiGru with attention.

C. Performance with only closed questions

The next set of experiments examine the performance of each model trained only with the closed-ended questions, and thus, with no textual feature, that is the model presented in section III-B compared to classical approaches. Table IV shows the results obtained by each classifier with only categorical variables.

Regarding the adult and neonate age-range, the Boosting Gradient model is the strongest, with the best results for every metric. In regards to the child age-range, we have divided

results. The Random Forest model gets the higher score in Precision and F-Score, while XGB performs better in terms of Recall and MRR.

The data show that the ensemble models, and to a less extent, also the linear, are in general stronger when dealing with categorical variables than the deep models. The ensembles perform better in every age-range. It is important to point out that the deep model without textual features is merely a single-layer feed-forward neural network.

TABLE IV
EXPERIMENTAL RESULTS FOR EACH MODEL AND SEGMENT TRAINED ONLY WITH CLOSE-QUESTIONS. PRECISION, RECALL AND F-SCORE ARE WEIGHTED AVERAGED.

CQ						
Age	Model	Acc.	Prec.	Recall	F-Score	MRR
Adult	NB - Statistical	38.2	38.0	38.2	37.3	54.3
	LR - Linear	41.4	41.5	41.4	41.1	57.8
	RF - Ensemble	48.0	45.6	48.0	45.9	62.6
	XGB - Ensemble	**50.6**	**48.4**	**50.6**	**48.9**	**64.9**
	BiGru - DL	35.4	39.2	35.4	32.1	49.3
Child	NB - Statistical	49.1	49.9	49.1	48.7	63.9
	LR - Linear	44.2	48.6	44.2	45.1	61.1
	RF - Ensemble	52.9	**55.6**	52.9	**52.9**	67.0
	XGB - Ensemble	**53.1**	49.8	**53.1**	50.4	**67.1**
	BiGru - DL	41.4	44.0	41.4	38.2	55.3
Neo	NB - Statistical	64.6	65.4	64.6	64.6	78.3
	LR - Linear	60.6	62.5	60.6	61.4	76.4
	RF - Ensemble	69.4	69.4	69.4	69.0	81.7
	XGB - Ensemble	**72.6**	**73.6**	**72.6**	**71.5**	**83.3**
	BiGru - DL	71.5	73.1	71.5	69.2	81.0

D. Performance with both open and closed responses

The final set of experiments compare the performance of every model trained jointly with the open-response and the closed-ended questions. Table V shows a summary of the results.

TABLE V
EXPERIMENTAL RESULTS FOR EACH MODEL AND AGE-RANGE TRAINED JOINTLY WITH OPEN-RESPONSE AND CLOSED-ENDED QUESTIONS. PRECISION, RECALL AND F-SCORE ARE WEIGHTED AVERAGED.

Dual						
Age	Model	Acc.	Prec.	Recall	F-Score	MRR
Adult	NB - Statistical	44.2	43.6	44.2	43.0	58.9
	LR - Linear	46.5	46.2	46.5	46.1	61.6
	RF - Ensemble	54.1	52.2	54.1	52.1	67.8
	XGB - Ensemble	**55.9**	**54.6**	**55.9**	**54.7**	**69.2**
	BiGru - DL	51.6	54.4	51.6	50.6	63.8
	BiGru|Att - DL	51.2	51.4	51.2	49.4	63.4
Child	NB - Statistical	54.6	53.0	54.6	53.0	68.8
	LR - Linear	64.5	48.1	50.9	48.1	48.7
	RF - Ensemble	56.1	54.2	56.1	53.0	**70.8**
	XGB - Ensemble	**56.3**	**54.5**	**56.3**	**54.0**	70.3
	BiGru - DL	52.4	54.3	52.4	51.2	64.9
	BiGru|Att - DL	53.8	51.6	53.8	51.0	66.7
Neo	NB - Statistical	66.2	65.8	66.2	65.6	79.9
	LR - Linear	68.6	68.4	68.6	68.5	80.8
	RF - Ensemble	70.5	69.9	70.5	69.0	82.1
	XGB - Ensemble	**73.7**	**73.3**	**73.7**	**72.0**	**84.2**
	BiGru - DL	69.7	68.4	69.7	68.4	81.2
	BiGru|Att - DL	69.7	68.4	69.7	68.6	80.9

The empirical results show that the performance of every model, on practically all metrics, improves due to the joint handle of textual and categorical variables. The models leveraging the open response and closed-ended questions are the most solid when predicting the CoD and also when generating ranking labels.

E. Model interpretability

The Bidirectional GRU with the attention mechanism allows to explore how much each fragment of the text contributed in the prediction. The studied models performed differently for each CoD, but we can leverage the model interpretability to gain insights into the models' behaviour discrepancies, focusing on the most divergent CoDs.

First, let us take as an example the *TB* and *Suicide* CoDs as those substantially better predicted by the BiGru model over the XGB. In Figure 4 we show a VA open response about a patient with *TB* and in Figure 5 a VA open response concerning a patient who committed suicide, both correctly predicted.

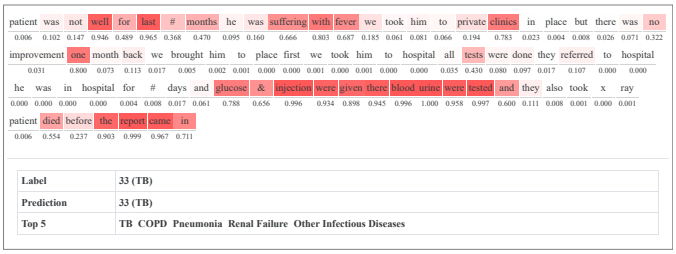

Fig. 4. BiGru attention weights over VA (*ID=850*) open response with gold CoD *TB* correctly predicted.

