



UNIVERSITY OF LEEDS

This is a repository copy of *Linguistic and statistically derived features for cause of death prediction from verbal autopsy text*.

White Rose Research Online URL for this paper:  
<http://eprints.whiterose.ac.uk/89830/>

Version: Accepted Version

---

**Proceedings Paper:**

Danso, S, Atwell, E [orcid.org/0000-0001-9395-3764](http://orcid.org/0000-0001-9395-3764) and Johnson, O (2013) Linguistic and statistically derived features for cause of death prediction from verbal autopsy text. In: Gurevych, I, Biemann, C and Zesch, T, (eds.) Language Processing and Knowledge in the Web. 25th International Conference, GSCL, 25-27 Sep 2013, Darmstadt, Germany. Lecture Notes in Artificial Intelligence (8105). Springer , pp. 47-60. ISBN 978-3-642-40721-5

[https://doi.org/10.1007/978-3-642-40722-2\\_5](https://doi.org/10.1007/978-3-642-40722-2_5)

---

© 2013, Springer-Verlag Berlin Heidelberg. This is an author produced version of a paper published in Language Processing and Knowledge in the Web. The final publication is available at Springer via [http://dx.doi.org/10.1007/978-3-642-40722-2\\_5](http://dx.doi.org/10.1007/978-3-642-40722-2_5). Uploaded in accordance with the publisher's self-archiving policy.

**Reuse**

Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.



[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>

# Linguistic and statistically derived features for Cause of Death prediction from Verbal Autopsy text

Samuel Danso<sup>1,2</sup>, Eric Atwell<sup>1,2</sup>, Owen Johnson<sup>2</sup>

<sup>1</sup>School of Computing, Language Research Group, University of Leeds

<sup>2</sup>Yorkshire Centre for Health Informatics, eHealth Research Group, University of Leeds  
{scsod, eric.atwell, owen.johnson}@leeds.ac.uk

**Abstract.** Automatic Text Classification (ATC) is an emerging technology with economic importance given the unprecedented growth of text data. This paper reports on work in progress to develop methods for predicting Cause of Death from Verbal Autopsy (VA) documents recommended for use in low-income countries by the World Health Organisation. VA documents contain both coded data and open narrative. The task is formulated as a Text Classification problem and explores various combinations of linguistic and statistical approaches to determine how these may improve on the standard bag-of-words approach using a dataset of over 6400 VA documents that were manually annotated with cause of death. We demonstrate that a significant improvement of prediction accuracy can be obtained through a novel combination of statistical and linguistic features derived from the VA text. The paper explores the methods by which ATC may lead to improved accuracy in Cause of Death prediction.

**Keywords:** Verbal Autopsy, Cause of Death Prediction, Features, Text Classification

## 1 Introduction

Not all deaths that occur annually are medically certified with Cause of Death (CoD). It is estimated that about 67 percent of the 57 million deaths that occur annually are not medically certified due to weak or negligible death registration systems, predominantly in low income countries [1]. Information about CoD is a means to revealing preventable illness; developing health interventions; and research for treatment of diseases [2]. In low income countries there is pressure to find cost effective but still accurate CoD information and the Verbal Autopsy technique is frequently employed to do this [1].

The Verbal Autopsy (VA) technique is now well established in a large number of low income countries and generally follows the same pattern. It involves interviewing individuals (such as relatives or caregivers) who were close to the deceased, and if possible, those who cared for the individual around the time of death, in order to document events that may have led to the individual's death. The interviews are captured

on a standard questionnaire or document that is then sent for analysis by physicians who agree on a Cause of Death (CoD) classification based on the World Health Organisation (WHO) International Classification of Diseases (ICD) coding standards. It is worth noting that the VA interview is carried out in local languages of the countries in which they are employed, translated into English and transcribed onto the VA document for physicians to review.

