Textual Data Augmentation for Patient Outcomes Prediction

Qiuhao Lu
University of Oregon
Eugene, OR, USA
luqh@cs.uoregon.edu

Dejing Dou
University of Oregon
Eugene, OR, USA
Baidu Research
Beijing, China
dou@cs.uoregon.edu

Thien Huu Nguyen
University of Oregon
Eugene, OR, USA
thien@cs.uoregon.edu

Abstract—Deep learning models have demonstrated superior performance in various healthcare applications. However, the major limitation of these deep models is usually the lack of high-quality training data due to the private and sensitive nature of this field. In this study, we propose a novel textual data augmentation method to generate artificial clinical notes in patients' Electronic Health Records (EHRs) that can be used as additional training data for patient outcomes prediction. Essentially, we fine-tune the generative language model GPT-2 to synthesize labeled text with the original training data. More specifically, We propose a teacher-student framework where we first pre-train a teacher model on the original data, and then train a student model on the GPT-augmented data under the guidance of the teacher. We evaluate our method on the most common patient outcome, i.e., the 30-day readmission rate. The experimental results show that deep models can improve their predictive performance with the augmented data, indicating the effectiveness of the proposed architecture.

Index Terms—data augmentation, GPT-2, readmission prediction, EHR

I. INTRODUCTION

Patient outcomes, including patients' readmission risk, mortality rate, and length of stay (LOS), have been examined as important measurements for evaluating the quality of hospital care [1]. As the most commonly reported health outcome in the United States, readmissions are estimated to cost Medicare $15 billion annually, of which $12 billion is potentially preventable, according to the Medicare Payment Advisory Committee [2]. This highlights the importance of identifying patients at high risk of readmission.

Over the past few years, there has been a surge of interest in making predictions on patient outcomes using deep learning techniques, such as readmission prediction [3], mortality prediction [4], length of stay prediction [5], etc. Most of these studies heavily rely on feature engineering, where they select statistically significant features from patients' Electronic Health Records (EHRs), and feed them into deep models like a LSTM-CNN network [3].

A common theme among these studies is that they all rely on numerical and time-series features of patients, while neglecting the clinical notes of EHRs which prove to be informative in such predictive tasks. This motivates recent studies to cast this task as text classification, where the contextual content of EHRs are leveraged to make predictions. For example, Lu *et al.* propose a graph-based method that converts clinical notes to multi-view graphs and use them to predict ICU patients' 30-day unplanned readmission risk, surpassing state-of-the-art numerical-based methods [6].

However, in real-world downstream applications, deep learning models often suffer from data limitation as they require large amounts of data for effective training. The situation is even worse in the biomedical domain due to the private and sensitive nature of this field. Despite data shortage, data imbalance is also an issue for patient outcomes prediction, e.g., only few patients are readmitted post-discharge. These data issues make patient outcomes prediction more challenging than general predictive tasks.

A natural solution to these problems is data augmentation, where new data is synthesized based on existing training data. This strategy has been actively applied in the field of computer vision, where researchers alter the training images to create a larger dataset by introducing random transformations such as translation, mirroring, rotation, and more [7]. However, these augmentation strategies that are successful in computer vision cannot be easily applied to textual data due to the inherent complexity of natural language [8], where the grammatical or semantic consistency of text could hardly be preserved after transformation [9]. As to the specific task of readmission prediction, such issues, e.g., data imbalance, are either ignored [10] or processed with sampling techniques [11], such as SMOTE [12] or ROSE [13] that do not cope with textual data.

Recently, natural language generation (NLG) techniques have been leveraged as a new means for textual data augmentation. With the development of large pre-trained generative language models like GPT-2 [14], researchers are able to generate high-quality and semantic-consistent textual data while preserving the annotated labels. This augmentation strategy has been applied in various NLP downstream tasks, such as event detection [15], relation extraction [16], commonsense reasoning [17], etc. However, in the biomedical field, leveraging GPT-2 to facilitate clinically-relevant predictive models is under-explored.

