Extracting Medical Data From Unstructured Clinical Texts

Shravya Gurrapu, Neha Srivathsa, Roshni Thachil

Introduction

- Medical professionals take significant time and effort to keep updated Electronic Medical Records (EMRs).
- Finding an automated way to enter and categorize patient data would make medical visits more efficient and patient-focused.
- If we assume that we can audio record and transcribe doctor-patient interactions, it would be extremely useful to be able to categorize each sentence into the section that it would fall under in the EMR.

Data Preprocessing

- Through Kaggle, we located a dataset containing ~2000 transcripts of hypothetical patient summaries, spoken by doctors.
- We separated the data by sentence, removed stop words, removed meaningless text, tokenized, and standardized the case.
- 2000 sentences were hand-labeled for our training set into the categories of 'Patient History', 'Diagnosis', 'Treatment', or 'Other', allowing for multiple labels per sentence.

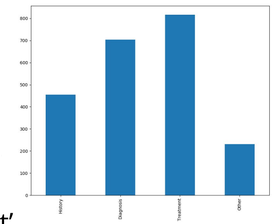


Figure 1. Number of sentences in each category after manual labelling

K-means Attempt

- To see whether sentences could self-cluster into defined categories, we attempted to use K-means Clustering.
- However, upon assignment of 4, 7, and 20 different clusters, there were no clear trends within clusters.

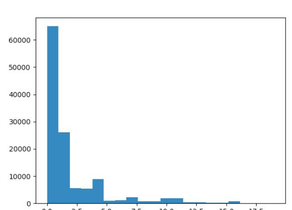


Figure 2. Number of sentences in each cluster for K = 20.

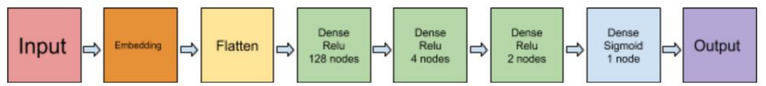
Classification Models

Logistic Regression

Using Keras embedding layer and Stochastic Gradient Descent

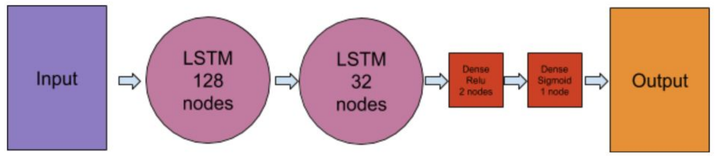
Test Accuracy: Patient History: .7717, Diagnosis: .6849, Treatment: .5831

Feed-Forward Neural Network



Test Accuracy: Patient History: .8784, Diagnosis: .7370, Treatment: .5831

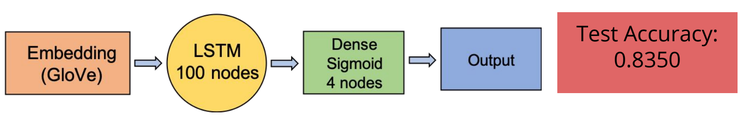
Recurrent Neural Network with LSTMs



Test Accuracy: Patient History: .8287, Diagnosis: .7816, Treatment: .8635

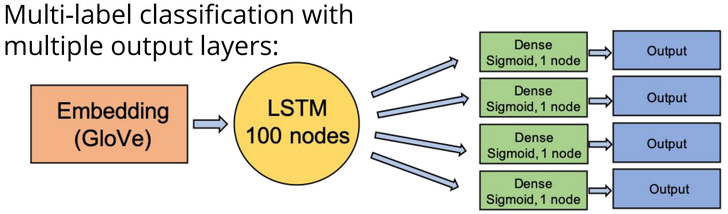
Multi-Label Classification

Multi-label classification with single output layer:



Test Accuracy: 0.8350

Multi-label classification with multiple output layers:



Test Accuracy: Patient History: .6271, Diagnosis: .6298, Treatment: .6368

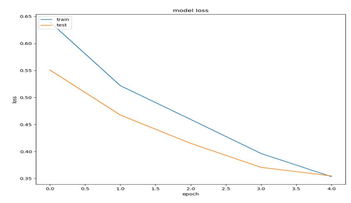


Figure 3, 4. Single output model loss and accuracy for multiclass model

Conclusions

- Logistic regression is not ideal for data with several non-linear features and elements.
- Feed forward neural networks allows for more complicated nonlinear associations and computations, better classifying our data.
- LSTMs have an overall edge because of their property of selectively remembering patterns for long durations of time.
- Multi-label classification methods allow us to predict multiple categories for a single sentence, taking into account associations between the categories.

Future Directions

- Overall, neural networks, specifically LSTMs, can be used to classify medical text into useful fields for EMRs.
- We should continue to design and train such complex models for such datasets.
- We should increase the size and variety of our training data and introduce clinical interviews with both physician and patient texts to continue tackling this problem.

References

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *JMLR workshop and conference proceedings*, *56*, 301–318.
Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*. doi: 10.1093/jamia/ocw112
Wei, Q., Chen, T., Xu, R., He, Y., & Gui, L. (2016). Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. Database, 2016. doi: 10.1093/database/baw140
Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z., & Xu, H. (2017). Entity recognition from clinical texts via recurrent neural network. *BMC Medical Informatics and Decision Making*, *17*(S2). doi: 10.1186/s12911-017-0468-7
Gupta, Tuchar. (2017). Deep Learning: Feedforward Neural Network. Towards Data Science. https://towardsdatascience.com/deep-learning-feedforward-neural-network-26a6705dbdc7
Ma, J. (2016, April 4). All of Recurrent Neural Networks. Retrieved from https://medium.com/@jianqiangma/all-about-recurrent-neural-networks-9e5ae2936f6e.
Srivastava, Pranjal. (2017). Essentials of Deep Learning : Introduction to Long Short Term Memory. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/
Bengio, Yoshua, et al. "A Neural Probabilistic Language Model." Journal of Machine Learning Research, vol. 3, Feb. 2003, pp. 1137–1155.