Automatic prediction of CoD from the VA documents presents a number of benefits over the traditional manual physician based approach which is characterised by several limitations: high cost; intra-physician reliability; repeatability; inefficiencies and time consuming which automatic approaches may help overcome[3]. The potential benefits to be derived from the automated approaches to VA analysis and classification of CoD are attracting research interest [3, 4]. The VA document captures information of responses to both closed questions and open questions that record a narrative history. However, the automatic approaches published so far have only made use of the closed question responses [5, 6], while physicians have access to and make use of both the closed question responses and the open narrative in order to complete their diagnosis [7].

Our research is motivated by the hypothesis that computational algorithms that can take into account information obtained from the VA open narrative may lead to an improved prediction accuracy, which may in turn contribute to the United Nations' Millennium Development Goals. The research has been formulated as a Text Classification problem and classifies the VA according to CoD categories. We have employed Computational Linguistics (CL) and Machine Learning approaches to identify various features to be able to classify VA documents. We want to explore whether information derived from the open narrative leads to an improved CoD classification accuracy over either the closed question responses or a combination of both. Our literature review and experience in the field indicates that this is the first research that seeks to explore this approach.

Classification of biomedical documents is witnessing a high rate of growth in research in the applications of CL technology [8-11]. Cohen [11] for example employed chi-square as a statistical technique to extract features for a Support Vector Machine algorithm to classify genomes in biomedical text. Pakhomov et al [8] also employed various Text Classification based approaches to develop predictive models that identify patients with risk of heart failures from clinical notes obtained from Electronic Health Records.

The studies mentioned above have mainly explored the data originated from the formal environmental settings of the biomedical domain, where use of language is standardized with limited vocabulary. However, limited research has explored the informal settings where there are no specific rules but rather colloquial language is predominantly used. Nikfarjam and Gonzalez [12] and Leaman et al [13] are few researchers who have explored colloquial text within the biomedical domain. Nikfarjam and Gonzalez [12] employ CL approaches to automatically classify whether users experience adverse reactions of a given drug. Using data generated from DailyStrength ([www.dailystrength.org](http://www.dailystrength.org)), they employed association rules to extract patterns of colloquial expressions that correlate with adverse reactions. Their work is

largely motivated by the works in the area of automatic analysis and classification of sentiment and opinions, which are mostly expressed in colloquial text [14-17]. Pang et al [17] for example employed CL approaches to determine whether a sentiment expressed about a movie is positive, negative or neutral. Using various lexical and statistical features derived from a sample of movie review text, they demonstrate the possibility of using this approach with a comparable results obtained by humans.

Despite the emerging interest in research and numerous studies focused on automatic classification of colloquial text in general and specifically text from the biomedical, domain this has not been extended to VA text, which is argued by Danso et al [18] that the text should be considered a rather unusual subtype of biomedical genre. The next section gives a brief description of the VA dataset which we have used for our research and summarise the argument that VA it is distinct as a subtype of biomedical text.

## 2 Dataset

Danso et al[18] provide a detailed description of the dataset. In brief, unlike a standard biomedical text generated from a discourse either between a non-health professional and a health professional, or between health professionals, the VA text is a transcription of a discourse between two non-health professionals, written for a health professional (usually a medical doctor) to review. The dataset contains a total of over 11,700 VA documents of stillbirth, infants and women of reproductive age. It was collected from Kintampo in Ghana as part of a multi-year, single centre study between the year 2001 and 2010, and funded by the United Kingdom’s Department for International Development. The VA text in this instance are electronic version of the interviews that were conducted in a local language called Twi, translated into English and transcribed onto the VA document by the non-medically trained interviewer. The dataset also contains the corresponding closed ended responses to each of the open narratives.

Figure 1 and 2 below are shown to demonstrate the difference that exist between the closed response and open narrative as found in a typical VA document.