One main challenge of using GPT-2 for textual data augmentation is noise control. Existing studies typically address

this issue in a isolated way, where they introduce heuristic filtering mechanisms to eliminate low-quality samples [9] and feed the rest to the downstream model. However, such filtering strategies are prone to coverage errors and thus inevitably make incorrect judgements on the generated samples [15], which would cause false inclusion of good samples or false exclusion of bad samples. Moreover, the combined data samples are treated equally by the to-be-trained downstream model, and this would negatively impact the model as a consequence.

To overcome this issue, we propose a conceptually different strategy where all the generated samples are involved during training. We preserve all the generated samples in the first place, and then introduce a teacher-student framework to regularize the representation learning of the generated samples with knowledge transferred from the original data. More specifically, we pre-train a teacher model on the original data and then train a student model on the combined data adaptively under the guidance of the teacher. The goal is to transfer the knowledge learned in the teacher model into the student model by enforcing a knowledge consistency between them, and that eventually the student model can be improved. We evaluate the framework with the state-of-the-art textual-based readmission prediction model [6], the results of which indicates the effectiveness of the method.

The contributions of this work can be summarized as follows:

- We propose a novel architecture that leverages GPT-2 for **Med**ical text **Aug**mentation (MedAug) in the task of patient outcomes prediction. Essentially, we introduce a teacher-student framework that aims to control the noise of the generated text by enforcing a knowledge consistency across the original and artificial texts.
- Taking the readmission prediction task as a case study, we specifically investigate the performance of MedAug with the state-of-the-art readmission prediction model as well as a baseline model. Extensive experiments demonstrate that both models can improve their performance with the augmented data, indicating the effectiveness of the proposed architecture.

II. METHODOLOGY

A. Notations

In this study, we focus on textual-based readmission prediction models where the prediction task is cast as a supervised binary text classification problem. We refer to the original training dataset as $D_{train} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where x_i is a clinical note and $y_i \in \{0, 1\}$ indicates whether the patient is readmitted or not. Note that D_{train} is imbalanced where negative samples are 3x more than the positive ones, as only few patients are readmitted post-discharge. We similarly denote the test set by D_{test} and the validation set by D_{valid}. We also denote the synthesized training set by $D_{synthetic}$, which is generated by the fine-tuned GPT-2 model G_{tuned}. We also combine the original and generated training data together to create a large training dataset $D_{combined} = D_{train} \cup D_{synthetic}$. Finally, we refer to the prediction method as $\mathcal{M}$.

Algorithm 1: MedAug

Input: D_{train}, $\mathcal{G}$, $\mathcal{M}$
Output: $\mathcal{M}_{student}$

1 Fine-tune $\mathcal{G}$ on D_{train} to obtain $\mathcal{G}_{tuned}$
2 Use $\mathcal{G}_{tuned}$ to generate $D_{synthetic}$ and combine it with D_{train} to obtain $D_{combined}$
3 Pre-train a teacher model $\mathcal{M}_{teacher}$ on D_{train}
4 Train the student model $\mathcal{M}_{student}$ on $D_{combined}$ under the guidance of $\mathcal{M}_{teacher}$
5 **Return** $\mathcal{M}_{student}$

B. Data Generation

We fine-tune the GPT-2 model $\mathcal{G}$ on the original training data D_{train} so that it can synthesize reasonable textual data that can be used for the training of $\mathcal{M}$. To preserve the class information, we prepend the class label y_i to each note in the training data, i.e., $y_i \text{SEP} x_i \text{EOS}$, where SEP and EOS are the separation and ending token, respectively. We then fine-tune GPT-2 on the processed training data with the objective of predicting the next token, the same way it was pre-trained [14]. The fine-tuned model is regarded as $\mathcal{G}_{tuned}$.

For generating new data, we use the class label along with a short context as the prompt to G_{tuned}, i.e., prompt = $y_1 \text{SEP} w_1 w_2$ where the first two tokens are included as context, as suggested in [9]. Since in our case the negative samples are 3x more than the positive ones, we only focus on generating positive samples to fulfill the gap, i.e., only the positive label y_1 is used for generation. We denote the generated training data by $D_{synthetic}$.

