Team Cadence at MEDIQA-Sum 2023: Using ChatGPT as a Data Augmentation Tool for Classifying Clinical Dialogue

Ashwyn Sharma[1], David I. Feldman[2,3]

[1]Cadence Solutions, USA
[2]Cadence Solutions, USA
[3]Massachusetts General Hospital, Harvard University , USA

Abstract

In this paper, we present Team Cadence's winning submission to Task A of the MEDIQA-Sum 2023 shared task, which focused on the classification of doctor-patient dialogues based on their associated topic or section header. The methodology we adopted was inspired by our previous work on the dialogue summarization task for MEDIQA-Chat 2023, where our data augmentation approach showed promising results. For this task, we leveraged **gpt-3.5-turbo**[1] to generate synthetic pairs of doctor-patient conversations and their corresponding section headers, subsequently augmenting the dataset. This augmented dataset was then utilized for fine-tuning the BART model (facebook/bart-large[2] checkpoint) for sequence classification. Results demonstrated that data augmentation improved classification accuracy for labels with scarce training data by 30%. Our submission ranked first on the Task A leaderboard, achieving an accuracy of 82%. Moreover, we analyzed the quality of synthetic data produced and the impact of augmentation on class imbalance.

Keywords

ChatGPT, Clinical dialogue classification, Data augmentation, MEDIQA-Sum

1. Introduction

In the context of NLP for clinical text, the ability to accurately classify sections of doctor-patient dialogues is crucial for understanding, summarizing, and managing healthcare conversations. This paper discusses Team Cadence's strategy for Task A of the MEDIQA-Sum 2023[1] shared task. Task A required participants to predict the topic or section header of given doctor-patient dialogues, a classification task that required a sophisticated understanding of medical language and contextual nuances.

Our approach was inspired by our previous work[2] on the dialogue summarization task of the MEDIQA-Chat 2023[3] shared task. In this prior endeavor, we developed a data augmentation approach that yielded promising results. Encouraged by these findings, we decided to apply a similar methodology to the Task A classification challenge.

[1]https://platform.openai.com/docs/models/gpt-3-5
[2]https://huggingface.co/facebook/bart-large

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
ashwyn@cadencerpm.com (A. Sharma); david.feldman@cadencerpm.com (D. I. Feldman)

We employed **gpt-3.5-turbo** to generate synthetic pairs of doctor-patient conversations and their corresponding section headers. We utilized this synthetic data to augment our training dataset, which was then used to train the BART[4] model (facebook/bart-large checkpoint) for sequence classification. This strategy demonstrated significant improvements, enhancing classification accuracy for labels with scarce training data by 30%.

Our submission achieved Rank-1 on the Task A leaderboard with an accuracy of 82% [1], demonstrating the effectiveness of our approach. Furthermore, we examined the quality of the synthetic data and the impact of this data augmentation technique on class imbalance. The detailed analysis and results of these experiments are discussed in the following sections.

2. Background and Related Work

Effective classification of doctor-patient dialogues has been recognized as a critical challenge in the clinical NLP field[5]. Accurate classification of these dialogues can provide insightful data for healthcare professionals and researchers, facilitating improved patient care and medical research[6, 7].

Doctor-patient dialogues form a rich data source about a patient's medical condition, symptoms, the prescribed treatment, and more. However, the unstructured nature of these dialogues poses significant challenges in efficiently extracting meaningful information. Over the years, NLP techniques have been increasingly used to tackle this issue, offering promising results in automatic information extraction, understanding, and management of these dialogues.

Several methods have been proposed for the classification of healthcare dialogues, typically involving a combination of traditional machine learning methods and, more recently, deep learning approaches. However, despite these efforts, accurately classifying healthcare dialogues remains a challenging problem due to the nuances and complexity of medical language and the inherent diversity in doctor-patient conversations[8].

Pre-trained transformer[9] models have achieved remarkable performance on several NLP tasks, including text classification. Particularly, the BART[4] model has been widely adopted due to its ability to effectively model the sequence of text data, making it suitable for tasks like the one posed by MEDIQA-Sum 2023[1].

