Predicting Psychiatric Readmission with BERT

Emmy Phung MTP363@NYU.EDU
NYU Center for Data Science

Kevin Yie KY848@NYU.EDU
NYU Langone School of Medicine

Abstract

Rapid psychiatric readmissions present a burden for both the patients as well as the hospitals. Due to the difficulties in assessing mental health status and complications inherent to mental illnesses, premature discharges as well as rapid decay of mental status after discharge is an issue many psychiatric facilities faces. We aim to create a BERT-based model that can assist clinicians in identifying patients as risk of psychiatric readmission upon discharge as well as extract abstract relationships from discharge summaries.
Keywords: BERT, psychiatric readmission prediction

1 Introduction

The motivation for this project comes from established research that have shown initial success on predicting readmission possibility using clinical notes. One example is predicting heart failure readmission from discharge notes using conventional embedding methods, such as word2vec, and basic models, such as Random Forest and 3-layer Convolutional Neural Network (CNN) [1]. The promising results of very simple model architecture motivates us to attempt predicting psychiatric readmission from discharge notes using more complex embedding approaches and deeper neural network.

Psychiatric readmission research have primarily focused on applying traditional linguistic methods such as LDA [2]. These methods have seen some success however, due to the naïve nature of LDA, it may be inefficient in capturing information within discharge summaries. More novel NLP methods such as BERT or ELMO have yet to see application within the psychiatric readmission domain. However, application of BERT in general readmission has shown great promise [3]. The BERT model is able to more effectively capture abstract relationships through the use of the attention modeling and its pre-training allows it to have some notion of the English language. This knowledge is enhanced when fine-tuned on discharge summaries in particular with models such as ClinicBERT and Bio Discharge BERT performing very well on medical NLP tasks such as NER and de-identification [4].

2 Hypothesis

We believe that the stages of a mental illness cannot be fully described through numerical measurements and are best captured in words. Based on this notion, we hypothesize that from clinical notes, which contains patients' medical history, presenting symptoms and stages when discharged, we can anticipate the possibility of readmission as well as extract other meaningful insights. For this purpose, we want to explore the most recent approaches in natural language processing, LSTM and BERT, which are known to be capable of learning from unstructured text data.

3 Data

3.1 Data Source

The MIMIC III dataset [1] contains information from patients staying at the ICU at Beth Israel Deaconess Medical Center. Of especial import for our use are the tables DIAGNOSIS_ICD and NOTEEVENTS which detail the ICD9 code diagnosis as well as all notes associated with the patient respectively on each visit. We filtered the diagnosis ICD9 codes for all codes within range 290-319 (the category for mental disorders) as well as E950-E959 (the category for self-harm). We also extracted information from discharge summaries within NOTEEVENTS by using regex patterns corresponding to history of past illness, discharge instructions, etc. The filtered dataset from both tables were joined at the level of granularity of per visit. Standard cleaning was performed on the dataset: eliminating escape characters, removing filler tags, converting to lowercase, etc.

3.2 Data

Our final dataset contains the discharge summaries from 7,958 unique admission of 7,050 patients. We label each admission as *readmission* if the admission is followed by another admission and as *no readmission* otherwise. *Readmission* is our positive class which accounts for around 22% of all observations. Our negative class, *no readmission*, is the majority class which accounts for nearly 78%. For each admission, data is extracted from the corresponding discharge summary using regex patterns.

	Total	Train	Validation	Test
# Patients	7,050	4,630	1,716	1,721
# Admissions	7,958	5,530	1,844	1,844
• # Readmission (positive class)	7,159	4,295	1,432	1,432
• # No readmission (negative class)	2,059	1,235	412	412

Table 1. Data Summary

4 Materials and Methods

4.1 Deep Learning Models

Baseline Models

We started with a baseline model with the following architecture: a randomly initialized embedding layer that is fed into an LSTM layer before a final fully connected layer is used for classification. LSTMs have seen many successful applications in text classification and its structure is a natural fit for dealing with text sequences [5, 6, 7]. Information is kept within cell states and updated in each subsequent cell with the potential for unimportant information to be forgotten via sigmoid functions.