C. Data Integration

As mentioned in the introduction, noise control is one of the main challenges for textual data augmentation. In this work, we propose a teacher-student framework for data integration so that all the generated samples are included for training. We first pre-train a teacher prediction model $\mathcal{M}_{teacher}$ on D_{train} to capture the inherent knowledge of the original clean training data. Then we train the student model $\mathcal{M}_{student}$ on the combined data $D_{combined}$ in a way that the teacher's knowledge can be used to guide the student learning. To achieve this, we aim to enforce a knowledge consistency between the student and the teacher, by incorporating a KL divergence penalty to push the representations learned in the student model close to that in the teacher. Essentially, we seek to jointly minimize the KL divergence between the predicted label probability distribution of the student and the teacher, along with the original training objective of the student, i.e., $\mathcal{L} = \mathcal{L}_{student} + \tau \mathcal{L}_{KL}$. It's also worth mentioning that in this study we use the KL divergence to control noise in the labeled data generated by GPT-2, which is different from knowledge distillation on unlabeled data [18]. The architecture is defined in Algorithm 1.

III. EXPERIMENTS

In this section, we evaluate the proposed framework on the task of ICU patients readmission prediction where we aim to show the effectiveness of MedAug. Essentially, we take as input the clinical note of patients' EHRs, and predict whether or not the patient will be readmitted within 30 days after discharge or transfer.

A. Dataset

The experiment is conducted based on the MIMIC-III Critical Care (Medical Information Mart for Intensive Care III) Database, which is a large, freely-available database composed of de-identified EHR data [19]. Following prior work [20], we extract the `Discharge Summaries` from EHRs as the data. For a fair comparison, we use the same data split with the baseline [6] where $48,393$ generated documents are split into training (80%), validation (10%), and testing (10%). Specifically, the original training set D_{train} consists of 7555 positive samples and 30247 negative samples which are denoted by $D_{train,1}$ and $D_{train,0}$, respectively.

B. Evaluation Metrics

We follow the prior work [6] and use the area under the receiver operating characteristics curve (AUROC), the area under the precision recall curve (AUPRC), and the recall at precision of 80% (RP80) for evaluation.

C. Prediction Models

We consider the following two prediction models for evaluation in this experiment. We evaluate with two prediction models to investigate how MedAug performs when equipped with a base model and an advanced model.

- ClinicalBERT. ClinicalBERT is a domain-specific BERT variant initialized from BioBERT v1.0 [21] and pre-trained on MIMIC notes [22]. In this study, we add a linear classification head on top of it and use it as a baseline.
- MedText. MedText is a textual-based readmission prediction model and reports state-of-the-art performance on this task [6].

D. Augmentation Baselines

We consider two augmentation baselines for comparison.

- base. The base strategy is a baseline that all generated samples are included while no noise control is applied.
- LAMBADA. LAMBADA is an augmentation method specified for text classification [9]. Basically, they pre-train a classifier on the clean data and use it to select confident samples.

E. Results

Table I shows the test performance of the two readmission prediction models, along with three augmentation strategies. We observe that without controlling the noise, i.e., base, both models demonstrate inferior performance, indicating the non-negligible level of noise in the generated samples. On the

TABLE I: Test performance on 30-day unplanned ICU patient readmission prediction.

Method	AUROC	AUPRC	RP80
ClinicalBERT	0.782	0.549	0.201
ClinicalBERT-base	0.779	0.550	0.221
ClinicalBERT-LAMBADA	0.782	0.543	0.196
ClinicalBERT-MedAug	**0.791**	**0.565**	**0.234**
MedText	**0.823**	0.632	0.319
MedText-base	0.803	0.599	0.290
MedText-LAMBADA	0.806	0.604	0.266
MedText-MedAug	0.822	**0.633**	**0.328**

TABLE II: Influence of $|D_{synthetic}|$ by MedAug.

$|D_{synthetic}|$	Method	AUROC	AUPRC	RP80
3k	ClinicalBERT	0.777	0.550	0.220
9k	ClinicalBERT	0.784	0.567	0.246
12k	ClinicalBERT	0.784	0.569	0.245
24k	ClinicalBERT	0.783	0.566	0.251
3k	MedText	0.812	0.621	0.329
9k	MedText	0.811	0.623	0.337
12k	MedText	0.806	0.611	0.311
24k	MedText	0.809	0.618	0.331

other hand, with MedAug, both models demonstrate better performance, and the improvement is significant comparing with the other two baselines, indicating the effectiveness of this framework.

