Empowering Citizens. Smarter Societies.

Insight
Centre for Data Analytics

EMBC 2019:

Outlier Detection in Health Record Free-Text using Deep Learning

Duncan Wallace, Tahar Kechadi

A World Leading SFI Research Centre

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union

European Union
European Regional Development Fund

Science Foundation Ireland For what's next

NUI Galway
OÉ Gaillimh

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI)under Grant Number SFI/12/RC/2289, co-funded under the European Regional Development Fund

Background

- Out-of-Hours' Care (OOHC) provides medicinal services during periods when family doctors are not available
- Common health platform in many countries
- Large telemedical element

- Cases are treated independent of one another
- Cases are predominantly recorded as free text

Problem Statement

Insight
Centre for Data Analytics

- Prediction of high volume users.
- Medical and operational motivations for detecting these patients.
- Data predominantly consists of medical domain natural language.
- Previous tests had indicated that RNNs featuring LSTM would be suitable for this task.
- This research was designed to determine:
  - Whether preprocessing would improve results
  - If Word2Vec or GloVe provided better word embeddings
  - If GRUs presented better alternative to LSTM
  - If adding parameterised features aided classification

Preprocessing

- Traditional machine learning techniques typically, significantly benefit from preprocessing
- Use is not necessarily as clear cut in neural network application.
- Approach used took medical context into account - aimed to increase value of domain specific terms and reduce feature volume
- Methods used detailed in (Wallace, 2016, *Retrieval and Clustering of Medicines Within Healthcare Data Records*) and (Wallace, 2017, *Abbreviation and Acronym Identification and Expansion in Medical Records*)

Given maxalon im as just taken stemitil po an hour previous. Script given for po stemitil and also serc 8 mg tds PRN. caredoc over weekend if any change
given **{MT}** metoclopramide im as just taken **{MT}** prochlorperazine **{TA}** per oral an hour previous **[ST]** script given for **{TA}** per oral **{MT}** prochlorperazine and also **{MT}** betahistine morphine **[NT]** **{AA}** mgs **{AA}** tds **{AA}** prnx **[ST]** caredoc over weekend if any change

Medication identification →	**Medication correction** →	**Ontological transformation**
solbutamol	salbutamol	Salbutamol
salanol	salomol	Salbutamol
vemntolin	ventolin	Salbutamol

Word Embeddings

Insight
Centre for Data Analytics

- Pretrained Word2Vec model (using Google News) compared with models trained on corpus using both Word2Vec and GloVe

PERFORMANCE RELATING TO WORD EMBEDDING METHOD

	t(24) P		t(24) P'		t(50) P		t(50) P'	
	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV
W2V	0.66±0.018	0.72±0.044	0.76±0.034	0.74±0.04	0.81±0.013	0.82±0.01	0.85±0.016	0.85±0.028
GloVe	0.64±0.077	0.75±0.073	0.72±0.022	0.76±0.03	0.82±0.022	0.81±0.019	0.84±0.026	0.81±0.022
Pretrained	0.62±0.037	0.69±0.051	0.69±0.043	0.74±0.05	0.76±0.026	0.79±0.009	0.75±0.037	0.80±0.016

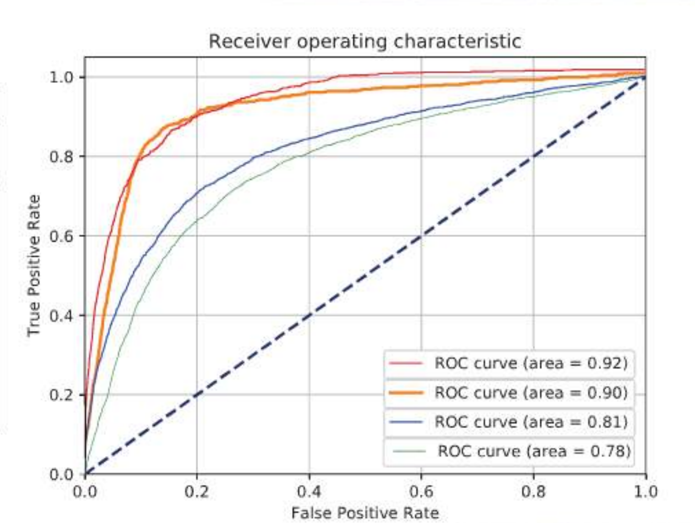
Classification

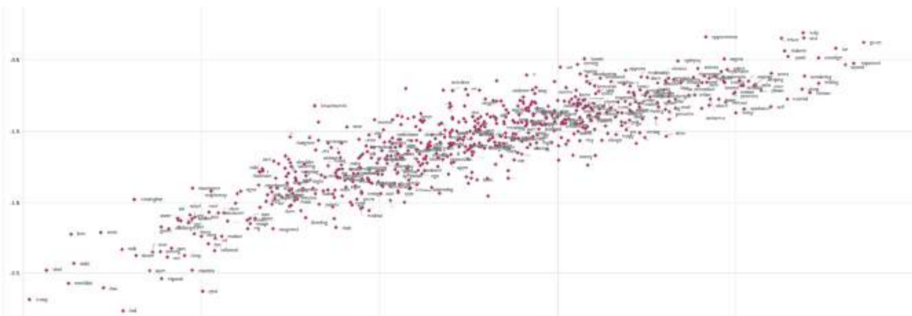
Insight
Centre for Data Analytics

- Classification compared LSTM and GRU usage in RNNs. Also a comparison between patient details (age, sex, etc.) added as features compared with noise input.

ARCHITECTURE AND INPUT COMPARISON

			LSTM				GRU			
			Features		Noise		Features		Noise	
T	Length	C	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV
24	100	100	0.71±0.006	0.76±0.005	0.75±0.033	0.75±0.047	0.69±0.067	0.76±0.01	0.74±0.035	0.74±0.035
24	100	200	0.73±0.021	0.75±0.003	0.72±0.047	0.76±0.013	0.74±0.009	0.75±0.015	0.72±0.011	0.73±0.022
24	200	100	0.71±0.018	0.72±0.012	0.54±0.21	0.89±0.081	0.7±0.014	0.74 ±0.008	0.75±0.015	0.74±0.01
24	200	200	0.78±0.098	0.63±0.21	0.38±0.33	0.91±0.042	0.78±0.083	0.63±0.21	0.67±0.16	0.82±0.007
50	100	100	0.88±0.009	0.78±0.008	0.86±0.012	0.88±0.014	0.89±0.015	0.8±0.013	0.86±0.023	0.83±0.027
50	100	200	0.85±0.017	0.83±0.026	0.88±0.019	0.84±0.025	0.78±0.018	0.89±0.011	0.81±0.044	0.86±0.041
50	200	100	0.80±0.083	0.69±0.23	0.79±0.015	0.71±0.17	0.80±0.019	0.65±0.26	0.8 ±0.32	0.83±0.063
50	200	200	0.89±0.048	0.75±0.165	0.82±0.14	0.72±0.21	0.83±0.22	0.77±0.08	0.78±0.26	0.62±0.052





Conclusion

Insight
Centre for Data Analytics

- LSTMs with Word2Vec proved best for our classification purposes
- But differences with GRUs and GloVe, respectively, could sometimes be quite small.
- Using a pretrained word embedding model provided inferior performance
- Adding noise to input layer for classification helped reduce overfitting, while adding demographic details showed little improvement.
- Classification of frequent users may help identify patients requiring specialised treatment.