A major breakthrough in modern NLP has been the introduction of vector representations of semantic information within text sequences. An example of a fundamental algorithm within this space is the popular Word2Vec. By training a neural network on a proxy task such as predicting a word given its surrounding words (CBOW) or predicting surrounding words given a hidden word (Skip-gram), the weights of the hidden layer of the network naturally encodes semantic information and can be extracted as vectors [8]. These vectors allow for mathematical operations to be conducted on words such as the well-known example of "king – man + woman = queen". However, a limitation of the Word2Vec algorithm is that it only captures information from a word's local

context. We implement pre-trained GloVe embeddings within our final baseline model to overcome this limitation. GloVe captures global context of a word by using co-occurrence probabilities of words and solving for a probability scalar to form an intermediate vector representation that can be extracted [9].

BERT

Our LSTM with GloVe embedding model has its own limitations, however. Although LSTM's sequentially update information, information entering each cell is only based on the previous cell's state and output. This could result in abstract relationships between distant words not being effectively captured, degradation of upstream information, and, although better than simple RNN's, LSTM's could still suffer from vanishing or exploding gradient problems with very long sequences [10, 11]. GloVe also has limitations in that every word has only one vector representation vector despite the potential of having different meanings. For example, the word "cell" in "prison cell" and "animal cell" have two different meanings but "cell" will only have one vector to represent all the multiple meanings.

The introduction of attention models and BERT overcome these limitations. The attention mechanism considers each token's contribution to each other and a positional embedding layer allows the model to differentiate between tokens in different contexts and locations. Integration of attention in models has been shown to have better performance across various tasks [12, 13, 14] and BERT in particular has consistently been amongst the best performing models for established NLP tasks within GLUE [15]. The BERT model has already been pre-trained on a large corpus of information from Wikipedia and BooksCorpus, providing it with a solid foundation in how the English Language is structured but fine-tuning the model for specific downstream tasks within different domains has shown improved performance within that space [16, 17]. ClinicBERT and Bio Discharge BERT [4] shows an increase in performance in medical NLP tasks such as NER and patient de-identification after being pre-trained on discharge summaries. We implement and train the base BERT model as both a classifier and a feature extractor as well as implementing Bio Discharge BERT.

4.2 Model Training

LSTM

We add a fully connected classification layer on top of our LSTM model and train the entire model with two different embedding methods. One is non-pretrained embedding and the other is pretrained GloVe embedding. The primary hyper parameter that we want to explore within LSTM's architecture is its hidden dimension. While the model's loss decreases when we shrink the output size of the last hidden layer, it does not give us any changes on the validation set.

BERT

We apply two approaches of transfer learning to train our BERT Base and BERT Discharge models, which are 1) using BERT as a feature extractor and train only the last classification layer and 2) finetuning BERT layer together with the classifier. [20] In the first approach, we freeze the parameters in the BERT layer and train the classifier until no significant loss reduction or increase in accuracy and AUC on validation set are obtained. The primary hyper parameters of interest are batch size (12 – 32) and learning rate ($2e^{-2}$ – $2e^{-5}$).

In the second approach, we finetune the full BERT Base/Discharge Classification model using a smaller range of batch size (6 – 12) due to limited computation resources. For learning rate, we also adopt a more narrow range ($2e^{-5}$ – $5e^{-5}$), which are standard to training BERT's attention layers. Since BERT tend to overfit the training set after 4 epochs, we only train the full model within 5 epochs at max. [21]

Loss Function

We use Cross Entropy as our primary loss function as the task in hand is a classification problem. When training BERT Base Classifier, we opt for the built-in Binary Cross Entropy with Logits Loss.