Did your child have fever?	1	2	8	For how long? (in days)		
				[99=NA]		

**Fig. 1.** Question and response options provided to respondent during interview

Can you tell me something about your pregnancy?
Movement of the baby in the womb started around the 6th month continuously till 9th month following the delivery. Although I did not encounter too many pregnancy complications, malaria persistently attacked me on the 7th month until I delivered. I suffered severely from anaemia which was diagnosed by a health worker when i visited hospital on the 8th month. Finally, I was not able to feed by self well when about a month to delivery due to lost of appetite. Sometime instead of feeding thrice a day, once daily becomes a problem for me.
Can you tell me something about your labour?
the labour started around 1pm in the night following the flow of water approximately 4hours. All of a sudden I felt the baby coming therefore I decided to try my best as much as possible to deliver at home. To my surprise the baby came with her both legs which really made it difficult to deliver myself. Therefore the TBA in the village was called to assist yet it proved futile. thus my husband had to go and arrange for vehicle to take me to the nearest hospital facility remarked by the TBA. before the vehicle arrived i had finally delivered.
Can you tell me something about the baby?
the baby landed without breathing or crying, therefore I enquired from the TBA to know what has happened to my baby but the woman assured me that the child is weak so I should lie down for a while and feel comfortable for everything will be alright. after she had finished with me she confirmed the baby landed dead.
Can you tell me what happened after delivery?
the baby neither cried or nor breath after delivery
Any signs and symptoms before the death of the child?
since the baby was very weak, he was put in an incubator but died after three hours of birth.

**Fig. 2.** A sample of open narrative question and responses from Infant Verbal Autopsy questionnaire

The above figures represent how both the open narrative and coded VA information are recorded. The open narrative is a verbatim account of the interview translated and transcribed by the interviewer and subsequently digitised. Figure 1 indicates the questions asked by the interviewer and the various options provided for the closed responses have the following meanings: 1="Yes", 2 = "No", 8 = "not known" and 99 = "Not application". The issues that characterised the text of the discourse of VA shown in figure 1 above are as catalogued in Table 1 below.

**Table 1.** categories of issues with the VA dataset

Type of issue	example
A non-standard grammatical and spelling errors	"Before labour waters, which look clear and without bad scent" "... she fell sick, which lauted for three days.."
Colloquial forms in expressing concepts	Baby came out Baby landed Gave birth } Delivery
Use of local terms to describe medical conditions	Asram, Anidane

A non-standard and fuzzy expressions of medical concepts	“I visited xxx hospital on Tuesday and was given one bottle of water. . .”
Abbreviations and acronyms	TBA = Traditional Birth Attendant ANC = Antenatal care . CS = caesarean session
Inappropriate use of punctuation marks	“Any time, she breaths, you see a hole”

For this paper we report on the use of a subset of 6407 infant (stillbirths included) VA documents, which is approximately 1.6 million words in total, taken from the full dataset and used as the basis for the experiments being reported here. Each document in this subset has already been annotated by a minimum of two physicians and the final agreed CoD classification assigned. Where there is a disagreement between the two physicians, a third physician is employed to decide on the final CoD. There are two separate features of the CoD to be categorised: Time-of-Death has five categories and Type-of-Death consists of 16 categories as detailed in Tables 2 and 3 below.

**Table 2.** breakdown of Time-of-Death categories

Time of Death	% distribution
Neonatal	31.3
Antepartum_stillbirth	21.5
PostNeonatal	19.1
Intrapartum_stillbirth	15.6
Non_stillbirth_unknown_cause	12.5

**Table 3.** breakdown of Type-of-Death categories

Type of Death	% distribution
Stillbirth- unexplained	22.1
Cause unknown	12.5
Birth asphyxia	10.9
Neonatal Infection	8.7
Stillbirth-obstetric complications	7.5
PostNeonatal - Other Infections	6.9
Neonatal - other causes	5.9
Prematurity	5.8
Pneumonia	5.6
Malaria	4.3
Stillbirth- maternal disease	3.2
Diarrhoea	2.4
stillbirth- maternal haemorrhage	1.9
Stillbirth - other causes	1.5
stillbirth-congenital abnormalities	0.5
Measles	0.1

The issues outlined above characterise a dataset with high level of sparseness and lexical diversity [18]. To further demonstrate the relative noisy nature of the VA text, Danso et al [18] selected a sample of the VA text, which was used to evaluate the accuracy of a PoS tagger that was trained using the Brown corpus [19]. The evaluation of the performance of the PoS tagger carried out by a linguistic expert suggested an accuracy of 88 percent, which is a clear departure from the expected accuracy of about 96 - 97 percent from a normal English text [18]. Additionally, there are also issues about imbalance across CoD categories as shown in Tables 2 and 3 above. These issues present various levels of challenges for Computational Linguistics and Machine Learning based approaches to Text Classification and employing standard techniques in the area of biomedical Text Classification may not be produce desirable results.

