Computer Science
UNIVERSITY OF TORONTO

NLP FOR CAUSE-OF-DEATH CLASSIFICATION IN HINDI

PARTH PARMAR, SERENA JEBLEE, GRAEME HIRST

BACKGROUND

The Million Death Study (MDS) is an ongoing study conducted in India to assign a probable cause to deaths that occurred outside of health facilities without any medical attention using verbal autopsy[1]. Verbal Autopsy is a method that helps determine probable causes of death (CoD) using written narratives recorded in the local language describing the events that preceded the death.

The primary goal of this project is to automatically determine CoD categories from free-text narratives written in Hindi by utilizing machine learning algorithms.

DATASET

Our main dataset comes from the Million Death Study Program. Since the majority of the the available records in MDS are scans of handwritten forms, we use a subset consisting of the records with narratives written in Hindi that have been manually transcribed using Latin characters into a digital format. Then, we automatically convert these narratives into Devanagari.

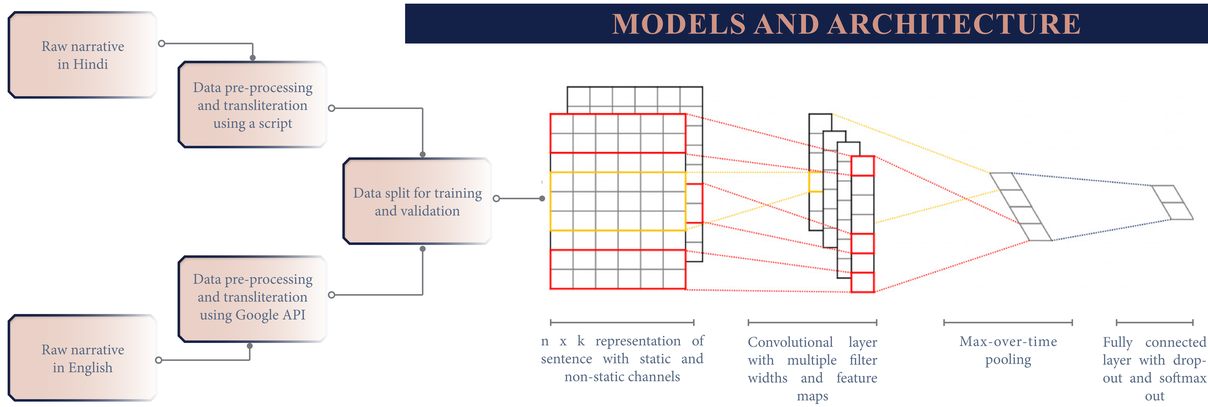
HINDI NARRATIVE IN LATIN

uttaradAtA kE anusAra mRRitaka kI kAphI umra hO chukI thI. usE pichalE takarIbana 15 varShOM sE “diabetes” kI bImArI thI. niyamita rUpa sE usakA IlAja va dawAI lEtI thI. usE dO bAra “mild heart attack” bhI AyA thA. IlAja bhI karawAyA “Admit” bhI rahIM magara Eka dina zOrO sE sInE mE darda hOnE lagA aura “3rd heart attack” A gayA jisakE kAraNa usakI ghara para hI mRRityu hO gayI.

HINDI NARRATIVE IN DEVANAGARI

उतरदाता के अनुसार मृतक की काफी उमर हो चुकी थी उसेपिचलेतकरीबन वरष से "diabetes" की बीमारी थी नियमित रप सेउसका ईलाज व दवाई लेती थी उसे दो बार "mild heart attack" भी आया था ईलाज भी करवा या "Admit" भी रहीअं मगर एक दिन सेसीनेमेदरद होनेलगा और "3rd heart attack" आ गया जिसके कारण उसकी घर पर ही मृतुहो गयी

MODELS AND ARCHITECTURE



The features that we used for CoD classification are word frequency counts for non-neural network based models and multi-dimensional vectors for neural network based models. Initially, we trained our classifiers using only 450 original Hindi data points. It was apparent that lack of training data lead to poor results from neural network based models.

In order to improve results, we translated 12,000 English narratives into Hindi using Google Translate API; which were then fed into a CNN implemented in PyTorch.

RESULTS

PERFORMANCE ON THE ORIGINAL HINDI DATA

Model	Precision	R	F1	PCCC	CSMF
SVM	0.440	0.389	0.365	0.346	0.758
RF	0.417	0.491	0.425	0.455	0.655
NB	0.313	0.423	0.337	0.382	0.603

PERFORMANCE ON THE ORIGINAL HINDI DATA + TRANSLATED DATA

Model	Precision	R	F1	PCCC	CSMF
CNN (only translated data)	0.725	0.733	0.721	0.717	0.914
CNN (translated data + original data)	0.657	0.616	0.618	0.591	0.883
SVM (translated data + original data)	0.421	0.361	0.365	0.318	0.803

DISCUSSION AND FUTURE WORK

- Medical symptoms are often described in local terms by the non - medical surveyors/ respondents. Moreover, different writing styles and inconsistent adherence to transliteration rules could lead to poor data quality.
- Some narratives are long and include background information that is not directly related to CoD; which caused some misclassification.
- As two different methods (script and Google Translate API) were used to convert narratives into Devanagari, it lead to poor performance.
- We would need to explore the effect of linguistic features (semantics, chronology) and alternative neural networks.

REFERENCES

1. Gomes M, Begum R, Sati P, Dikshit R, Gupta PC, Kumar R, et al. Nationwide Mortality Stud ies to Quantify 402 Causes of Death: Relevant Lessons from India’s Million Death Study. Health Affairs. 2017;36(11):1887–1895.
2. Jeblee S, Gomes M, Jha P, Rudzics F, Hirst G. Automatically determining the cause of death from verbal autopsy narratives
3. Britz D. Implementing a CNN for text classification in TensorFlow. Dec 11, 2015. Web.