IV. ANALYSIS

In this section, we investigate three potential issues that might have influenced the performance of MedAug, i.e., the number of synthesized samples $|D_{synthetic}|$, the fine-tuning and generation strategy for GPT-2, and the version of GPT-2.

A. Number of Synthesized Samples

Table II shows the validation performance of different $|D_{synthetic}|$, demonstrating the influence of the size of the synthetic training set. With the increasing of synthesized samples, the general performance appears to have reached a peak and then begin to drop slightly. We conjecture that there is a trade-off between the size and the performance, and it is determined by the augmentation strategy.

B. GPT-2 Fine-tuning Strategy

It is common that patient outcomes demonstrate an imbalanced distribution, e.g., only few patients are readmitted after discharge. In our case, negative samples are 3x more than the positive ones, i.e., $D_{train,0} = 4 \times D_{train,1}$. Therefore, when fine-tuning GPT-2 using the original training data, we explicitly make it balanced to prevent the negative samples from misleading GPT-2, by performing random under-sampling over D_{train}. As to the prompt to GPT-2 in generating new samples, we compare two options, i.e., w/ and w/o context, where context refers to the first two tokens of the text.

We investigate the two issues and show the comparison results on the validation set in Table III. Note that to avoid

TABLE III: Influence of GPT-2 fine-tuning/generation strategies.

Prompt	Balanced	Method	AUROC	AUPRC	RP80
w/o ctx	N	ClinicalBERT	0.771	0.535	0.205
w/o ctx	Y	ClinicalBERT	0.773	0.536	0.216
w/ ctx	N	ClinicalBERT	0.767	0.531	0.198
w/ ctx	Y	ClinicalBERT	0.775	0.551	0.226
w/o ctx	N	MedText	0.791	0.589	0.296
w/o ctx	Y	MedText	0.791	0.595	0.313
w/ ctx	N	MedText	0.791	0.593	0.296
w/ ctx	Y	MedText	0.795	0.602	0.318

TABLE IV: Influence of the version of GPT-2.

GPT-2 version	Method	AUROC	AUPRC	RP80
small	ClinicalBERT	0.784	0.567	0.246
medium	ClinicalBERT	0.783	0.568	0.252
small	MedText	0.811	0.623	0.337
medium	MedText	0.811	0.623	0.339

the impact from augmentation strategies, we use the base method, i.e., simply include all the samples, in this experiment. Generally, a balanced training set and a prompt with context are the best options for fine-tuning and generation with GPT-2 in this task.

C. GPT-2 Version

Finally, we investigate the version of GPT-2 and its influence over the quality of synthesized samples. We test with GPT-2-small and GPT-2-medium and show the results in Table IV. Generally, we observe that GPT-2-medium has a minor advantage over GPT-2-small. However, considering the training cost and efficiency, we choose to use GPT-2-small for all the experiments in this study.

V. RELATED WORK

Readmission prediction is challenging task and has attracted a lot of attention over the years. Lin *et al.* select numerical chart event features over a 48-hour time window and feed them to a deep LSTM-CNN network [3] and achieve much better performance than traditional methods. Zhang *et al.* propose CC-LSTM that encodes external knowledge into text representations and outperforms Lin's work [20]. Afterwards Lu *et al.* propose to convert clinical notes to multi-view graphs and process them with graph convolution networks [6]. These studies demonstrate the value of textual content in EHRs and motivate us to apply textual data augmentation to this task.

Recently, using GPT-2 for augmenting textual training data has been studied for a variety of tasks in the NLP field, such as such as event detection [15], relation extraction [16], commonsense reasoning [17], spoken language understanding [23], extreme multi-label classification [24], etc. However, none of these works has leveraged GPT-2 for patient outcomes prediction. This highlights the importance of this study and motivates us to explore more of this direction.

VI. CONCLUSION

In this paper, we propose MedAug, a framework that leverages GPT-2 to synthesize artificial training data for patient outcomes prediction. We evaluate the method on task of ICU patients readmission prediction, the results of which demonstrate that either a baseline or an advanced prediction model can benefit from the synthesized training data, under the framework of MedAug. Essentially, to control the noise in the synthesized data, we propose a teacher-student architecture that enforces a knowledge consistency across the original and artificial texts. We introduce a mechanism for knowledge consistency enforcement to mitigate noises from generated data based on KL divergence.