The VA (*ID=850*) open response belongs to a patient with tuberculosis. The respondent summarized the evolution of the patient, explaining key issues such as past symptoms ("he was suffering with fever"), where the care was sought ("we took him to private clinics"), conducted diagnostic tests ("blood urine were tested"), and applied treatment ("glucose & injection were given"). Notice the heatmap over tokens according to the attention scores, which effectively detect these issues as being the most significant text fragments of the CoD ascertainment. Note that the CoD is not explicitly mentioned, but is correctly inferred by the symptoms, tests and treatment applied to the patient.

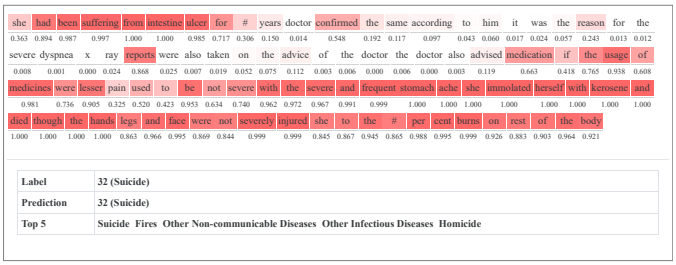

Fig. 5. BiGru attention weights over VA (*ID=7752*) open response with gold CoD *Suicide* correctly predicted.

The VA (*ID=7752*) open response relates to a patient who committed suicide. The interviewee gave a brief history of the patient, mentioning past diseases ("had been suffering from intestine ulcer"), symptoms ("severe and frequent stomach ache"), and prescription ("doctor also advised medication").

Note that in this case, the CoD is *Suicide*, which can be denoted literally or describing a suicide. This open response corresponds to the latter ("she immolated herself with kerosene"), which is the harder situation for an algorithm, as it needs to have natural language understanding capabilities. Regarding the attention scores, the model pays attention to patient history and death description, but it manages to correctly predict the CoD focusing on the suicide narrative while disregarding the patient history information.

Now, let us take as an example the *Colorectal Cancer* and *Epilepsy* as those CoDs were considerably better predicted by the XGB model over the BiGru. In Figure 6 we show a VA open response concerning a patient with *Colorectal Cancer* and in Figure 7 a VA open response about a patient with *Epilepsy*, both incorrectly predicted.

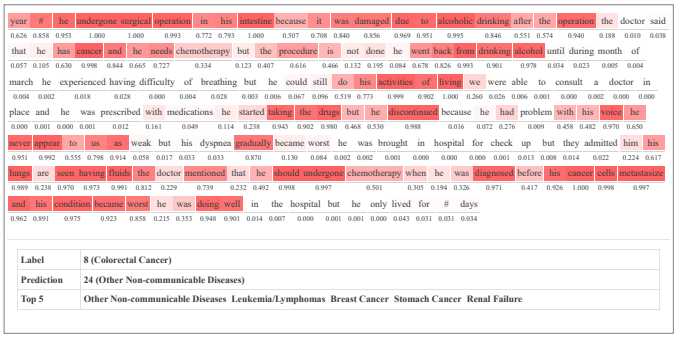

Fig. 6. BiGru attention weights over VA (*ID=2603*) open response with gold CoD *Colorectal Cancer* and predicted as *Other Non-communicable Diseases*.

The VA (*ID=2603*) open response concerns a patient with *Colorectal Cancer* and describes the patient CoD from the beginning of the illness. The respondent started by reporting an undergone surgical operation on the patient's intestine, in which the doctor found cancer cells. Then, he or she reported treatment ("he needs chemotherapy"), life habits ("it was damaged due to alcoholic drinking") and other symptoms ("his lungs are seen as having fluids"). The open response does not specify the CoD, but it suggests that it was cancer, and as the cancer was found during an intestinal operation, the prediction of colorectal cancer is coherent. Nonetheless, the gold CoD is *Other Non-communicable Diseases*, which indeed includes neoplasms in general and colorectal cancer in particular.

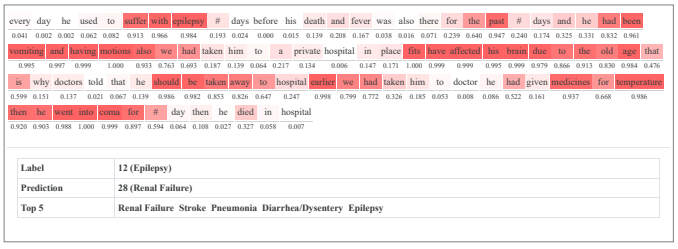

Fig. 7. BiGru attention weights over VA (*ID=3985*) open response with gold CoD *Epilepsy* and predicted as *Renal Failure*.

The VA (*ID=3985*) open response concerns a patient with *Epilepsy*. At the start of the open response, the interviewee explicitly mentions the CoD ("he used to suffer with epilepsy"), and then briefly explains the patient status the days prior to the death, mentioning several symptoms ("had been vomiting and having motions, went into coma") or treatments ("medicines for temperature"). The attention scores highlight the CoD and the symptoms, but finally, the prediction is biased into a CoD consistent with the symptoms but incorrect, *Renal Failure*.

