DataSets: <https://www.who.int/publications/m/item/verbal-autopsy-standards-the-2016-who-verbal-autopsy-instrument>

**TITLE: USING MACHINE LEARNING TO IMPROVE THE QUALITY OF VERBAL AUTOPSY**

Presented by NKESHIMANA Carmel

Nelson Mandela African Institution of Science and Technologies

Embedded and Mobile System

Supervisor Dr. Neema Mduma (Ph.D.)

1. Question

* How is it boring to feel a verbal autopsy document?
* Which important features do we have in the verbal autopsy?
* How can we use a machine-learning algorithm to reduce the size of the document and steel get the same accuracy in our predictions of death?

1. Introduction

Many low and middle-income countries (LMIC) cannot track essential registration data, such as the causes of death, which are critical inputs to health and development decision-making[1].

In developing countries, if a death occurs away from health facilities, a field-worker interviews a relative of the deceased about the circumstances of the death; this verbal autopsy can be reviewed off-site[2]

The verbal autopsy (VA) instrument is suitable for routine use. The instrument is designed for all age groups, including maternal and perinatal deaths and deaths caused by injuries. However, it suffers multiple limitations including many similar questions which require a lot of time to conduct the interview and can also cause overfitting in the prediction of the cause of death. The main objective of this study will be to contribute to the ongoing effort to reduce the questionnaire in the coming tools.

Normally, a verbal autopsy is a document containing many columns (features). Those columns are filled one by one up to the end and sometimes there are some which are not filled. But it takes a lot of time for the health agent who does it. And there is information that is not necessary. That is why with this research, I want to resume the formula and remain with the only important columns, and with the remaining ones in the dataset, I will still be able to predict the cause of death with the same accuracy as a normal dataset with all columns combined. This study will improve the timeliness of data collection and the accuracy of predictive models.

1. Literature review

Samuel Danso et al. uses Natural Language Processing (NLP) to try to forecast the cause of death by extending various computer methodologies[1].

The support vector machine approach was found to be the best performing algorithm and the most suitable for vocal autopsy text (Samuel Danso et al.). However, when exploring binary feature representation, Nave Bayes outperformed the support vector machine and Random Forest, which may be appropriate for data with restricted vocabulary size, such as the VA closed section. The experiment also demonstrates that using a locally semi-supervised strategy to feature reduction improved accuracy significantly [2].

Abraham D. Flaxman et al. demonstrate that for adult and child Verbal Autopsy with and without health care experience and neonatal Verbal Autopsy without health care experience, their random forest approach beat the physical-certified verbal autopsy method in terms of chance-corrected concordance and cause-specific mortality fraction accuracy. In terms of time and cost, it is also superior to a physical-certified verbal [3].

Machine learning (ML) techniques have been used extensively in the domain of public health. S Idicula et al. used Random Forest (RF) Method to forecast death, assessed and validated their results at the individual level using linear regression[4].

S. Idicula-Thomas et al. built models for the prediction of death from verbal autopsy by using six prominent ML-based algorithms: support vector machine, gradient boosting modeling, C5.0, artificial neural network, nearest neighbor, classification, and regression tree[4].

ST Green et al. combine corpus linguistics and natural language processing with machine learning approaches to predict death caused by verbal autopsies[5].

Using a narrative-based machine learning classifier performs as well as classifiers based on structured data at the individual level showed by Serena Jeblee et al. [6].

Alberto Blanco et al. demonstrated that when using Bayesian models with their models, the method that provided the best overall performance was BiGru and that this approach was competitive as an automatic classification system and also as a decision support system if they allowed the expert to choose the cause of death from a ranked list [7].

Conclusion

This proposed research will especially cover the verbal autopsy document and improve the time used to complete or register a new one. The data collection will not take much time because they will fill the only important and precious column with data.

References

[1] S. Danso *et al.*, “A verbal autopsy corpus for machine learning of cause of death.”

[2] S. Danso, E. Atwell, and O. Johnson, “A Comparative Study of Machine Learning Methods for Verbal Autopsy Text Classification.”

[3] A. D. Flaxman, A. Vahdatpour, S. Green, S. L. James, and C. J. L. Murray, “Random forests for verbal autopsy analysis: Multisite validation study using clinical diagnostic gold standards,” *Population Health Metrics*, vol. 9, Aug. 2011, doi: 10.1186/1478-7954-9-29.

[4] S. Idicula-Thomas, U. Gawde, and P. Jha, “Comparison of machine learning algorithms applied to symptoms to determine infectious causes of death in children: national survey of 18,000 verbal autopsies in the Million Death Study in India,” *BMC Public Health*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12889-021-11829-y.

[5] S. T. Green and A. D. Flaxman, “Machine Learning Methods for Verbal Autopsy in Developing Countries,” 2010. [Online]. Available: www.aaai.org

[6] S. Jeblee, M. Gomes, P. Jha, F. Rudzicz, and G. Hirst, “Automatically determining cause of death from verbal autopsy narratives,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, Jul. 2019, doi: 10.1186/S12911-019-0841-9.

[7] A. Blanco, A. Perez, A. Casillas, and D. Cobos, “Extracting cause of death from verbal autopsy with deep learning interpretable methods,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1315–1325, Apr. 2021, doi: 10.1109/JBHI.2020.3005769.