On the other hand, as a preliminary exploration of this direction, we do observe that the improvement for the advanced model is less significant than the baseline model, which motivates us to investigate further in the future work.

REFERENCES

[1] B. A. Davison, M. Metra, S. Senger, C. Edwards, O. Milo, D. M. Bloomfield, J. G. Cleland, H. C. Dittrich, M. M. Givertz, C. M. O'Connor *et al.*, "Patient journey after admission for acute heart failure: length of stay, 30-day readmission and 90-day mortality," *European journal of heart failure*, vol. 18, no. 8, pp. 1041–1050, 2016.

[2] G. Hackbarth, "Reforming america's health care delivery system," *Statement before the Senate Finance Committee Roundtable on Reforming America's Health Care Delivery System*, p. 5, 2009.

[3] Y.-W. Lin, Y. Zhou, F. Faghri, M. J. Shaw, and R. H. Campbell, "Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory," *PloS one*, vol. 14, no. 7, p. e0218942, 2019.

[4] H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan, "Multi-task learning and benchmarking with clinical time series data," *CoRR*, vol. abs/1703.07771, 2017.

[5] X. Ma, Y. Si, Z. Wang, and Y. Wang, "Length of stay prediction for icu patients using individualized single classification algorithm," *Computer methods and programs in biomedicine*, vol. 186, p. 105224, 2020.

[6] Q. Lu, T. H. Nguyen, and D. Dou, "Predicting patient readmission risk from medical text via knowledge graph enhanced multiview graph convolution," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1990–1994.

[7] N. McLaughlin, J. M. Del Rincon, and P. Miller, "Data-augmentation for reducing dataset bias in person re-identification," in *2015 12th IEEE International conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2015, pp. 1–6.

[8] A. Amin-Nejad, J. Ive, and S. Velupillai, "Exploring transformer text generation for medical dataset augmentation," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4699–4708. [Online]. Available: https://aclanthology.org/2020.lrec-1.578

[9] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling, "Do not have enough data? deep learning to the rescue!" *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 7383–7390, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6233

[10] Q. Lu, N. de Silva, S. Kafle, J. Cao, D. Dou, T. H. Nguyen, P. Sen, B. Hailpern, B. Reinwald, and Y. Li, "Learning electronic health records through hyperbolic embedding of medical ontologies," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019, pp. 338–346.

[11] A. R. B. Junqueira, F. Mirza, and M. M. Baig, "A machine learning model for predicting icu readmissions and key risk factors: analysis from a longitudinal health records," *Health and Technology*, vol. 9, no. 3, pp. 297–309, 2019.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[13] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data mining and knowledge discovery*, vol. 28, no. 1, pp. 92–122, 2014.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[15] A. P. B. Veyseh, V. Lai, F. Dernoncourt, and T. H. Nguyen, "Unleash gpt-2 power for event detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 6271–6282.

[16] Y. Papanikolaou and A. Pierleoni, "Dare: Data augmented relation extraction with gpt-2," *ArXiv*, vol. abs/2004.13845, 2020.

[17] Y. Yang, C. Malaviya, J. Fernandez, S. Swayamdipta, R. Le Bras, J.-P. Wang, C. Bhagavatula, Y. Choi, and D. Downey, "G-daug: Generative data augmentation for commonsense reasoning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 1008–1025.

[18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[19] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[20] X. Zhang, D. Dou, and J. Wu, "Learning conceptual-contextual embeddings for medical text." in *AAAI*, 2020, pp. 9579–9586.

[21] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[22] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. [Online]. Available: https://www.aclweb.org/anthology/W19-1909

[23] B. Peng, C. Zhu, M. Zeng, and J. Gao, "Data augmentation for spoken language understanding via pretrained models," *arXiv e-prints*, pp. arXiv–2004, 2020.

[24] D. Zhang, T. Li, H. Zhang, and B. Yin, "On data augmentation for extreme multi-label classification," *arXiv preprint arXiv:2009.10778*, 2020.