### **3 Methodology**

A brief description the methods employed in predicting CoD from VA is given here.

#### **3.1 Evaluation**

We employ the standard Precision, Recall and F1- measure metrics[20] to evaluate the performance of the classification methods against the physician CoD classification as a gold standard. Macro-averages as opposed to Micro-averages are used since Macro-averages tend to be suitable for highly skewed multi-class dataset which allows equal weights to be computed for each CoD category [21]. All averages are obtained based on 10 fold cross-validation [22].

#### **3.2 Pre-processing**

The texts were converted to lower case and tokenized by whitespaces. All punctuations were removed. Our initial exploration of the feature space pointed to the fact that removing the standard English stop-words affects tend to adversely affect the performance accuracy which falls below baseline of bag of words. All words were used in their natural form as they appeared and no further processing such as spell checking and corrections were carried out. With regards to the responses to closed questions part of the dataset, all information pertaining to symptoms, history of cares sought and treatment were extracted and separately stored. These were further discretised to ensure each category of response was appropriately used and not treated as a numeric value for the feature which has implications for the machine learning algorithm. For example question “did your child have fever” as indicated in figure 1 above, has three options 1=“Yes”, 2 = “No” and 8= “don’t now” of responses and the numeric values are captured.

### 3.3 Classification algorithm

Danso et al [23] previously showed in an experiment aimed at exploring the VA space in order to identify the suitable algorithm for this task. The results suggest Support Vector Machine (SVM) as the most suitable Machine Learning algorithm. The Sequential Minimal Optimisation (SMO) algorithm, which is an implementation of SVM in WEKA Machine Learning software[24], was therefore used in this experiment.

### 3.4 Features for classification

Danso et al [23] explored a baseline bag-of-words and how feature values must be represented for a classification algorithm. The results suggest a Normalised Term Frequency value as the best feature value representation. Normalized Term Frequency value was computed as the frequency of a given term, divided by the total number of terms in a given document. The experiments in this paper therefore employed the same scheme of representation. The various features that were explored in this experiment are as outlined as detailed in the table 4 below.

**Table 4.** list of features explored in this experiment

label	Features
A	Unigram (bag-of-words)
B	Unigrams + PoS Trigrams
C	Unigrams + PoS Trigrams + Relative Word positions
D	Unigrams + PoS Trigrams + Relative Word positions + Noun Phrases
E	Unigrams + PoS Trigrams + Relative Word positions + Noun Phrases + collocation bigrams (top 1 collocates )
F	Unigrams + PoS Trigrams + Relative Word positions + Noun Phrases + collocation bigrams (top 1 collocates + collocation bigrams (top 2 collocates )
G	Unigrams + PoS Trigrams + Relative Word positions + Noun Phrases + collocation bigrams (top 1 collocates + collocation bigrams (top 2 collocates )+ collocation bigrams (top 3 collocates )
H	Closed response
J	Closed response + Unigrams + PoS Trigrams + Relative Word positions + Noun Phrases + collocation bigrams (top 1 collocates + collocation bigrams (top 2 collocates )+ collocation bigrams (top 3 collocates )

Our motivations for employing the above features in our experiments are explained in turn below.



### 3.4.1 Linguistics features

Building on the results obtained from the lexical features (bag-of-words) the following linguistic features have also been explored.

Part of Speech (PoS) information obtained from a part of speech tagger have been employed as features in several Text Classification works. This approach is considered as a crude form of determining the correct sense of a given word in a text [25]. The PoS information were obtained using the Natural Language Processing Tool Kit's PoS tagger [26] trained using the Brown corpus [19] to tag the Verbal Autopsy dataset.