We find that the BiGru model's greatest performance improvements come from its natural language understanding ability to infer CoD from symptoms, diagnostic tests, treatments and non-specific descriptions, such as in the presented examples of *TB* and *Suicide*. Nevertheless, we notice that the XGB model outperformed the BiGru in handling overlapping classes, like *Colorectal Cancer* and *Other Non-communicable Diseases*. This is also the case when the CoD is explicitly specified, arguably due to the BiGru bias on predicting based on other indicators such as symptoms, instead of simply focusing on the CoD, if it is mentioned.

V. Discussion

Overall, we adapted emerging neural models to cope with the dual input (closed-ended questionnaire and open response) available in this work. While related works focus mainly on information extraction from the closed-ended questionnaire, we found that the open response conveys valuable information. The performance of the system trained with only the open response remains close to that attained with the closed questions. Given that the closed-ended questionnaire is designed by virtue of expert knowledge, the information yielded by the open response (a spontaneous narrative provided by a non-expert) encompasses valuable information such as keywords (e.g., fever, snake or even the CoD itself or a near synonym) that should not be discarded. For example, in the case of the child segment, both the open response and the closed questions yield similar results, with a $MRR \approx 66.5$ meaning that, on average both systems rank the correct CoD in position $1/MRR = 1.5$. Note that the open response only comprises a few dozen words uttered by a non-expert while there are more than 300 closed questions explicitly designed with care in an attempt to discard CoDs.

Nevertheless, an improvement was achieved combining the dual input, and we found that the closed questions are the most informative part of the VAs, while the open response is almost as informative as the closed questions. Remarkably, the dual input is more informative than the individual contribution of each of these.

As reported in the results, up to 40% of the narratives were considered non-informative with answers such as "Respondent has nothing to add" or "no comment". Given the technological advances in VA implementation with the wide use of tablets to collect the information, countries could consider a number of strategies to improve the usefulness of the VA narrative. First, recording the narrative (as opposed to having it summarized by the interviewer and being transcribed automatically) by means of state of the art ASR approaches achieves a Percentage Error below 6%. Thus, they are highly reliable [36], [37], [38]. In addition, structured narratives are easier to analyse as experience shows with the analysis of medical records [39].

In this work, we compared neural networks with non-neural approaches and we found that neural approaches do not necessarily offer the best performance. In fact, deep neural networks tend to achieve high performance with big data provided that high computation resources are available. Nevertheless, in this task, we struggle with data scarcity, amidst a high lexical variability, as revealed by OOV words and the label-density imbalance shown in Tables I and II. Moreover, in the context of this task, we also strive to develop techniques that can be run in mid and low-end machines. Deep neural networks are often criticised for being used as a black box that lacks interpretability [40], [41]. To cope with this issue, current trends are promoting Explainable Artificial Intelligent (XAI). From XAI strategies applied to text classification, the attention mechanisms [24] are an approach with a good cost-effectiveness trade-off in terms of computation and required amount of training data. The calculation of the attention weights is cheap since it does not imply an overload in the training process[42], but at the same time, it reduces the opacity of the models considerably.

In addition, we assessed the models as a decision support system, in contrast to antecedent works [43], [44], as our systems do rank the CoDs and provide interpretable outputs as the most important words, evaluating interpretability as a key factor.

A. Related work

In recent years different methods have been investigated for automatically assigning CoD from Verbal autopsy and different comparative studies have been published. In [35] the authors concluded that three automated diagnostic methods, Tariff, Simplified Symptom Pattern (SSP), and Random Forest (RF) performed better than physician review. The authors used accuracy as the evaluation metric.

A neural network that consists of a gated recurrent unit layer (GRU) is used in [45] for CoD classification. In this case, they classified free-text narratives of the VA records of the Million Death Study (MDS) [46]. The prediction layer of the GRU outputs the probabilities of 18 CoDs and the most likely code is chosen.

An expert group led by the World Health Organization (WHO) generated the 2016 WHO VA instrument [6], based on the 2012 instrument, and they proposed three analytical software tools, fully compliant with the 2016 VA instrument. We found that algorithms InterVA, InSilicoVA and SmartVA outperformed the Bayesian approach. **InterVA** [43] is an algorithm based on Bayes theorem and calculates the conditional probability of each particular CoD given a set of events. The events are the responses of the interview questionnaire in the form of binary indicators, representing whether the event happens or not. **InSilicoVA** [44] is also based on Bayes theorem and the authors argue that their version improves the results obtained by InterVA. They have used accuracy to measure the behavior of both methods on the VA data generated by the Agincourt Health and Demographic Surveillance System (HDSS) [47]. The **SmartVA** [12] algorithm was designed and validated by the Population Health Metrics Research Consortium (PHMRC) Gold Standard VA database [14]. It is a probabilistic model calculated as a function of the fraction of deaths for each item that has a positive response. In [13] the authors concluded that InSilicoVA performed better than all other algorithms.

In [48] the authors explained six VA algorithms that have been successfully tested: the aforementioned InterVA, SmartVA (Tariff), and InSilicoVA, jointly with the Naïve Bayes classifier (NB) and the King-Lu algorithm. They concluded that the performance of VA algorithms does not present a consensus that clearly identifies the best algorithm.

Closed-ended questions and open-ended narratives of the VAs are used as input of the SmartVA algorithm [12]. The results are significantly better when the open-ended narrative is taken into account. From the narrative, key words are identified and words with the same or similar meanings are grouped. Our models are also capable of extracting meaningful information from the open response, and even leverage the dual input, as the best performance is achieved with both open response and closed-ended questions.