Our team's earlier work for MEDIQA-Chat 2023 involved the use of data augmentation for dialogue summarization tasks. Inspired by the promising results of that approach, we aimed to adapt it for dialogue classification. Data augmentation has been increasingly recognized as an effective technique for improving the performance of machine learning models, especially in medical scenarios where the available data is limited or imbalanced[10].

This paper presents a data-augmentation-first approach to dialogue classification, integrating the use of synthetic data generation via **gpt-3.5-turbo** for data augmentation and the BART model for sequence classification. By doing so, we not only build upon the existing body of work but also introduce a methodology that we hope will inspire continued research in this crucial area.

[1] https://github.com/ashwyn/MEDIQA-Sum-2023-Cadence

Table 1
Section headers and their number of training examples.

Label	Count
FAM/SOCHX	351
GENHX	282
PASTMEDICALHX	118
CC	77
PASTSURGICAL	63
ALLERGY	60
ROS	60
MEDICATIONS	54
ASSESSMENT	34
EXAM	23
DIAGNOSIS	19
DISPOSITION	15
PLAN	11
EDCOURSE	8
IMMUNIZATIONS	8
IMAGING	6
GYNHX	5
PROCEDURES	3
OTHER_HISTORY	2
LABS	2

3. Methods

3.1. Dataset

The MEDIQA-Sum 2023[1] dataset forms the foundation for our work in this paper. Participants were provided with conversation snippets between a doctor and a patient and were tasked with identifying the associated section header or topic. These section headers represent one of twenty normalized common section labels, such as Assessment, Diagnosis, Exam, Medications, and Past Medical History, among others. The complete list of section headers is provided in Table 1.

The training set comprised 1,201 pairs of conversations and their corresponding section headers. The validation and test sets included 100 and 200 examples, respectively.

3.2. Data augmentation

Two major challenges encountered in this task were the relatively small size of the training dataset and class imbalance. To overcome this, we adopted a data augmentation approach inspired by the results of our work in MEDIQA-Chat 2023[3]. In that shared task, data augmentation proved to be a promising technique for summarizing clinical dialogues[2].

We utilized the **gpt-3.5-turbo** model to generate synthetic pairs of clinical conversations and the corresponding section headers. Leveraging a few-shot learning approach, we used the

Table 2
Hyperparameters used for fine-tuning the classifier.

Parameter	Value
learning_rate	2E-05
per_device_train_batch_size	8
per_device_eval_batch_size	8
weight_decay	0.01
num_train_epochs	30
fp16	TRUE
gradient_accumulation_steps	4
gradient_checkpointing	TRUE
max_source_length	1024
num_examples	10

following prompt as an input to the model:

"Given this training dataset for a classifier that predicts the section_header given a dialogue between a patient and a doctor, generate {num_examples} more examples for section_header {label}. Please follow the format of the given training dataset and output a csv. Training data: {samples}."

We generated 10 synthetic pairs for each section header ten times using the prompt above. After filtering out invalid examples, we obtained 2585 examples that were added to 1201 original training examples to create an augmented dataset of 3786 examples.

3.3. Classification

The BART[4] model, specifically the facebook/bart-large checkpoint, was fine-tuned using this augmented dataset. Our decision to use BART was based on the impressive results it yielded in our previous work at MEDIQA-Chat 2023.

The BART model was selected for its versatility in handling complex sequence-based tasks. We used the *BARTForSequenceClassification* wrapper provided by HuggingFace[11], which adds a classifier head to the BART model, making it suitable for the classification task. Fine-tuning was performed using the Trainer API offered by HuggingFace [11] and the hyperparameters used for fine-tuning the classifier are detailed in Table 2.