Part of Speech Trigrams: PoS tags have have been shown to be useful feature in numerous Text Classification problems. Gamon [14]for example, demonstrated the use of PoS trigrams in sentiment classification. We however explored various PoS tags in our experiment which include unigram, bigrams and trigrams of which PoS trigrams was found to be the best and results presented here.

Relative Word Position: our motivation to experiment this feature is based on one of the criticism of the famous bag-of-words approach to Text Classification that bag-of-words approach tends to ignore the order and syntactic relations between the words that occur in a sentence [27]. The adaption of the relative positions of words in text as possible features for Text Classification has been explored by various researchers [28, 29. Matsumoto et al [29] for example, demonstrate the usefulness of this feature by extracting word sub-sequences and dependency sub-trees from sentences to classify movie reviews. We however adapt a simplified approach by exploring the relative positions of the words as they appear in the text to capture the sequential order of events within the context of the VA.

Our approach treats the entire content of a VA document as a single string of words with an imaginary grid, where each cell represents a word which is a member of the string. Each cell is serially allocated a unique number and that represents the position of the word with respect to the entire string. The position number of the word captured is divided by the number of the string (number of cells) to obtain the relative word position with respect to the entire words in the VA document. The hypothesis here is that there may be a logical order of event in the history that led to the death of an individual, which may be a major factor in case profiling in an investigation process, and this feature may help in capturing this order. This is illustrated by an example, which is taken from a VA document.

*“In the second month of the pregnancy, ...labour started which I was at home in the morning...”*

If the order of these words is ignored one possible reading could be

“In the second month labour started....”

and this presents different scenarios and may mean a different outcome from medical perspective. The proposed relative word position features is to avoid this situation and preserve the in which the words appear in the VA document.

Noun Phrases: having obtained PoS information for every word in the text, chunking techniques as implemented by the regular expression below was used to extract noun phrases:

```
r'''NounPhrase: {<DT>?<JJ>*<NN>*}
```

The decision to explore these features was inspired from the fact that domain concepts are mostly expressed using multiword structures [30]. For example “a normal labour” was used to describe a type of labour a mother experienced during pregnancy, which is domain specific information. A generalised approach to capturing these types of mentions in the text is through extraction of noun phrases, which are derived from the PoS tagset and were represented as single terms.

### 5.3.3 Statistically derived collocation as features

Statistically derived features are considered to be some form of phenomenon that tend to occur in the use of a language but that are not predictable. As observed by [31] that “each word has a particular and roughly stable likelihood of occurring as argument, or operator, with a given word, though there are many cases of uncertainty disagreement among speakers, and change through time”. Collocation is one technique that can be used to capture the phenomenon described above. Collocation can be employed to capture word-pairs and phrases that frequently occur in the use of a language with no regard to their semantic or syntactic rules of use; and are also known to be dialect or language specific[32]. Collocation has been employed in many applications by lexicographers to carry out word sense disambiguation and semantic analysis of text[33]. This therefore suggests an imperative investigation into the potential use of collocation as a feature to identify patterns of co-occurrence of words that could as indicative of phrase or an expressions of CoD considering the peculiar nature of colloquial text contained in the VA corpus.

We have employed statistical methods based on log-likelihood estimation to determine the likelihood of co-occurrence words and phrases [34]. The log-likelihood estimation was based on the entire corpus and estimated the likelihood of two words co-occurring as defined by bigram log-likelihood statistics association measure[35]. The limitation associated with this approach however is that it usually take into account the only two word-collocate (bigrams) that co- occur in the corpus [36]. To address this limitation, we explored the levels of associations observed from the corpus as ranked by the bigram log-likelihood statistics association measure algo-

rithm. The topmost collocation bigrams were experimented in turns and their impacts on performance were obtained. We illustrate our idea with the following example.

during | 'labour'= 4150,'pregnancy'= 2901 'my' = 1785

In the example shown above the word during, which is mentioned in a given VA document retrieves the three words with the strongest association with their corresponding likelihood values as ranked by the algorithms. These words are retrieved and added as part of the feature set.