There are different studies in which the objective is to examine if VA is suitable for predicting CoDs. [49], [50] concluded that VA is a suitable instrument for that objective. Bearing this fact in mind, the automated classification of VAs according to the CoD is a task that requires enhancing and can be approached using NLP techniques. It is important to consider different alternatives regarding two aspects: i) VA representation and ii) classification methods. Works such as [35] concluded that the open response, in addition to the closed response, contains relevant information that helps in the classification task. We have examined the influence of using either one or both parts of the VA to automatically predict the CoD. Most of the proposed solutions for automatically deducing the CoD are based on statistical tools but in [45] they tested neural networks obtaining encouraging results. We have proved both classical approaches, some of them based on statistical methods, and neural networks, and we found that every model can leverage the open response. However, Deep Learning based models provide the most informative models regarding interpretability and explainability by virtue of emerging XAI approaches [40], [51] that have access to big data and computation resources.

VI. Conclusions

In this work, we tackle the **task** of assigning the correct Cause of Death to a given Verbal Autopsy. This task is essential in impoverished regions whose languages are often not broadly studied. We feel that NLP can help to classify VAs automatically, and then monitor the evolution of CoD per region and age-module. Therefore, it is important to extend the models to other languages (e.g., Swahili, Vietnamese, Malaysian). The VAs comprise two sources of information: categorical information from the closed-ended questionnaire, and textual information from the open response. However, it is important to note that all the information is given by non-professionals, usually relatives and relatives and people close to the deceased.

Previous works approached CoD classification mainly with classical algorithms that require feature engineering (this is why the closed-ended questionnaire was the main source of information), and turned to Bayesian solutions. In summary, these were mainly focused on single input and traditional classification approaches. We found it important to compare the emerging **Deep Learning** approach with the traditional approaches developed in related work of VA classification. One of the benefits of Deep Learning rests on the ability of the algorithm to discover latent features, avoiding the feature engineering required to leverage the closed-ended questionnaire. We did this with single input (either closed-ended questionnaire or open response) and also combining the dual input. Our **research question** concerns the information conveyed by the open response, and aims to to discover if both are synergistic. Specifically, we assessed six classification algorithms. For the baseline model, we chose the classical approach Naïve Bayes (NB) served with TF-IDF and one-hot input, as it is the main trend in related tools. Besides this, we explored both linear models and ensemble models based on trees.

We **learnt** that, curiously enough, open response (only text in natural language) conveys nearly the same information as the closed-ended questionnaire. Thus, there is no need to dismiss the open response. Moreover, this encourages us to ask the interviewer to collect VAs with meaningful text to minimize the number of stop-texts. Additionally, dual input is not redundant and is of value. Even though the combination strategy implemented in this work was quite simplistic (bear in mind that concatenation was used), the results are promising. This opens a new line of research on how to combine the information available in the VAs and influence, accordingly, the design of forthcoming VA standards.

While the dominant approach in the antecedents were Bayesian models, with our models, the method that provided, overall, the best performance was BiGru. Our **contribution** is showing that the performance of this approach is competitive as an automatic classification system and also as a decision support system if we let the expert choose the CoD from a ranked list. What is more, BiGru can be provided with attention mechanisms that can help the expert understand why the system made its prediction. In our view, interpretability is a key factor for the success of a model in its real use phase. We do not consider that previous research have focused on this aspect.

Deep Learning with **attention mechanisms** provides interpretable models with word attention, which is useful in several ways. On the one hand, word attention allows the performing of an error analysis with enough detail to point towards some frequent causes of misclassification, such as i) CoDs with different levels of specificity (e.g., assign *Other Non-communicable Diseases* instead of *Colorectal Cancer* and vice-versa). ii) Mention of another CoD (e.g., the open response contains the mention of a CoD which is not the gold CoD, due to multi-causality, similarity or respondents' confusion or ignorance). On the other hand, physicians find it useful to understand what motivated the models' predictions when used as a decision support system.

Although the performance of the classifiers have room for improvement, we have already had some appealing results. The amount of available training data is a notable limitation in increasing the performance of this task, and it seems that inference ability is limited by the class imbalance. We argue that the incorporation of more training data, along with the state-of-the-art NLP approaches and techniques, could advance the VA automatic classification field.

There are appealing options for **future work**. In the short term, we are working on algebraic operations to leverage dual input combination and alternative representations for traditional classifiers, and aspire to compare our best performing models with the existing algorithms. In the midterm, we intend to analyze other existing algorithms and strive to improve their performance with the inclusion of the open response. We are also determined to address active learning in an attempt to incorporate the human in the loop. That is, we found that the system can be fine-tuned in the first stage, with the system applied as a DSS, where the human can choose the correct output. With that strategy, the inference algorithm can learn from the errors made (in the cases in which the system did not provide the correct CoD in the first place). To this end, attention mechanisms shall be re-approached. Next, in a second stage, the system would work fully automatically. We are thinking of carrying out this through a crowd-sourced project against a web-based prototype. In the long term, we consider that it is important to study attention mechanisms in order to continue developing the design of VAs leading to the next VA standard with attention to the open response, such as avoiding texts without meaningful inputs.