Optimizer

Our choice for the optimizer is Adam (for LSTM) and Adam with weight decay (AdamW) for BERT. Adaptive optimizers like these two have been preferred over other conventional optimizers like stochastic gradient descent with momentum (SGDm) or Root Mean Square Propagation (RMS Prop) to train deep learning models for its ability to scale the learning rate using squared gradients (like RMS Prop) and take advantage of momentum by using moving average of the gradient (like SGDm).[18] As a combination of the SGDm and RMS Prop, Adam and AdamW optimizers have been proven to generalize well and lead to more rapid convergence in recent research. [19]

Evaluation Metrics

For training purpose, we evaluate our models' capacity to learn based on cross entropy loss, accuracy, and area under the receiver operating characteristic curve (AUROC) across epochs. Final model performance on test set will be evaluated based on AUC, precision & recall, and F1

The rationale for selecting these metrics is that we face class imbalance problem. The model may obtain a fairly good accuracy (nearly 78%) and 0.0 precision if it predicts all observations to be the majority class. Therefore, AUC, precision, and F1 are more reliable metrics for model evaluation in our case.

5 Results

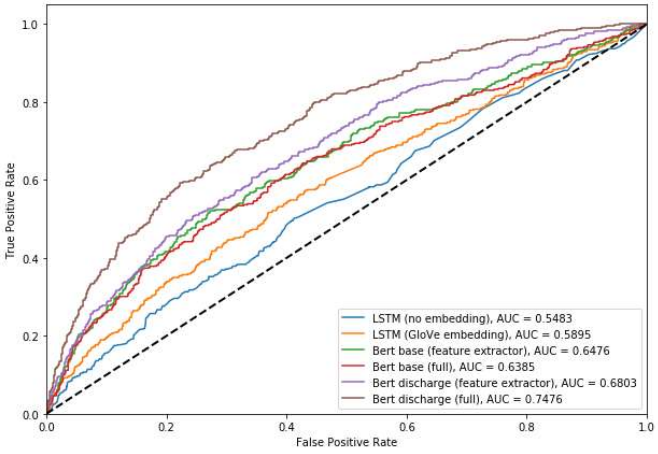


Figure 1. ROC plot on test set

Figure 1. shows the Receiver Operating Characteristics curves of our six models and their corresponding AUROC evaluated on test set. We can observe that contextual embedding improves the AUROC of our LSTM baseline model but LSTM still underperforms BERT in this task. Finetuned BERT Discharge model that used pretrained embedding on MIMIC's discharge notes outperforms our BERT Base model as expected. Lastly, between the two applications of transfer learning, our second approach, finetuning the full BERT Classification model gives a higher AUROC. Full performance results of the six models are summarized in Table 2.

The precision and F1 results demonstrate that our LSTM models predicts all observations to be the majority class. Compared to LSTM, BERT Base models have some capacity to learn but perform poorly on the test set. A significant improvement on the BERT Base Classification model is the adoption of pretrained embeddings on MIMIC's discharge notes, known as BERT Discharge. Interestingly, using BERT Discharge for feature extraction gives us the best precision (0.59) with some sacrifice on accuracy and recall. Evaluating across all five metrics, we conclude that BERT Discharge Full model gives us the best prediction results on the test set, achieving F1 of 0.66 and AUROC of nearly 0.75.

	Accuracy	AUC	Precision	Recall	F1
LSTM (No embedding)	0.7776	0.5483	0.0	1.0	0.0
LSTM (GloVe embedding)	0.7776	0.5895	0.0	1.0	0.0
BERT (Base – feature extractor)	0.7798	0.6476	0.1335	0.963	0.2345
BERT (Base – full)	0.7777	0.6385	0.3835	0.9225	0.5418
BERT (Discharge Summary – feature extractor)	0.6464	0.6803	0.585	0.6641	0.6220
BERT (Discharge Summary – full)	**0.7402**	**0.7476**	**0.5704**	**0.7891**	**0.6622**