#### 5.3.4 Combined feature set

In order to explore whether the information obtained from both closed response and open narratives parts could improve classification of CoD in VA, all features derived from both parts were combined.

## 4 Results:

The results presented here are based on the features described above. Many combinations were explored but for brevity only the best performing combinations are presented here. To give a better perspective of the results given here, we give the result obtained by a simple majority baseline algorithm ZeroR in WEKA as captured in table 4 below. We differentiate between time-of death and type-of-death feature labels by adding (1) and (2) to the respective labels.

**Table 5.** baseline results from a simple majority

	Category	Precision	Recall	F-measure
O <sub>1</sub>	Time-of-Death	0.098	0.313	0.149
O <sub>2</sub>	Type-of-Death	0.049	0.221	0.08

**Table 6.** Results from various feature sets - Time of Death categories

	Precision	Recall	F-measure
A <sub>1</sub>	0.414	0.434	0.416
B <sub>1</sub>	0.473	0.428	0.339
C <sub>1</sub>	0.56	0.59	0.517
D <sub>1</sub>	0.56	0.59	0.517
E <sub>1</sub>	0.613	0.599	0.559
F <sub>1</sub>	0.637	0.618	0.559
G <sub>1</sub>	0.643	0.629	0.582

**Table 7.** Results from various feature sets- Type-of Death categories

	Precision	Recall	F-measure
A <sub>2</sub>	0.248	0.288	0.251
B <sub>2</sub>	0.22	0.256	0.142
C <sub>2</sub>	0.314	0.376	0.285
D <sub>2</sub>	0.314	0.376	0.285
E <sub>2</sub>	0.33	0.391	0.304
F <sub>2</sub>	0.311	0.395	0.306
G <sub>2</sub>	0.35	0.406	0.322

**Table 8.** Results from various closed and narrative combined - Time of Death categories

	Precision	Recall	F-measure
H <sub>1</sub>	0.826	0.836	0.827
J <sub>1</sub>	0.826	0.835	0.828

**Table 9.** Results from various closed and narrative combined - Type of Death categories

	Precision	Recall	F-measure
H <sub>2</sub>	0.575	0.616	0.583
J <sub>2</sub>	0.591	0.616	0.587

## 5 Discussion and future work

The meaning of a word is best known by the context in which it exists and this was evident in the results obtained from these experiments. Using various linguistic and statically derived features which have extra information about the individual words has shown that there is significant increase in performance accuracy over the single words (bag-of-words) in predicting of CoD. Notable among these are our novel collocation and relative word positions features introduced. There were also marginal gains in terms of precision obtained from the PoS trigrams feature set. One feature set, which was however not found useful, is the noun phrase as shown in experiment G<sub>1</sub> and G<sub>2</sub> in table 6 and 7 respectively. The result remained unchanged when noun phrases were introduced. This is surprising considering our hypothesis about the potential usefulness of noun phrases in capturing multiword concepts. The result is however consistent with the findings in the literature that noun phrases tend not be useful features in ATC29] and require further investigations.

The results obtained from the combined set of features from both closed and open parts in Table 8 and 9 tend to suggest that the closed response part achieves better performance accuracy than the narrative part. The marginal gains in accuracy from the combined set in experiments J<sub>1</sub> and J<sub>2</sub> also tend to suggest that there is a marginal benefit in predicting CoD using a combination of the narrative and the closed responses. A detailed examination of the output at the individual level of CoD however suggests some benefits in combining information from both closed and narrative text.

An example is ‘intra-partum stillbirth’ category where F-measure values of 0.36, 0.49 and 0.87 were recorded for narrative, closed and combined respectively suggesting the close response missed some relevant information from available in the narrative.

The relatively low F-measure recorded particularly for the Type-of-Death categories in either or both narrative and closed response parts of the VA suggest:

(i) Data skew: few samples for many CoDs, making them hard to classify with either or both of closed and open parts of the data; and

(ii) Complexity of diagnosis: some of the examples suggest that CoD is determined by the physicians using a complex combination of information from the overall VA document, and this is hard to capture in simplistic models used in Machine Learning classifiers.