APPENDIX A
DEEP LEARNING WITH ATTENTION: PER CLASS EVALUATION

The Figures 8-10 show the results of the BiGru|Att model presented in Table V, desegregated by class, along with the test set class frequency, for reference. In the first figure, on the top, in the series of Figures 8-10, we find that some CoDs are quite frequent while 41% appear fewer than 25 times (this is a reflect of the class skew in the data). In the adult segment (Figure 8), the worst performing classes are those that are less frequent and cancer-related, due to the presence of the *Other Non-communicable Diseases* class. Regarding the child segment (Figure 9), the difficult classes are also the less frequent and the infectious diseases, due to the labelling issues found in the error analysis (page 5). In the neonate segment (Figure 10), there are only six classes and the performance is correlated with the class support, exhibiting a remarkable ability to discriminate between similar classes such as *Stillbirth*, *Preterm Delivery* and *Birth asphyxia*. However, even though the CoD is skewed, the system was able to make valuable predictions on each class (not only on those with high support). Confusion matrices provided in Figure 3 show, explicitly, the classes confounded by the system e.g. Pneumonia was often misclassified with Other Infectious Diseases.

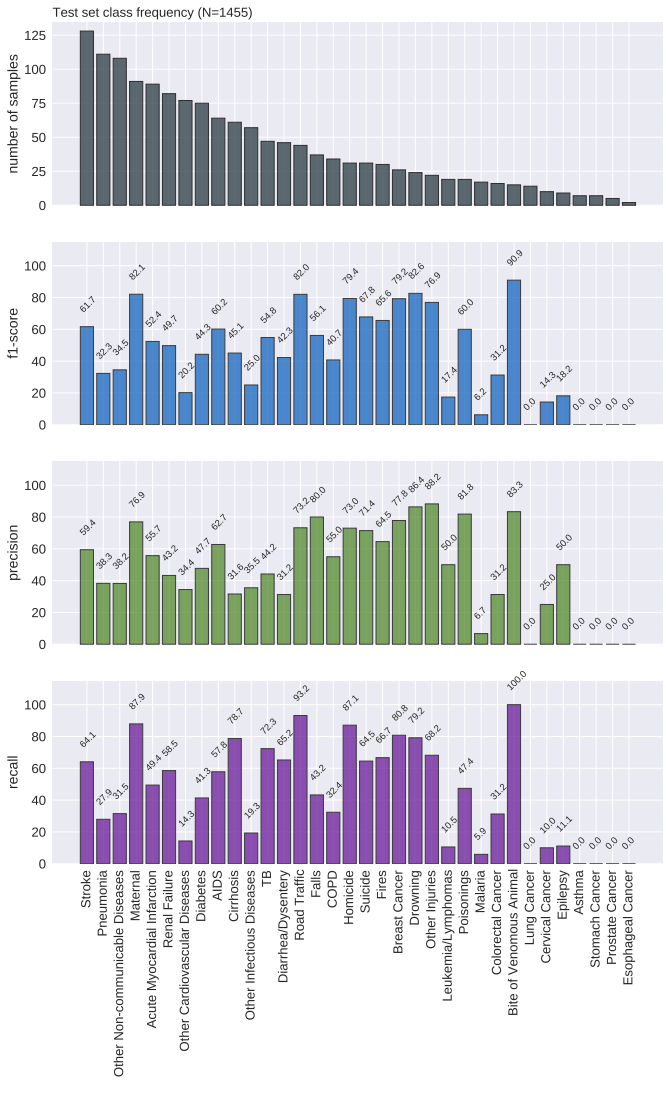

Fig. 8. Per-class evaluation showing weighted Precision, Recall and F-Score of BiGru|Att model with both open and closed responses, along with test set class frequency, for **adult** segment.

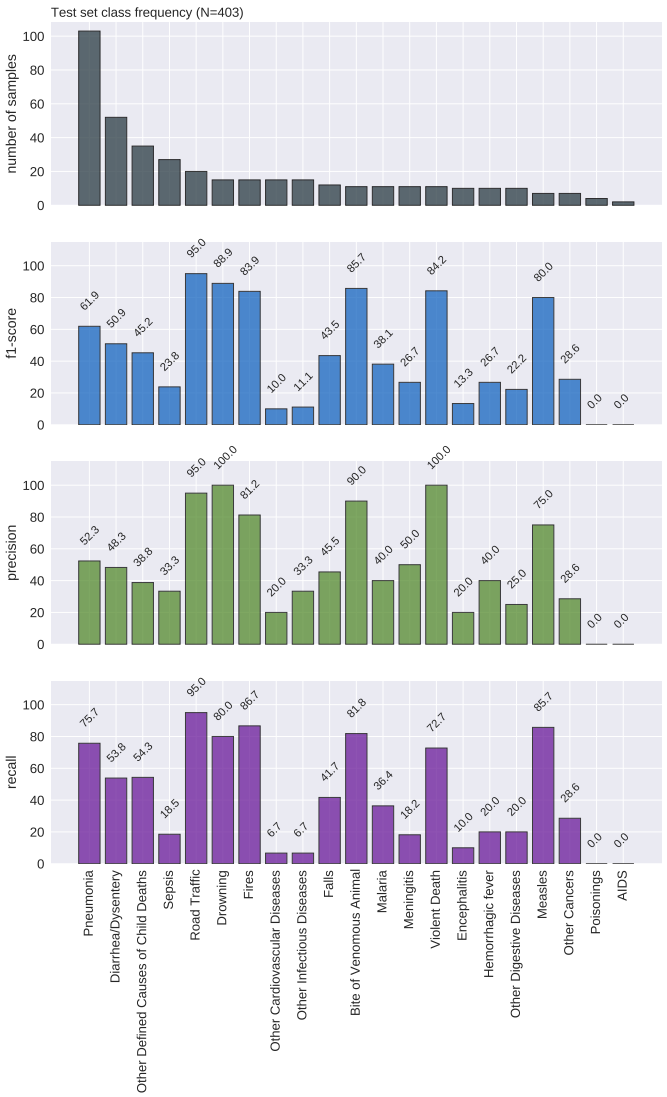

Fig. 9. Per-class evaluation showing weighted Precision, Recall and F-Score of BiGru|Att model with both open and closed responses, for **child** segment.