Table 2. Psychiatric readmission prediction performance

While predictive power is one of our primary goals, we also want to be able to derive insight into how and why our model learns information. This is one of the biggest advantages that attention-based models have provided. Unlike the complex and abstract vectors that the LSTM uses to encode cell state, attention weights serve as a readily interpretable depiction of the connections that our model forms between various tokens. Through the use of the BertViz tool [23], the attentions between tokens within BERT can be visualized in an easily interpretable format. We've randomly chosen 2 sentence couplings without any post hoc selection influences for demonstration purposes (Appendix Figures A & B). These examples demonstrate BERT's ability to learn abstract concepts such as the relationship between symptoms/diseases and treatments as well as between drugs and their side effects/treatment purposes.

In the first example, we see that the patient presents with shortness of breath and hypoxia and was given oxygen as a treatment. We also see that nebulized medicine was administered for his pneumonia (Appendix A1). BERT consists of 12 attention layers with 12 attention heads each. We expect that the deeper layers of the model should be responsible for learning deeper and more abstract relationships while the early layers will form more basic connections. Visualization of the attention weights after the two sentences were passed into our model shows that the first few attention layers are not learning strong connections between any of the tokens as expected (Appendix A2). The lasts few layers of the network however reveals that strong connections are drawn to the separation token tag as well as the word oxygen and nebulizer (Appendix A3). The strong connections to the separation tokens or the classification tokens is an expected and often occurring pattern in attention models.

A closer look at some of the heads that compose the latter layers shows that some of the individual heads are responsible for forming very specific connections that relay relationship. The words breath and hypoxia are both lending a lot of weight towards the token oxygen, a pattern only present for those words. This shows that this head was able to extract that abstract relationship between administering oxygen to those patients that lack it. It was also able to draw connections between pneumonia and the administering of nebulized medicine, demonstrated by the strong connections in that head (Appendix A4).

A second example shows the ability for BERT to detect even more abstract relationships like those between drugs and their side effects/treatment purposes: a feat that even humans would be unable to do without previous knowledge of the drug. Sentence A shows that the patient suffers from benzodiazepine withdrawal while the second sentence shows the results of that abuse and withdrawal: agitation, anxiety, and psychosis (Appendix B1). We see from the visualization that strong connections between the drug and anxiety form, likely as a result of the drug being often used to sedate agitated patients (Appendix B2). We also see that the drug is strongly attending to psychosis which the drug is often used to sedate as well as abuse of benzodiazepines causing psychosis.

6 Discussion

We hypothesized that our model would be able to predict for psychiatric readmission as well as identify and learn abstract relationships between medical concepts. Our final model shows that it was able to improve upon the accuracy of both baseline LSTM models thanks to the integration of the attention mechanisms as well as the deeper network present in BERT. However, a limitation of this study is that our final precision and recall scores are still not at a level that can be deployed. The lack of strong performance within the model could be indicative of the difficulty in the task of predicting solely psychiatric readmission. Another difficulty lies in using the MIMIC III dataset as the patients were from a specific hospital and were admitted to the ICU first. This causes heavy bias within the data as most of the patients who were included were likely severe cases and likely had other co-occurring issues as well to warrant admission to the ICU. Lastly, the present study included only discharge summaries however, a more comprehensive view into a patient's history will likely lead to better performance as mental illnesses are a complex subject.

However, despite the underperforming models, a key finding of this study is that BERT is still capable of forming and constructing abstract relationships regardless of task performance. In a similar fashion as how vectorization algorithms train their models on a proxy task for the ultimate goal of extracting embeddings, so too can we train BERT on proxy medical tasks to generate relationships as defined by attention. A potential application of this approach could be in recognizing novel drug side-effects or interactions by extracting tokens that drug's give a lot of attention to. We could also potentially see a different usage for readmission prediction using BERT by analyzing relationships between words that are strongly linked to "readmission" to identify risk factors. Limitations of these applications, however, is determining cutoffs signifying "significant" relationships as well as automating a method to manually comb through the massive amounts of relationships that are present within the attention network, most of which is meaningless.