4. Experiments and Results

4.1. Classification metrics

In addition to accuracy, balanced accuracy[2] [12] and validation loss, we also evaluated the ability of the classifier to predict labels for classes where the given training data was limited. To be specific, we computed the mean accuracy for section headers where the number of training examples were less than 10. We call this metric *mean scarce accuracy* and such section

[2]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html

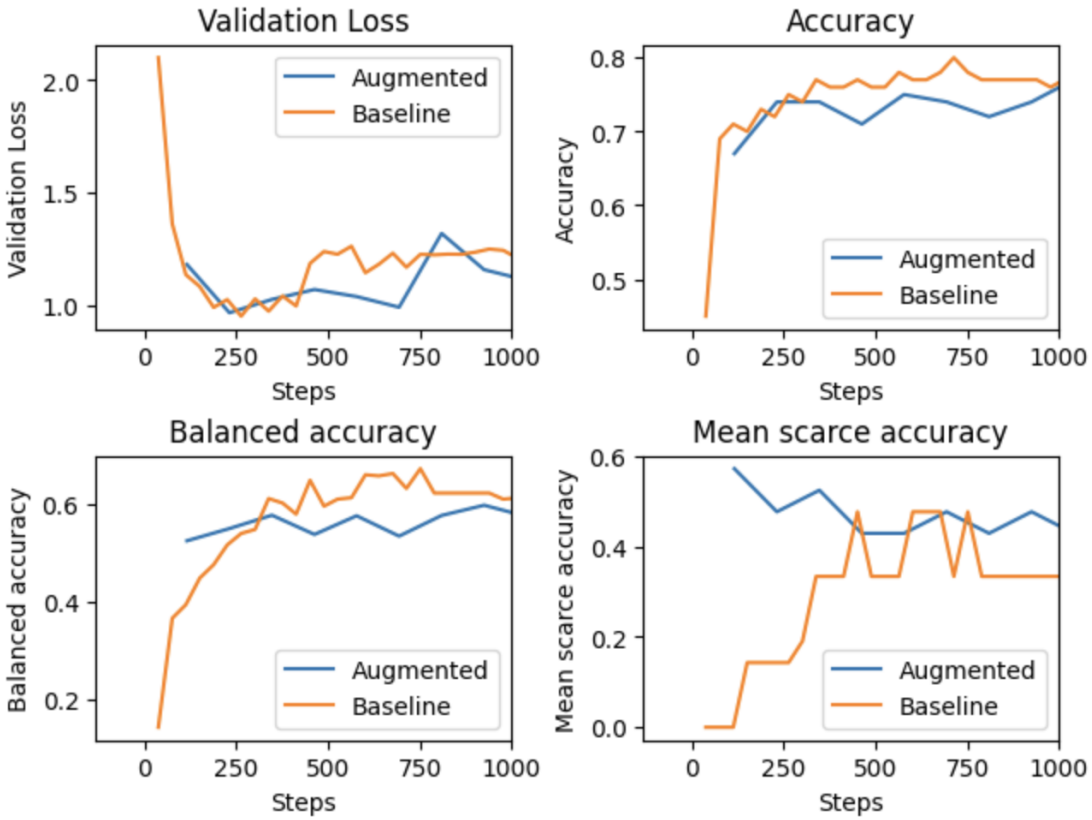

Figure 1: Comparison of classification metrics illustrating the impact of data augmentation.

Table 3
Scarce section headers with less than 10 training examples.

Label	Count
EDCOURSE	8
IMMUNIZATIONS	8
IMAGING	6
GYNHX	5
PROCEDURES	3
OTHER_HISTORY	2
LABS	2

headers are listed in Table 3. The goal here is to evaluate if data augmentation can help mitigate challenges associated with class imbalance in the training data.