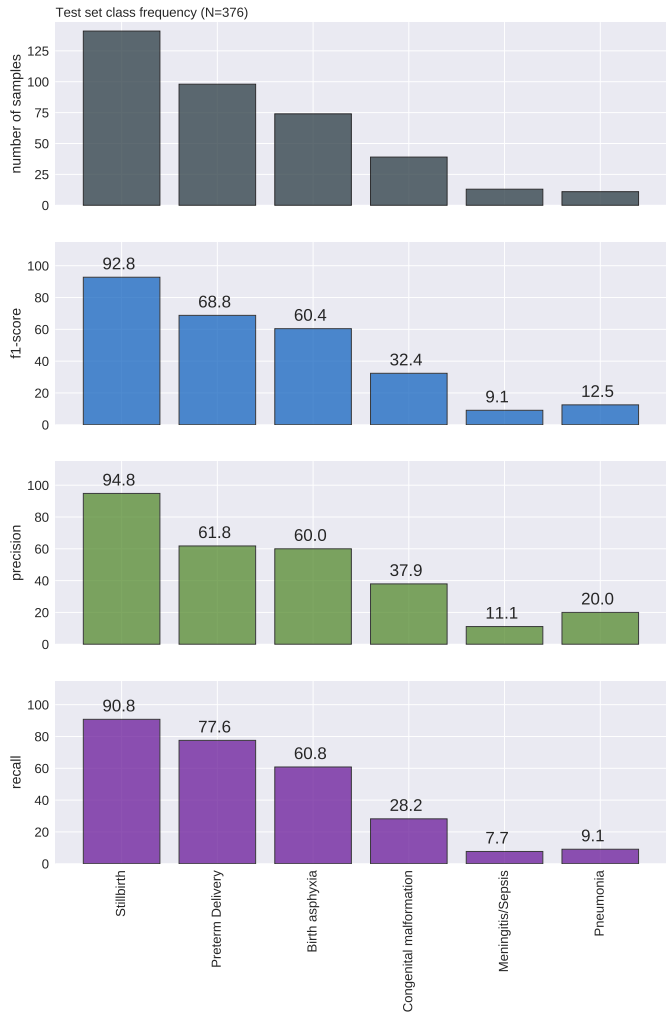

Fig. 10. Per-class evaluation showing weighted Precision, Recall and F-Score of BiGru|Att model with both open and closed responses, along with test set class frequency, for **neonate** segment.

APPENDIX B
CLASSIFICATION PERFORMANCE FOR EACH MODEL AND AGE-RANGE

Table VI combines the results shown in Tables III, IV and V in a single table with the aim of helping the reader compare the performance of various models and data inputs at a glance.

TABLE VI
RESULTS FOR EACH MODEL AND AGE TRAINED WITH EACH DATA INPUT (**OPEN-RESPONSE**, **CLOSE-QUESTIONS** AND **DUAL**). PRECISION, RECALL AND F-SCORE ARE WEIGHTED AVERAGED.