Finally, despite our models achieving suboptimal prediction power, this does not suggest that deep learning/machine learning does not have a place in psychiatric readmission prediction. Models such as XLNet has shown promise over BERT for certain NLP tasks and is an avenue for further exploration. Another possible method of improving power could be in text summarization algorithms. BERT is only capable of taking 512 tokens however, discharge summaries often have far more tokens. By using text summarization algorithms, we could filter out unnecessary information.

References

1. Liu, X., Chen, Y., Bae, J., Li, H., Johnston, J., & Sanger, T. (2019). Predicting heart failure readmission from clinical notes using deep learning. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. doi:10.1109/bibm47256.2019.8983095

2. Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., Mccoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry*, *6*(10). doi: 10.1038/tp.2015.182

3. Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *ClinicalNLP Workshop at NAACL 2019*.

4. Alsentezer, E., et. al., (n.d.). Publicly Available Clinical BERT Embeddings. Retrieved from https://arxiv.org/pdf/1904.03323.pdf

5. Overview of the MIMIC-III data. (n.d.). Retrieved from https://mimic.physionet.org/gettingstarted/overview/

6. Nowak, J., Taspinar, A., & Scherer, R. (2017). LSTM Recurrent Neural Networks for Short Text and Sentiment Classification. *Artificial Intelligence and Soft Computing Lecture Notes in Computer Science*, 553–562. doi: 10.1007/978-3-319-59060-8_50

7. Rao, G., Huang, W., Feng, Z., & Cong, Q. (2018). LSTM with sentence representations for document-level sentiment classification. *Neurocomputing*, *308*, 49–57. doi: 10.1016/j.neucom.2018.04.045

8. Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. Retrieved from https://arxiv.org/pdf/1801.06146.pdf

9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Retrieved from https://arxiv.org/pdf/1310.4546.pdf

10. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. doi: 10.3115/v1/d14-1162

11. Culurciello, E. (2019, January 10). The fall of RNN / LSTM. Retrieved from https://towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0

12. Lamba, H. (2019, May 9). Intuitive Understanding of Attention Mechanism in Deep Learning. Retrieved from https://towardsdatascience.com/intuitive-understanding-of-attention-mechanism-in-deep-learning-6c9482aecf4f

13. Gao, L., Guo, Z., Zhang, H., Xu, X., & Shen, H. T. (2017). Video Captioning With Attention-Based LSTM and Semantic Consistency. *IEEE Transactions on Multimedia*, *19*(9), 2045–2055. doi: 10.1109/tmm.2017.2729019

14. Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based LSTM for Aspect-level Sentiment Classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. doi: 10.18653/v1/d16-1058

15. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. doi: 10.18653/v1/p16-2034

16. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.

17. Han, X., & Eisentein, J. (2019). Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling. Retrieved from https://arxiv.org/pdf/1904.02817.pdf

18. Beltagy, I., Lo, K., & Cohan, A. (2019). SCI BERT: A Pretrained Language Model for Scientific Text. Retrieved from https://arxiv.org/pdf/1903.10676.pdf

19. Bushaev, V. (2018, October 22). Adam — latest trends in deep learning optimization. [Web log post]. Retrieved from https://towardsdatascience.com/adam-latest-trends-in-deep-learning-optimization-6be9a291375c

20. Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR Conference*.