However, considering the noisy and rather an unusual type of text being dealt with, there is the possibility that the features employed so far may not be effective enough in discriminating between CoDs. There is therefore the need for further exploration within the feature space of the narratives in order to increase the performance accuracies obtained. This may include adaptation of the standard PoS taggers for this particular type of text. This was clearly demonstrated by the PoS tagging experimental results obtained by Danso et al [18] as indicated in section 2. It may however be argued that the choice of Brown corpus for training the PoS tagger was inappropriate considering the difference in text, which may have resulted in the poor performance of the linguistic features extracted from the output of the PoS tagger. It must be pointed out that the choice was purely based on convenience as there was no linguistically annotated corpus readily available for both the VA and the biomedical domain. Future work could therefore explore the possibilities of training the PoS tagger with corpus that has linguistic annotations from either the VA or biomedical domain. Future work will also explore features that will be targeted at the minority categories to increase the potential of improving the performance accuracies for these categories.

We believe this work may help reduce the cost and increase the accuracy of predicting CoD from VAs and therefore addresses vital global health challenges that confront developing countries in particular and the WHO at large.

### **Acknowledgement**

We thank Professor Betty Kirkwood of the London School of Hygiene and Tropical Medicine and the entire members of the trial management team of the ObaapaVita and Newhints projects that generated the dataset used in carrying out this study. We furthermore thank the Director and Staff of Kintampo Health Research Centre, especially the head, Mr Seeba Amenga-Etego, and staff of the computer centre, who worked around the clock to get this corpus digitised. Our appreciation also goes to Centre for International Health and Development, University College London Institute of Child Health for their nomination of an award of a scholarship and the Commonwealth Scholarship Commission for their funding support for this research.

## References:

1. World Health Organization, Geneva: World Health Organization. WHO Handbook for Reporting Results of Cancer Treatments (WHO Offset Publication No. 48), 2004.
2. Kahn, K., Tollman, S.M Garenne, M and Gear, J.S, Validation and application of verbal autopsies in a rural area of South Africa. *Tropical Medicine & International Health*, 2000. **5**(11): p. 824-831.
3. Byass, P, Kathleen K, Edward F, Mark A. C and Stephen M. T. Moving from Data on Deaths to Public Health Policy in Agincourt, South Africa: Approaches to Analysing and Understanding Verbal Autopsy Findings. *PLoS Medicine*, 2010. **7**(8).
4. King, G., Y. Lu, and K. Shibuya, Designing verbal autopsy studies. *Population Health Metrics*, 2010. **8**(1): p. 19.
5. Byass, P, Edward F, Dao Lan H, Yamene B, Tumani C, Kathleen K, Lulu M and Do Duc Van, Refining a probabilistic model for interpreting verbal autopsy data. *Scandinavian Journal of Public Health*, 2006. **34**(1): p. 26-31.
6. Murray, C J L, Alan D L, Dennis F, Shannon T P and Gonghuan Y, Validation of the symptom pattern method for analyzing verbal autopsy data. *PLOS Medicine*, 2007. **4**: p. 1739 - 1753.
7. Soleman, N., D. Chandramohan, and K. Shibuya, WHO Technical Consultation on Verbal Autopsy Tools. Geneva, 2005.
8. Pakhomov, S., Shah, N., Hanson, P., Balasubramaniam, S., and Smith, S. A. Automatic quality of life prediction using electronic medical records. 2008. American Medical Informatics Association.
9. Pakhomov, S., Weston S.A, Jacobsen, S.J, Chute C.G, Meverden, R., and Roger, V. L. Electronic medical records for clinical research: application to the identification of heart failure. *The American journal of managed care*, 2007. **13**(6 Part 1): p. 281.
10. Cohen, A.M. and W.R. Hersh, A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 2005. **6**(1): p. 57-71.
11. Cohen, A.M. An effective general purpose approach for automated biomedical document classification. in *AMIA Annual Symposium Proceedings*. 2006. American Medical Informatics Association.
12. Nikfarjam, A. and G.H. Gonzalez. Pattern mining for extraction of mentions of adverse drug reactions from user comments. in *AMIA Annual Symposium Proceedings*. 2011. American Medical Informatics Association.
13. Leaman, R Wojtulewicz, L. Sullivan, R. Skariah, A., Yang J., and Gonzalez, G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. 2010. Association for Computational Linguistics,...
14. Gamon, M., Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis, in *Proceedings of the 20th international conference on Computational Linguistics 2004*, Association for Computational Linguistics: Geneva, Switzerland. p. 841.
15. Oberlander, J. and S. Nowson. Whose thumb is it anyway?: classifying author personality from weblog text. in *Proceedings of the COLING/ACL on Main conference poster sessions*. 2006. Association for Computational Linguistics.
16. Turney, P.D., Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics 2002*, Association for Computational Linguistics: Philadelphia, Pennsylvania. p. 417-424.