OR						
Age	Model	Acc.	Prec.	Recall	F-Score	MRR
Adult	NB - Statistical	36.4	37.7	36.4	35.8	51.9
	LR - Linear	44.5	44.9	44.5	44.1	59.2
	RF - Ensemble	44.9	45.4	44.9	44.4	**59.5**
	XGB - Ensemble	**45.6**	**46.0**	**45.6**	**44.7**	59.2
	BiGru - DL	44.7	45.7	44.7	43.5	55.9
	BiGru|Att - DL	43.7	43.8	43.7	42.8	55.8
Child	NB - Statistical	50.1	48.2	50.1	47.3	65.2
	LR - Linear	49.4	49.8	49.4	**49.1**	64.4
	RF - Ensemble	50.9	50.8	50.9	47.9	**66.3**
	XGB - Ensemble	46.9	44.5	46.9	43.7	62.7
	BiGru - DL	52.1	**51.9**	52.1	48.5	61.7
	BiGru|Att - DL	**53.3**	47.7	**53.3**	48.8	63.1
Neo	NB - Statistical	61.7	57.9	61.7	59.0	76.4
	LR - Linear	62.0	58.7	62.0	59.8	**76.5**
	RF - Ensemble	60.4	54.3	60.4	56.6	75.5
	XGB - Ensemble	59.3	54.2	59.3	55.3	74.6
	BiGru - DL	60.4	57.0	60.4	58.1	74.9
	BiGru|Att - DL	**63.0**	**63.2**	**63.0**	**61.3**	76.4
CQ						
Adult	NB - Statistical	38.2	38.0	38.2	37.3	54.3
	LR - Linear	41.4	41.5	41.4	41.1	57.8
	RF - Ensemble	48.0	45.6	48.0	45.9	62.6
	XGB - Ensemble	**50.6**	**48.4**	**50.6**	**48.9**	**64.9**
	BiGru - DL	35.4	39.2	35.4	32.1	49.3
Child	NB - Statistical	49.1	49.9	49.1	48.7	63.9
	LR - Linear	44.2	48.6	44.2	45.1	61.1
	RF - Ensemble	52.9	**55.6**	52.9	**52.9**	67.0
	XGB - Ensemble	**53.1**	49.8	**53.1**	50.4	**67.1**
	BiGru - DL	41.4	44.0	41.4	38.2	55.3
Neo	NB - Statistical	64.6	65.4	64.6	64.6	78.3
	LR - Linear	60.6	62.5	60.6	61.4	76.4
	RF - Ensemble	69.4	69.4	69.4	69.0	81.7
	XGB - Ensemble	**72.6**	**73.6**	**72.6**	**71.5**	**83.3**
	BiGru - DL	71.5	73.1	71.5	69.2	81.0
Dual						
Age	Model	Acc.	Prec.	Recall	F-Score	MRR
Adult	NB - Statistical	44.2	43.6	44.2	43.0	58.9
	LR - Linear	46.5	46.2	46.5	46.1	61.6
	RF - Ensemble	54.1	52.2	54.1	52.1	67.8
	XGB - Ensemble	**55.9**	**54.6**	**55.9**	**54.7**	**69.2**
	BiGru - DL	51.6	54.4	51.6	50.6	63.8
	BiGru|Att - DL	51.2	51.4	51.2	49.4	63.4
Child	NB - Statistical	54.6	53.0	54.6	53.0	68.8
	LR - Linear	64.5	48.1	50.9	48.1	48.7
	RF - Ensemble	56.1	54.2	56.1	53.0	**70.8**
	XGB - Ensemble	**56.3**	**54.5**	**56.3**	**54.0**	70.3
	BiGru - DL	52.4	54.3	52.4	51.2	64.9
	BiGru|Att - DL	53.8	51.6	53.8	51.0	66.7
Neo	NB - Statistical	66.2	65.8	66.2	65.6	79.9
	LR - Linear	68.6	68.4	68.6	68.5	80.8
	RF - Ensemble	70.5	69.9	70.5	69.0	82.1
	XGB - Ensemble	**73.7**	**73.3**	**73.7**	**72.0**	**84.2**
	BiGru - DL	69.7	68.4	69.7	68.4	81.2
	BiGru|Att - DL	69.7	68.4	69.7	68.6	80.9

ACKNOWLEDGMENT

This work was partially supported by the Spanish Ministry of Science and Technology (PAD-MED PID2019-106942RB-C31) and by the Basque Government (Elkartek KK-2019/00045, IXA IT-1343-19, Predoctoral Grant PRE-2019-1-0158)

REFERENCES

[1] C. AbouZahr, D. De Savigny, L. Mikkelsen, P. W. Setel, R. Lozano, E. Nichols, F. Notzon, and A. D. Lopez, "Civil registration and vital statistics: progress in the data revolution for counting and accountability," *The Lancet*, vol. 386, no. 10001, pp. 1373–1385, 2015.

[2] Department of Economic and Social Affairs, *Principles and recommendations for a vital statistics system*. New York: United Nations, 2014.

[3] P. W. Setel, S. B. Macfarlane, S. Szreter, L. Mikkelsen, P. Jha, S. Stout, C. AbouZahr, M. of Vital Events (MoVE) writing group *et al.*, "A scandal of invisibility: making everyone count by counting everyone," *The Lancet*, vol. 370, no. 9598, pp. 1569–1577, 2007.

[4] L. Mikkelsen, D. E. Phillips, C. AbouZahr, P. W. Setel, D. De Savigny, R. Lozano, and A. D. Lopez, "A global assessment of civil registration and vital statistics systems: monitoring data quality and progress," *The Lancet*, vol. 386, no. 10001, pp. 1395–1406, 2015.

[5] O. Sankoh and P. Byass, "Cause-specific mortality at INDEPTH health and demographic surveillance system sites in Africa and Asia: concluding synthesis," *Global health action*, vol. 7, no. 1, p. 25590, 2014.

[6] E. K. Nichols, P. Byass, D. Chandramohan, S. J. Clark, A. D. Flaxman, R. Jakob, J. Leitao, N. Maire, C. Rao, I. Riley *et al.*, "The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0," *PLoS medicine*, vol. 15, no. 1, p. e1002486, 2018.

[7] World Health Organization *et al.*, "Verbal autopsy standards: the 2012 WHO verbal autopsy instrument," *Geneva: WHO*, 2012.

[8] P. Serina, I. Riley, A. Stewart, A. D. Flaxman, R. Lozano, M. D. Mooney, R. Luning, B. Hernandez, R. Black, R. Ahuja *et al.*, "A shortened verbal autopsy instrument for use in routine mortality surveillance systems," *BMC medicine*, vol. 13, no. 1, p. 302, 2015.

[9] C. J. Murray, A. D. Lopez, D. M. Feehan, S. T. Peter, and G. Yang, "Validation of the symptom pattern method for analyzing verbal autopsy data," *PLoS Medicine*, vol. 4, no. 11, p. e327, 2007.

[10] World Health Organization *et al.*, "Civil registration and vital statistics 2013: challenges, best practice and design principles for modern systems," *WHO/HMN, December*, 2013.

[11] N. Desai, L. Aleksandrowicz, P. Miasnikof, Y. Lu, J. Leitao, P. Byass, S. Tollman, P. Mee, D. Alam, S. K. Rathi *et al.*, "Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low-and middle-income countries," *BMC medicine*, vol. 12, no. 1, p. 20, 2014.