Figure 1 illustrates the impact of data augmentation by comparing these metrics for the models fine-tuned on the augmented dataset and the original training data. It can be seen that the augmented version exhibits a lower validation loss for the most part and consistently outperforms the baseline model in *mean scarce accuracy*. The baseline model struggles with

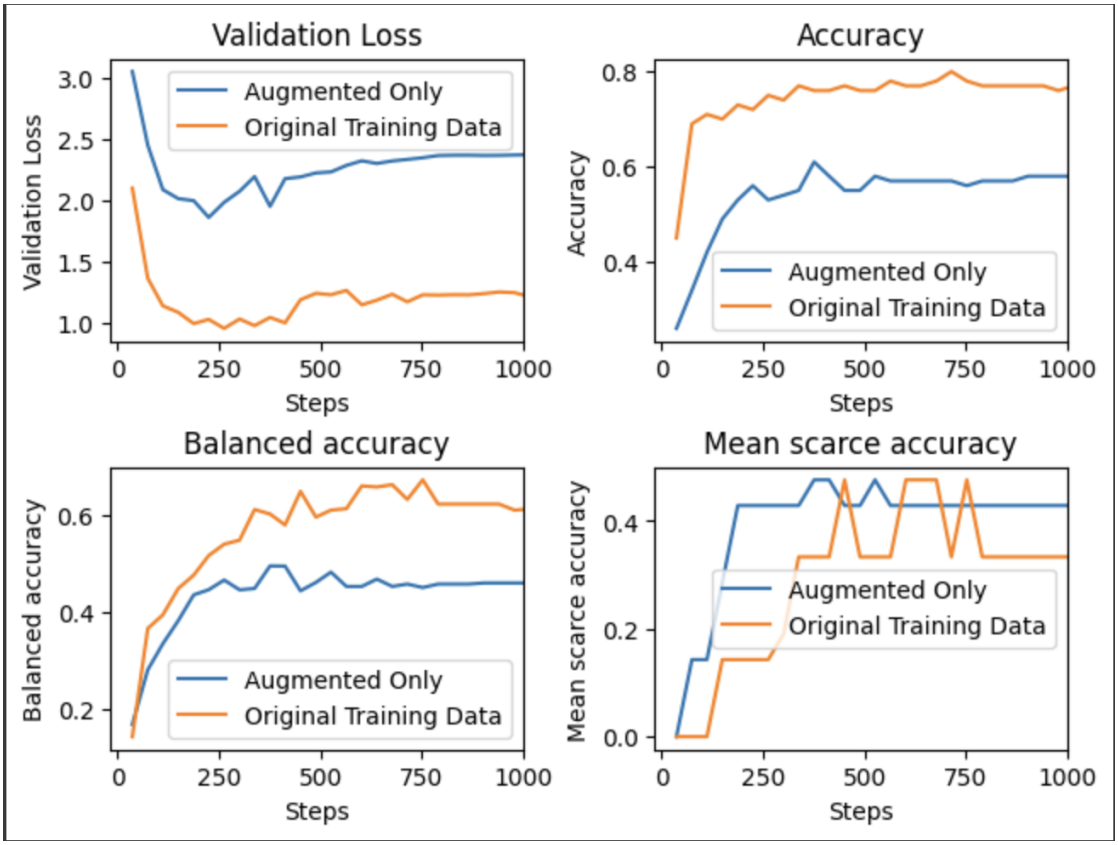

Figure 2: Classification metrics when the model is fine-tuned with synthetic data only.

mean scarce accuracy initially but improves as the model is fine-tuned for 250 steps and beyond. A similar trend can be observed with balanced accuracy of the baseline model as compared to its augmented counterpart. It's clear that augmentation helps the classifier perform better for classes with limited examples, especially when the model has been trained for fewer than 250 steps. Also, data augmentation doesn't seem to have a significant impact on overall accuracy of the model evaluated against the validation set. The model is fine-tuned to minimize the loss over all classes and it can afford to have a low accuracy for scarce classes as long as the accuracy for common classes is high. Since the ground truth for the test data was not released, we were not able to confirm these trends with the test set and had to limit the analyses to the validation set.

4.2. Quality of synthetic data

In Figure 2, we study the quality of the synthetic data generated by **gpt-3.5-turbo** by fine-tuning the classifier on synthetic data only. We can see that the model fine-tuned solely on synthetic data exhibits poor performance across all metrics except *mean scarce accuracy*. This suggests that the synthetic data generated by **gpt-3.5-turbo** falls short of exactly representing

the distribution of the training data. This behavior could be attributed to the limited training data included in the data augmentation prompt shown to the model.