17. Pang, B. and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. in Annual Meeting-Association For Computational Linguistics. 2005.
18. Danso, S; Atwell, ES; Johnson, O; ten Asbroek, A; Soromekun, S; Edmond, K; Hurt, C; Hurt, L; Zandoh, C; Tawiah, C; Fenty, J; Etego, S; Agyei, S; Kirkwood, B. 2013. A semantically annotated Verbal Autopsy corpus for automatic analysis of cause of death. ICAME Journal of the International Computer Archive of Modern English, vol.37.[In press]
19. Francis, W.N. and H. Kucera, Brown corpus manual. Letters to the Editor, 1979. 5(2): p. 7.
20. Scott, S. and S. Matwin. Text classification using WordNet hypernyms. in Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference. 1998.
21. Forman, G. A pitfall and solution in multi-class feature selection for text classification. 2004. ACM.
22. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. 1995. Lawrence Erlbaum Associates Ltd.
23. Danso, S; Atwell, ES and Johnson, O. A Comparative Study of Machine Learning Methods for Verbal Autopsy Text Classification. International Journal of Computer Science Issues, vol.10.[In press].
24. Witten, I.H. and E. Frank, Data Mining: Practical machine learning tools and techniques 2005: Morgan Kaufmann.
25. Pang, B., L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. 2002. Association for Computational Linguistics.
26. Loper, E. and S. Bird, NLTK: the Natural Language Toolkit, in Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 12002, Association for Computational Linguistics: Philadelphia, Pennsylvania. p. 63-70.
27. Wilks, Y. and M. Stevenson. Word sense disambiguation using optimised combinations of knowledge sources. in Proceedings of the 17th international conference on Computational linguistics-Volume 2. 1998. Association for Computational Linguistics.
28. Moschitti, A. and R. Basili, Complex linguistic features for text classification: A comprehensive study. Advances in Information Retrieval, 2004: p. 181-196.
29. Matsumoto, S., H. Takamura, and M. Okumura, Sentiment classification using word sub-sequences and dependency sub-trees. Advances in Knowledge Discovery and Data Mining, 2005: p. 21-32.
30. Scott, S. and S. Matwin. Feature engineering for text classification. in Machine Learning-International Workshop Then Conference-. 1999.
31. Harris, Z.S., Methods in structural linguistics. 1951.
32. McKeown, K.R. and D.R. Radev, Collocations. Handbook of Natural Language Processing. Marcel Dekker, 2000.
33. Pearce, D. and B. Qh. Using conceptual similarity for collocation extraction. in Proceedings of the Fourth annual CLUK colloquium. 2001
34. Dunning, T., Accurate methods for the statistics of surprise and coincidence. Computational Linguistics., 1993. 19(1): p. 61-74.
35. Seretan, V., L. Nerima, and E. Wehrli. Extraction of multi-word collocations using syntactic bigram composition. in Proceedings of the Fourth International Conference on Recent Advances in NLP (RANLP-2003). 2003.

36. Pearce, D. A comparative evaluation of collocation extraction techniques. in Proceedings. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002). 2002.