[12] P. Serina, I. Riley, A. Stewart, S. L. James, A. D. Flaxman, R. Lozano, B. Hernandez, M. D. Mooney, R. Luning, R. Black *et al.*, "Improving performance of the Tariff method for assigning causes of death to verbal autopsies," *BMC medicine*, vol. 13, no. 1, p. 291, 2015.

[13] T. H. McCormick, Z. R. Li, C. Calvert, A. C. Crampin, K. Kahn, and S. J. Clark, "Probabilistic cause-of-death assignment using verbal autopsies," *Journal of the American Statistical Association*, vol. 111, no. 515, pp. 1036–1049, 2016.

[14] C. J. Murray, A. D. Lopez, R. Black, R. Ahuja, S. M. Ali, A. Baqui, L. Dandona, E. Dantzer, V. Das, U. Dhingra *et al.*, "Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets," *Population health metrics*, vol. 9, no. 1, p. 27, 2011.

[15] A. D. Flaxman, L. Harman, J. Joseph, J. Brown, and C. J. Murray, "A de-identified database of 11,979 verbal autopsy open-ended responses," *Gates open research*, vol. 2, 2018.

[16] S. Santiso, A. Casillas, and A. Pérez, "The class imbalance problem detecting adverse drug reactions in electronic health records," *Health informatics journal*, pp. 1–11, 2018.

[17] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *International Jorunal of Mathematical Models and Methods in Applied Sciences*, vol. 1, no. 2, p. 1, 2007.

[18] E. Collins, N. Rozanov, and B. Zhang, "Evolutionary data measures: Understanding the difficulty of text classification tasks," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 380–391. [Online]. Available: http://www.aclweb.org/anthology/K18-1037

[19] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[25] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," *arXiv preprint arXiv:1604.06737*, vol. 1, 04 2016.

[26] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, no. 2, p. 3, 2004.

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

[28] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[29] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[30] L. KUNCHEVA, *PATTERN RECOGNITION AND NEURAL NETWORKS*. S.l: LULU COM, 2019.

[31] C. D. Manning, C. D. Manning, and H. Schütze, *Foundations of statistical natural language processing*. London: MIT press, 1999.

[32] J. Nam, J. Kim, E. Loza Mencía, I. Gurevych, and J. Fürnkranz, "Large-scale multi-label text classification - revisiting neural networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8725 LNAI, no. PART 2, pp. 437–452, 2014.

[33] N. Craswell, *Mean Reciprocal Rank*. Boston, MA: Springer US, 2009, pp. 1703–1703. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_488

[34] Y. Goldberg, "Neural network methods for natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 10, no. 1, pp. 1–309, 2017.

[35] C. J. Murray, R. Lozano, A. D. Flaxman, P. Serina, D. Phillips *et al.*, "Using verbal autopsy to measure causes of death: the comparative performance of existing methods," *BMC Medicine*, vol. 12, no. 1, p. 5, Jan 2014. [Online]. Available: https://doi.org/10.1186/1741-7015-12-5

[36] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim *et al.*, "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

[37] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[38] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5934–5938.

[39] N. Hong, A. Wen, F. Shen, S. Sohn, S. Liu, H. Liu, and G. Jiang, "Integrating structured and unstructured ehr data using an fhir-based type system: a case study with medication data," *AMIA Summits on Translational Science Proceedings*, vol. 2018, p. 74, 2018.

[40] D. Gunning and D. W. Aha, "Darpa's explainable artificial intelligence program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[41] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, 2017.

[42] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[43] Z. R. Li, T. H. McCormick, and S. J. Clark, "Interva4: An R package to analyze verbal autopsy data," 2014.

[44] S. J. Clark, T. H. McCormick, Z. R. Li, and J. C. Wakefield, "InSilicoVA: A method to automate cause of death assignment for verbal autopsy," *arXiv preprint arXiv:1504.02129*, 2015.

[45] S. Jeblee, M. Gomes, and G. Hirst, "Multi-task learning for interpretable cause of death classification using key phrase prediction," in *Proceedings of the BioNLP 2018 workshop*, 2018, pp. 12–17.

[46] L. Aleksandrowicz, V. Malhotra, R. Dikshit, P. C. Gupta, R. Kumar, J. Sheth, S. K. Rathi, W. Suraweera, P. Miasnikof, R. Jotkar *et al.*, "Performance criteria for verbal autopsy-based systems to estimate national causes of death: development and application to the indian million death study," *BMC medicine*, vol. 12, no. 1, p. 21, 2014.

[47] K. Kahn, M. A. Collinson, F. X. Gómez-Olivé, O. Mokoena, R. Twine, P. Mee, S. A. Afolabi, B. D. Clark, C. W. Kabudula, A. Khosa *et al.*, "Profile: Agincourt health and socio-demographic surveillance system," *International journal of epidemiology*, vol. 41, no. 4, pp. 988–1001, 2012.

[48] S. J. Clark, "A guide to comparing the performance of VA algorithms," *arXiv preprint arXiv:1802.07807*, 02 2018.

[49] S. S. Ganapathy, K. Y. Yi, M. A. Omar, M. F. M. Anuar, C. Jeevananthan, and C. Rao, "Validation of verbal autopsy: determination of cause of deaths in Malaysia 2013," *BMC public health*, vol. 17, no. 1, p. 653, 2017.

[50] H. T. Tran, H. P. Nguyen, S. M. Walker, P. S. Hill, and C. Rao, "Validation of verbal autopsy methods using hospital medical records: a case study in Vietnam," *BMC medical research methodology*, vol. 18, no. 1, p. 43, 2018.

[51] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.