5. System Specification

In the spirit of reproducibility, we share details of the systems used to run these experiments. The models were fine-tuned on an *A100* Google Colab notebook instance [3]. HuggingFace's Python package transformers [11] version 4.27.1 was used in a Python3.8 environment. Reported results were aggregated from 4 different runs using 4 different random seeds.

6. Limitations

While our method demonstrated strong performance on the classification task in MEDIQA-Sum 2023, we acknowledge that there are limitations to our approach. Firstly, our data augmentation technique relies on a third-party API, which may pose challenges to HIPAA compliance when dealing with real-world medical data.

Another potential limitation involves the sensitivity of our method to the chosen prompt for data generation. Although we designed our prompt with careful consideration, the nature of few-shot learning using large language models often requires substantial experimentation and exploration to identify an optimal prompt that results in the most effective synthetic data.

Lastly, we note that the BART model has a maximum input length of 1024 tokens. Given that the MEDIQA-Sum 2023 dataset contained snippets of clinical conversations, this constraint did not pose a problem in our case. However, this approach might not generalize well to longer or full-length clinical conversations, which could exceed this token limit.

7. Conclusion

Our work in this paper underscores the power of Large Language Models (LLMs) and their potential when combined with data augmentation techniques, particularly in generating synthetic data for clinical NLP tasks. Despite the limitations noted above, we believe our approach offers a valuable contribution to the field of NLP in healthcare, as it demonstrated significant improvements in the classification of conversations with limited number of annotations for Task A of MEDIQA-Sum 2023.

Our results are a testament to the effectiveness of this approach—using a combination of **gpt-3.5-turbo** for data augmentation and the BART model for sequence classification, Team Cadence achieved Rank-1 on the MEDIQA-Sum 2023 leaderboard for Task A.

Moving forward, we anticipate that further improvements could be made by exploring other data augmentation techniques, optimizing prompt design, and investigating models that can handle longer text sequences. We hope our work will inspire further research and development in this important and rapidly evolving field.

[3] https://research.google.com/colaboratory/faq.html#whats-colaboratory

References

[1] W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, Overview of the mediqa-sum task at imageclef 2023: Summarization and classification of doctor-patient conversations, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[2] A. Sharma, D. I. Feldman, A. Jain, Team cadence at mediqa-chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models., in: ACL-ClinicalNLP 2023, 2023.

[3] A. Ben Abacha, W. wai Yim, G. Adams, N. Snider, M. Yetisgen, Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations, in: ACL-ClinicalNLP 2023, 2023.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[5] M. A. Alkureishi, W. W. Lee, M. Lyons, V. G. Press, S. Imam, A. Nkansah-Amankra, D. Werner, V. M. Arora, Impact of electronic medical record use on the patient–doctor relationship and communication: a systematic review, Journal of general internal medicine 31 (2016) 548–560.

[6] M. M. van Buchem, H. Boosman, M. P. Bauer, I. M. Kant, S. A. Cammel, E. W. Steyerberg, The digital scribe in clinical practice: a scoping review and research agenda, NPJ digital medicine 4 (2021) 57.

[7] X. Zhu, G. Penn, Summarization of spontaneous conversations, in: Ninth International Conference on Spoken Language Processing, 2006.

[8] T. Knoll, F. Moramarco, A. P. Korfiatis, R. Young, C. Ruffini, M. Perera, C. Perstl, E. Reiter, A. Belz, A. Savkov, User-driven research of medical note generation software, arXiv preprint arXiv:2205.02549 (2022).

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[10] B. Chintagunta, N. Katariya, X. Amatriain, A. Kannan, Medically aware gpt-3 as a data generator for medical dialogue summarization, in: Machine Learning for Healthcare Conference, PMLR, 2021, pp. 354–372.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.

[12] K. H. Brodersen, C. S. Ong, K. E. Stephan, J. M. Buhmann, The balanced accuracy and its posterior distribution, in: 2010 20th international conference on pattern recognition, IEEE, 2010, pp. 3121–3124.