21. Ruder, S. (2019, August 18). The State of Transfer Learning in NLP [Web log post]. Retrieved from https://ruder.io/state-of-transfer-learning-in-nlp/

22. McCormick, C., & Ryan, N. (2019, July 22). BERT Fine-Tuning Tutorial with PyTorch [Web log post]. Retrieved from https://mccormickml.com/2019/07/22/BERT-fine-tuning/

23. Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. Retrieved from https://arxiv.org/pdf/1906.05714.pdf

Detailed implementation of the project can be found at:
https://github.com/Emmyphung/psychiatric-readmission-prediction

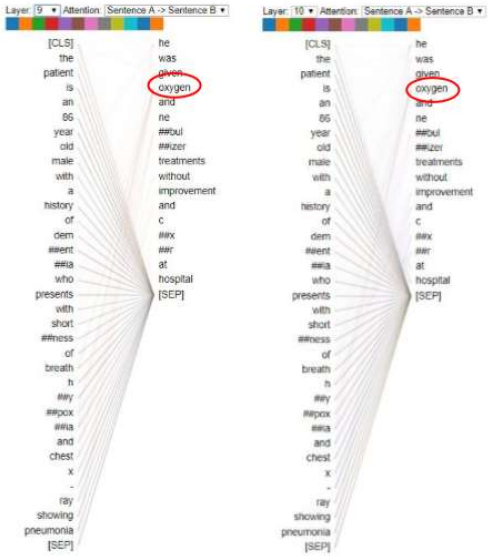
Appendix A:

1)

Sentence A: The patient is an 86 year old male with a history of dementia who presents with shortness of breath, hypoxia, and chest x-ray showing pneumonia.
Sentence B: He was given oxygen and nebulizer treatments without improvement and XCR at Hospital.

2)

3)



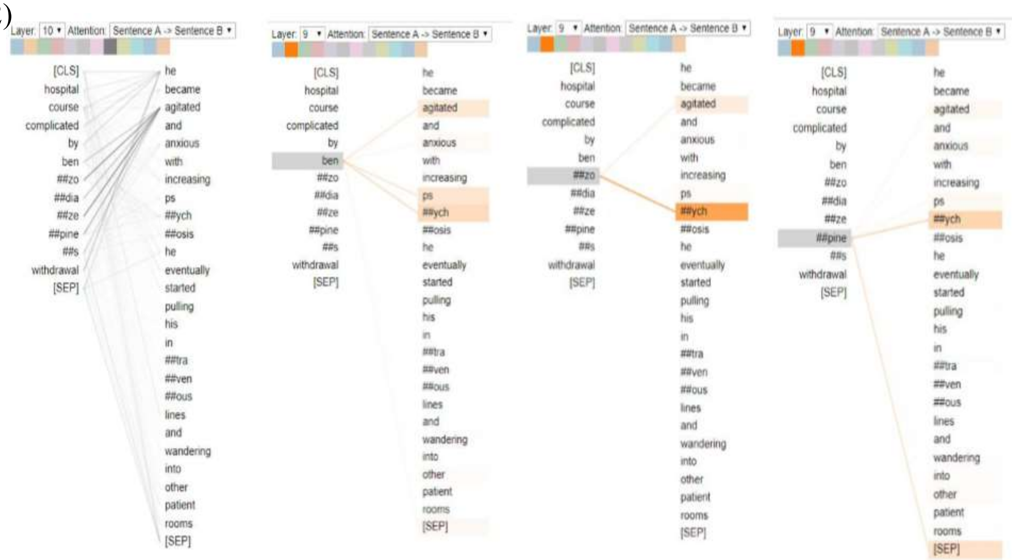
4)

Appendix Figure A.
1) Sample sentence showing relation between symptoms/disease and treatment
2) First few layers of BERT
3) Last few layers of BERT
4) Specific attention heads within layer 10 of BERT

Appendix B:

1)
Sentence A: Hospital course complicated by **benzodiazepines withdrawal**.
Sentence B: He became **agitated** and **anxious** with increasing **psychosis**. He eventually started pulling his intravenous lines and wandering into other patient rooms.

2)



Appendix Figure B:
1) Example sentences showing drug and its side effects and treatment purposes.
2) Attentions focused at agitation and psychosis from benzodiazepines